

©2012

**Tung T. Nguyen**

ALL RIGHTS RESERVED

SYSTEMS BIOLOGY APPROACHES TO CORTICOSTEROID  
PHARMACOGENOMICS AND SYSTEMIC INFLAMMATION

by

TUNG THANH NGUYEN

A Dissertation submitted to the  
Graduate School-New Brunswick  
Rutgers, The State University of New Jersey  
in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Ioannis P. Androulakis, PhD

and approved by

---

---

---

---

New Brunswick, New Jersey

January, 2012

## ABSTRACT OF THE DISSERTATION

# **Systems Biology Approaches to Corticosteroid Pharmacogenomics and Systemic Inflammation**

By TUNG THANH NGUYEN

Dissertation Director:

Ioannis P. Androulakis, PhD

Despite increasing knowledge about pathophysiological pathways and cellular processes involved in diseases, the molecular mechanisms and physiological significance are not fully understood. Consequently, within this exploratory research we wish to lay the foundations for developing bioinformatics tools and systems biology approaches towards the analysis and modeling of transcriptional dynamics and the understanding of gene transcriptional regulatory program. Two *in vivo* models, namely corticosteroid pharmacogenomics in rat and human endotoxemia in human, have been investigated to gain insights into (1) adverse-effects, tissue-specificity, and circadian effects under corticosteroid treatment, (2) temporal regulatory programs in acute inflammation, and (3) cellular variability and synchronization as well as time-dependent systemic responses under acute stress.

In order to pursue these goals, the hypothesis that informative components of the genome-wide transcriptional dynamics are composed of genes which are either co-expressed and co-functional or co-expressed across multiple conditions has been pursued

to identify significant genome-wide transcriptional signatures. Concepts from consensus clustering and meta-analysis have been explored to avoid the bias and assumption of each single clustering method/metric and handle challenges in the analysis of microarray data from heterogeneous sources. Subsequently, the mysteries and complexities of transcriptional regulation have been explored by using two main strategies, namely phylogenetic foot-printing and context-specific CRM search, to identify relevant transcriptional regulators and examine the putative temporal transcriptional regulatory program. Additionally, an *in silico* multi-level agent-based model of human endotoxemia has been constructed to gain insights into cellular behaviors and circadian effects under acute stress. The model captures stochastic transcriptional dynamics and critical aspects of the *in vivo* physiological human endotoxemia model. By defining novel hypothetical quantities such as the variability-based fitness and the synchronization level, we provided a step forward to the exploration of cell-to-cell variability and stochastic dynamics of cellular behaviors as well as predictive implications inferred from cellular variability.

In summary, our work aims at (i) identification of critical transcriptional signatures and regulatory controls to provide a better understanding of system behaviors and (ii) simulation to understand the cellular behaviors and circadian effects within specific contexts.

## ACKNOWLEDGEMENTS

First and foremost I would like to express my truthful gratitude to my advisor, Dr. Ioannis Androulakis who has spent a lot of time during the past five years to supervise and guide me from the very early stage of my research. Besides providing me extraordinary experiences and guidance to make my Ph.D. experience more productive and stimulating, he also taught me how to ask a question or how to express an idea in the scientific area as well as the persistency to accomplish a goal. I am heartily thankful to him more than he knows and I hope to have more opportunities to work with him in the future.

I have been fortunately to work in an interdisciplinary research area and collaborate with Dr. Steven Calvano and Dr. Stephen Lowry at Surgery Department, RWJMS – UMDNJ, with Dr. Patrizia Casaccia at Mount Sinai School of Medicine, and with Dr. Siddharth Sukumaran, Dr. Richard Almon, and Dr. William Jusko at Biological Sciences – SUNY Buffalo, to whom I owe great appreciation. I would like to thank all of them for taking time to share with me their valuable comments on the results I obtained and great insights into my research direction. Further, I would also acknowledge the rest of my committee members – Dr. Gyan Bhanot and Dr. Troy Shinbrot who have provided useful comments on my science as a whole.

I would like to acknowledge the funding sources which supported me throughout my graduate student life and made this work possible. Support has been provided by the

National Institutes of Health (NIH R01-GM082974), the National Science Foundation (NSF-BES 0519563) and the EPA (GAD R 832721-010). Also, many thanks go in particular to my colleagues, IPA lab, including Jeremy Scheff, Qian Yang, Pantelis Mavroudis, John Mattick, Eric Yang (Alumni), Peggy Foteinou (Alumni), Mehmet Orman (Alumni), Meric Ovacik (Alumni), and Kaiyuan He (Alumni).

Lastly but not least, I would like to thank my beloved mother (Mrs. Loi Lai), my aunt (Ms. Loc Lai), and my darling (Ms. Trang Bui). Words fail me to express my appreciation to them but they are all in my mind. Without their encouragement and love, I would not be who I am now. Thank you!

# Contents

<b>ABSTRACT.....</b>	<b>ii</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>iv</b>
<b>List of Tables .....</b>	<b>x</b>
<b>List of Figures.....</b>	<b>xi</b>
<b>Chapter 1 – Background and significance .....</b>	<b>1</b>
<b>1.1. Corticosteroid pharmacogenomics.....</b>	<b>1</b>
<b>1.2. Systemic inflammation .....</b>	<b>3</b>
1.2.1. Clinical relevance.....	3
1.2.2. Human endotoxemia model .....	6
<b>1.3. Significance.....</b>	<b>8</b>
<b>Chapter 2 – Materials and data .....</b>	<b>12</b>
<b>2.1. Expression data .....</b>	<b>12</b>
2.1.1. Corticosteroid pharmacogenomics.....	12
2.1.2. Human endotoxemia model .....	14
2.1.3. Synthetic data.....	15
<b>2.2. Promoter data.....</b>	<b>17</b>
2.2.1. Promoter sequences .....	17
2.2.2. TF profiles.....	18

<b>Chapter 3 – Identification of critical transcriptional modules .....</b>	<b>19</b>
<b>3.1. The ‘true’ expression profiles .....</b>	<b>19</b>
3.1.1. Background .....	19
3.1.2. The statistical model .....	22
3.1.3. The clustering effectiveness of the ‘true’ expression profiles .....	25
<b>3.2. Consensus clustering.....</b>	<b>28</b>
3.2.1. Background .....	29
3.2.2. The agreement matrix .....	33
3.2.3. The optimal suggestive number of clusters .....	34
3.2.4. Clusterable data.....	37
3.2.5. Consensus clustering.....	39
3.2.6. Method evaluation .....	41
<b>3.3. Multi-plus clustering.....</b>	<b>49</b>
3.3.1. Background .....	50
3.3.2. Problem definition .....	52
3.3.3. The pre-processing step .....	54
3.3.4. Construction of the meta-agreement matrix .....	55
3.3.5. Selection and clustering .....	56
3.3.6. Merging similar patterns.....	61
3.3.7. Method evaluation on synthetic data .....	62
<b>3.4. Results from corticosteroids pharmacogenomics model .....</b>	<b>65</b>
3.4.1. Critical transcriptional modules.....	67
3.4.2. Functional characterization of critical transcriptional modules.....	70



3.5.	Results from human endotoxemia model .....	74
3.6.	Summary.....	78
<b>Chapter 4 – Reconstruction of the transcriptional regulatory program .....</b>		<b>81</b>
4.1.	Introduction to transcriptional regulation .....	81
4.2.	Gene promoter structure.....	83
4.2.1.	Gene structure .....	83
4.2.2.	Promoter elements .....	84
4.2.3.	Promoters identification.....	87
4.3.	Binding site representation .....	90
4.4.	Discovery of ‘physical’ transcription factor binding sites.....	94
4.5.	Phylogenetic footprinting .....	98
4.6.	Context-specific transcriptional regulators.....	101
4.6.1.	cis-regulatory modules (CRMs).....	103
4.6.2.	Discovery of TFBSs and promoter profiles .....	105
4.6.3.	Common cis-regulatory modules .....	106
4.6.4.	Statistical significance of CRMs.....	107
4.6.5.	Other relevant issues .....	109
4.7.	Putative transcriptional regulatory program.....	112
4.7.1.	Results from corticosteroids pharmacogenomics model .....	112
4.7.2.	Results from human endotoxemia model .....	115
4.8.	Limitations and advantages .....	127

<b>Chapter 5 – Cellular variability and circadian control in human endotoxemia.....</b>	<b>131</b>
<b>5.1. Introduction.....</b>	<b>131</b>
<b>5.2. The in silico model of human endotoxemia .....</b>	<b>134</b>
5.2.1. The system dynamics model.....	134
5.2.2. Agent rules .....	139
5.2.3. Agent movements and interactions.....	141
5.2.4. Model parameters.....	144
<b>5.3. Qualitative assessment of model behaviors with experimental observations</b>	<b>149</b>
<b>5.4. Cellular variability and stochastic behaviors .....</b>	<b>152</b>
5.4.1. Variability-based fitness .....	152
5.4.2. Synchronization .....	156
<b>5.5. Other relevant issues.....</b>	<b>158</b>
5.5.1. Time-dependent effects under endotoxin treatment .....	158
5.5.2. Sensitivity analysis.....	160
<b>5.6. Conclusions.....</b>	<b>161</b>
<b>Chapter 6 – Summary and Future Perspectives .....</b>	<b>163</b>
<b>6.1. Summary.....</b>	<b>163</b>
<b>6.2. Future perspectives.....</b>	<b>167</b>
<b>Bibliography .....</b>	<b>171</b>

## List of Tables

<b>Table 3.1:</b> Prediction the number of clusters by the process automatically .....	44
<b>Table 3.2:</b> Accuracy of the selection and clustering on the synthetic class structure .....	47
<b>Table 3.3:</b> Accuracy of running one clustering element on the entire dataset .....	48
<b>Table 3.4:</b> Friedman-Rafsky test for clusterability on high-noise synthetic sets .....	49
<b>Table 3.5:</b> The clustering effectiveness of the approach .....	63
<b>Table 3.6:</b> Effectiveness of the approach on synthetic data .....	64
<b>Table 3.7:</b> Characterization of significant transcriptional modules .....	67
<b>Table 3.8:</b> Connecting CS transcriptional modules to enriched gene ontology terms .....	71
<b>Table 3.9:</b> Connecting CS transcriptional modules to enriched biological pathways.....	73
<b>Table 3.10:</b> Pathway enrichment in four selected patterns.....	76
<b>Table 4.1:</b> Selected resources and relevant tools for <i>in silico</i> TFBS identification. ....	96
<b>Table 4.2:</b> Data information and inflammation-relevant significant functions.....	117
<b>Table 4.3:</b> Critical transcription factors in human endotoxemia model.....	120
<b>Table 4.4:</b> Statistical significance of selected <i>cis</i> -regulatory modules .....	125
<b>Table 4.5:</b> Critical transcription factors identified from the <i>in vitro</i> endotoxin study ...	126
<b>Table 5.1:</b> Model components .....	138
<b>Table 5.2:</b> Model rules .....	140
<b>Table 5.3:</b> Model production parameters .....	147
<b>Table 5.4:</b> Effects of production parameters on system behaviors .....	161

## List of Figures

<b>Figure 3.1:</b> The ‘true’ expression profiles are more robust than the average ones.....	24
<b>Figure 3.2:</b> The performance of typical clustering methods on different error-measurement integrated approaches .....	26
<b>Figure 3.3:</b> Schematic overview of microarray data analysis using multiple clustering runs to select a ‘clusterable’ subset – the subset which contains genes that are either highly coexpressed or non-coexpressed with a confidence level $\delta\%$ . ....	32
<b>Figure 3.4:</b> An example of the agreement matrix. The left is the results from N clustering runs. The right shows the corresponding agreement matrix that each entry $M_{ij}$ is the frequency of gene i and gene j grouped into the same cluster. ....	34
<b>Figure 3.5:</b> Histogram of AM entries (left) and the corresponding CDF curve (right) from the AM in Figure 3.4.....	36
<b>Figure 3.6:</b> The gene selection process .....	38
<b>Figure 3.7:</b> Illustration of the consensus clustering on the AM in Figure 3.4 .....	40
<b>Figure 3.8:</b> Examining the agreement matrix. (a) Average histograms of the AM on 100 random datasets (b) Histogram of the AM (top) and the CDF lines, AUC curve, as well as Gap curve on high-noise synthetic set and (c) on low-noise synthetic sets.....	42
<b>Figure 3.9:</b> Illustration the selection and clustering as well as the effect of different confidence levels $\delta$ .....	46
<b>Figure 3.10:</b> A brief look on the synthetic data and selected genes from low-noise set 1 (top) and high-noise set 1 (bottom).....	46
<b>Figure 3.11:</b> The flowchart of the multi-plus clustering approach .....	54

<b>Figure 3.12:</b> The procedure of selection and clustering. The left is the pseudo-code algorithm of the procedure. The right is an example to illustrate the process with a specific AM.....	59
<b>Figure 3.13:</b> Estimating the cluster significance threshold given a user-defined p-val. ..	60
<b>Figure 3.14:</b> Critical transcriptional modules of CS pharmacogenomic effects. Each module is characterized by the average gene expression profile of the corresponding cluster in the acute and the chronic data..	68
<b>Figure 3.15:</b> Selected genes and patterns from LPS dataset. The initial number of clusters is six and we picked out five distinguished patterns (3 up- and 2 down- regulation) in which cyan pattern can be omitted since it is not significant .....	75
<b>Figure 4.1:</b> Basic structure of promoter classes. (a) A general structure of an eukaryote gene; (b)(c)(d) Typical structures of promoters classes.....	83
<b>Figure 4.2:</b> Class II promoter structure and relevant regulatory elements. (a) Typical regulatory elements of a gene; (b) A detailed structure of a core promote; (c) Four typical types of distal regulatory elements and their corresponding effects.....	86
<b>Figure 4.3:</b> Data complexities in TFBS prediction. (a) Alternative promoters for genes in higher eukaryotes (b) Alternative sets of combinatorial TFs regulate the transcription process even though only one promoter is activated in these contexts.....	89
<b>Figure 4.4:</b> Binding site representation. (a) Illustration of several motif models for human factor ETS1. (b) A brief look on the history of binding motif models. ....	93
<b>Figure 4.5:</b> Identification of promoter conserved regions and common physical TFBSs. (a) Estimation of conserved regions on a single promoter (the red one) based on Dialign's alignment scores from a set of orthologous promoters. (b) Finding common physical	

TFBSs accounting for the case that genes may have multiple alternative promoters. TFBSs present on the conserved regions of any alternative promoter of a gene are also considered as putative TFBSs for that gene.....	100
<b>Figure 4.6:</b> Flowchart of the CRM searching process. ....	105
<b>Figure 4.7:</b> Statistical significance thresholds of CRMs.....	109
<b>Figure 4.8:</b> Putative regulation of CS transcriptional modules by enriched TFBSs. Those TF families with ‘blue’ border lines consist of transcription factors that are affected under corticosteroid administration in this study. The results show a putatively dynamic perspective of regulation between transcriptional regulators and involved sets of genes. .....	114
<b>Figure 4.9:</b> Putative temporal regulatory program in human endotoxemia plus schematic illustration of the integrated computational framework.....	123
<b>Figure 5.1:</b> <i>In silico</i> human endotoxemia model accounting for circadian variability. (a) The system dynamic model. (b) A snapshot of the implemented model. Molecules are displayed with solid circles (P: red-; A: magenta-; F: blue-; M: cyan-; NFkB: yellow-; E: green-; TLR & GR: white-; Ikb, IKK, NFkB.Ikb: black- circles). Cells are displayed with solid squares where green squares represent for cells with an approximate number of P and A, red squares for those with the number of P greater than 1.5 fold of the number of A and magenta squares for those with A more than 1.5 fold of P.....	137
<b>Figure 5.2:</b> Dynamics patterns of selected components under circadian control. Circadian control is regulated by the rhythms of cortisol (F) and melatonin (M) which in turn drive the patterns of other components in the system.....	150

<b>Figure 5.3:</b> Correspondence between <i>in vivo</i> - and <i>in silico</i> - system responses to endotoxin. The patterns between <i>in vivo</i> - and <i>in silico</i> - responses are matched to define the time-scale for the system.....	152
<b>Figure 5.4:</b> Stochastic dynamics in cell population. The stochastic behaviors of pro-inflammatory cytokines (a) and anti-inflammatory cytokines (b) in three different cells are shown in the top-panel. (c) Cells are displayed with solid squares where green squares represent for cells with an approximate number of P and A, red squares for those with the number of P much greater than the number of A and magenta squares for those with $A \gg P$ . .....	154
<b>Figure 5.5:</b> Cellular variability and synchronization behaviors. The top-panel displays the pattern of variability-based fitness of a simulated day in the homeostatic system and of the day where endotoxin is treated at 9:00AM. The bottom panel shows the synchronization level of all cells in the system. ....	155
<b>Figure 5.6:</b> Time-dependence system responses to endotoxin administration. The strength of the inflammatory response or the vulnerability of the host fitness is characterized by (a) the maximal peak of pro-inflammatory cytokines ( $P_{\max}$ ) and (b) the minimum peak of variability-based fitness versus the time of endotoxin treatment. ....	159

## **Chapter 1 – Background and significance**

### **1.1. Corticosteroid pharmacogenomics**

Glucocorticoids (GC) are a class of steroid hormones present in almost every animal cell, playing a central role in a wide range of physiological responses [1]. Because of their potent anti-inflammatory and immunosuppressive effects, synthetic glucocorticoids referred as corticosteroids (CS) (e.g. methylprednisolone - MPL) have been used widely in pharmacology as a therapeutic option for a wide range of autoimmune and inflammatory diseases [2, 3]. However, beneficial effects are derived from magnifying the physiological actions of endogenous glucocorticoids, causing a variety of side effects following long-term treatment with this class of drugs e.g. hyperglycemia, dyslipidemia, arteriosclerosis, muscle wasting, and osteoporosis [4-7]. The physiological and pharmacological effects of corticosteroids are complex and manifest themselves with expression changes of many genes across multiple tissues [8-10]. It has been observed that even in a single tissue different dosing regimens of CS administration can induce different patterns of expression [11-13]. As such genes with similar expression profiles under acute CS administration may not exhibit similar expression patterns during continuous infusion, pointing to the possibility of alternative regulatory mechanisms. Therefore, a better understanding of corticosteroid pharmacogenomic effects from multiple dosing regimens are very valuable not only to reveal the transcriptional dynamics under different patterns of input perturbations but also to provide an insight into the underlying molecular mechanisms of action, for both beneficial and detrimental effects, and thus for the optimization of clinical therapies.



However, it has been noted that genes affected by CS include both immunosuppressive genes, mostly associated with therapeutic effects, and metabolic genes often associated with adverse effects whose regulation is mainly controlled by glucocorticoid receptor gene mediated pathways [6]. Unbound CS binds with cytosolic free glucocorticoid receptors (GR) releasing it from the heat shock complex allowing dimerization and translocation into the nucleus where it binds to glucocorticoid response element (GRE) of the target genes, leading to enhancement or inhibition of the target gene expression. As a result, long-term treatment with corticosteroids results in sustained up- or down-regulation of numerous genes, leading to a new steady state which might be the basis for occurrence of adverse effects. However, it has also been noted that chronic infusion of CS causes a sustained down-regulation of the receptor (mRNA and thus protein) [14, 15]. While several alternative mechanisms have been proposed [16-18] it is still not understood why drug effects remain strong although GR mRNA is down-regulated to the point of almost being eliminated. A plausible explanation is that besides direct regulation through GRE binding sites in the 5' regulatory regions of genes, there are changes in expression that are also the indirect results of effects due to changes in expression of transcription factors (TFs) that act as secondary biosignals directly or indirectly modulating the transcription of genes [15, 19, 20]. Therefore, identification of putative regulatory control structures is also an essential step towards understanding corticosteroid effects. Consequently, our studies in this aspect focus on exploring the complexity of high-dimensional transcriptional expression profiles to discover critical transcriptional modules and regulatory control structures as well as certain sets of genes that are responsible for corticosteroid side-effects.

## **1.2. Systemic inflammation**

### **1.2.1. Clinical relevance**

The systemic inflammatory response syndrome (SIRS) often accompanies critical illness but is evoked by many stimuli e.g. infection, trauma, invasive surgery and biological stressors in general [21]. While the host inflammatory response is essential in controlling the stimulus, it also has a central pathogenic role in the development and severity of sepsis syndromes [22-24]. In the United States (US), more than 700,000 patients per year develop sepsis with an estimated rising incidence of ~1.5% per year [25, 26]. The average costs per case were ~\$22,100, resulting in an economic burden of nearly \$17 billion annually in the US alone [25]. Despite more than 20 years of extensive research, sepsis and SIRS remain the chief causes of death, killing 30 to 50 percent of severely affected patients. It is a leading cause of mortality among patients in non-cardiac intensive care units (ICUs) [27] and the 10th leading cause of death overall in the US [28]. Although the overall mortality rate among patient with sepsis is declining in recent years, the incidence of sepsis and the number of sepsis-related cases are still increasing [29-31]. Additionally, sepsis substantially reduces the quality of life of survivals [32, 33]. In an attempt to search for an efficacious therapy that reduces mortality, a lot of therapeutic strategies for the treatment of sepsis have been developed [34-36]. Glucocorticoids (GC), one of the most traditionally and exhaustively studied therapies for sepsis, have been shown to have anti-inflammatory properties and improve the vascular reactivity. Although the administration on animal models of sepsis provided improved outcomes, early clinical trials using short courses of high-dose steroids (up to 600 mg/kg of steroids over 24 hours) revealed harmful effects to patients [37]. Furthermore,

prolonged glucocorticoid therapy may result in deleterious side effects [6]. However, recent clinical trials using low-dose of steroids have demonstrated the beneficial effects of glucocorticoids to patients with vasopressor-dependent septic shock [38]. The second most widely known therapeutic strategy is the antiserum and antibodies against endotoxin from Gram-negative bacteria, lipopolysaccharide (LPS). This therapeutic target received greater interest as it was hypothesized that it may not only be responsible for sepsis but also be a mediator in all forms of shock [35]. Nonetheless, treatment with antiserum/antibodies was shown to be beneficial in animal models but most patients with sepsis often fail in responding to LPS-inhibitors [39, 40]. Among possibilities is that phagocytes may be subject to an endogenous stimulation in which heparan sulfate appears to trigger the same downstream signals as endotoxin [41, 42]. Moreover, different infectious pathogens can stimulate different mechanisms of the host response [43]. The next therapeutic strategy in sepsis is mediator-specific anti-inflammatory agents that can reduce or inhibit strong pro-inflammatory cytokines found in septic patients (e.g.  $\text{TNF-}\alpha$ , IL-1) [44]. In the similar manner with anti-LPS agents, anti-inflammatory mediators (e.g. anti-TNF antibodies, IL-1 receptor antagonists) were proved to show beneficial effects to the development of pathophysiological changes associated with sepsis and survival on animal studies. However, it is still unclear why these agents become less beneficial and even harmful in human clinical trials [45, 46]. Besides, the coagulation system has also been an important target for clinical therapies of sepsis. Of anti-coagulant agents, activated protein C (APC) has demonstrated its benefits on survival rates and was approved in the US as the first drug for clinical use to patients with severe sepsis and high risk of death [47, 48]. Additionally, a number of other late-

acting mediators have also been explored and considered as potential targets for developing novel therapeutic strategies in sepsis e.g. macrophage migration inhibitory factor (MIF) [49], high-mobility group B1 protein (HMGB1) [50], complement C5a [51, 52], as reviewed extensively elsewhere [36, 53, 54].

All in all, despite increasing knowledge about pathophysiological pathways and processes involved in sepsis as well as promising results on animal studies and preclinical trials, the vast majority of large, randomized clinical trials to patients showed little success in reducing the high mortality rates [25, 34, 55]. The fundamental rationale of such trials was that a mediator which is observed to be persistently elevated and detectable in septic patients should be blocked [56]. However, clinical trials have failed to show a significant improvement in survival, calling into a question whether modulating a particular pathway or mediator of the inflammatory response should be reduced i.e. shifting the perspective from the component-level to the system-level where inter-relationships among components and dynamic patterns of change are noticed as important factors [57, 58]. To address such problem as well as the rising cost of production and approval of new drug candidates for all diseases, the US Food and Drug Administration recently stated that ‘A new product development tool kit-containing powerful new scientific and technical methods such as animal or computer-based predictive models, biomarkers for safety and effectiveness, and new clinical evaluation techniques – is urgently needed to improve predictability and efficiency along the critical path from laboratory concept to commercial product’ [59].

### 1.2.2. Human endotoxemia model

Inflammation and activation of innate immunity are essential defense responses against invading pathogens and endogenous danger signals. The innate immune response involves the initial recognition of conserved pathogen-associated molecular patterns by members of the Toll-like receptor (TLR) family [60]. The exposure of the host to gram negative bacteria, simulated by lipopolysaccharide (LPS) recognized by TLR-4, triggers intracellular signalling cascades which eventually release a lot of pro- and anti-inflammatory cytokines [61]. While the host inflammatory response is essential to resolve the infection or repair the damage and restore the system homeostasis, it also plays a central pathogenic role in a wide spectrum of diseases including sepsis [62]. Under healthy circumstances, inflammatory responses are activated, clear the pathogen in the case of infection, initialize a repair process and then abate [23]. However when anti-inflammatory processes fail, an amplified inflammation can turn what is normally a beneficial reparative process into a detrimental physiological state with severe, uncontrolled systemic inflammation [24].

***In vivo model of human inflammation:*** To learn more about the mechanisms associated with the host inflammatory response, human endotoxemia models have been proposed in which a single intravenous bolus of *E. coli* endotoxin (LPS) is given to healthy human subjects. The model results in many similar physiological host responses that characterize Gram-negative bacteria infection [63], providing an invaluable source for the systemic identification of biological features representing the complex dynamics of a host undergoing inflammatory responses [64, 65]. Studies involving experimental human endotoxemia have reported rapid intravenous infusion in doses of 2-4ng/kg body weight,

which effectively induces an acute systemic inflammatory condition that mimics the early flow phase of injury and infection [63, 66-69]. In human peripheral blood leukocytes, intravenous administration of endotoxin elicits dynamic and reproducible changes in the circulating leukocyte population as well as significant changes in blood leukocyte gene expression patterns [64]. This perturbation of leukocyte gene expression involves several thousands of transcripts and accompanies the systemic physiological responses during inflammation, which peaks ~4-6 hours after endotoxin exposure and resolves within 24 hours, compatible with a large and dynamic regulatory network.

***In silico model of human endotoxemia:*** In parallel with *in vivo* model, *in silico* models of inflammation have also been developed to study the complex interplay between beneficial and harmful arms relevant to inflammatory responses [70-73]. With the primary goal of optimizing clinical practice, mechanistic simulations have been advocated to understand and predict the systemic behaviors seen in clinical settings [74-76]. As such, computational models offer a promising possibility for improving the interpretation of quantitative experimental data as well as generating and exploring simultaneously multiple hypotheses [75-78]. It is a means of knowledge representation that can help to (i) better understand the underlying molecular mechanism of the host inflammatory response, (ii) examine systemic effects under different initial experimental conditions, (iii) perform successively different and/or simultaneously multiple experiments coupled with different testable hypotheses, and eventually (iv) discover common features leading to distinct outcomes if applicable. Multiple modeling methods have been developed in the state-of-the-art but generally they can be classified into two main categories – equation-based [79, 80] and agent-based modeling [70, 71, 77], each of

which has its own strengths and limitations but the two disciplines ignored each other's literature almost entirely although the study area is significantly overlapped [81]. The most popular approach of the equation-based category in characterizing inflammation is using ordinary differential equations (ODEs) [82-87]. However, ODEs is fully deterministic with respect to the systemic behavior given a certain set of initial conditions and assumes the homogeneity and perfect mixing within compartments as well as ignores the spatial aspect [73, 88], requiring the employment of alternative approaches e.g. partial differential equations [89], stochastic differential equations [90, 91], and eventually the other category – agent-based modeling [70, 75, 92-94].

### **1.3. Significance**

Life science is being at the age of transition from descriptive to mechanistic approaches that explore underlying principles from molecules to cells, organs and their interactions across multiple scales of biological organization. Among key concepts to these analyses is the concept of 'network' [95]. In the context of biological systems, the implication is that macroscopic responses of a system are the results of propagating information, in the form of disturbances, across an intricate web of interactions at multiple biological scales. These interactions define elementary building motifs that are organized into intracellular pathways and regulatory structures, which in turn are integrated into interacting modules that eventually give rise to an organism's response.

As modulating gene expression levels is among the key regulatory responses of an organism to changes in its environment and/or external stimuli, identifying biologically relevant transcriptional regulators and their putative regulatory interactions with target genes is an essential step towards the study of the complex dynamics of gene regulatory

network. It has been hypothesized that one of the primary mechanisms for gene regulation is via transcription factor binding in which a protein (transcription factor) binds to certain sites in the genome [96-98]. The discovery of gene regulatory elements requires the synergism between computational and experimental techniques in order to reveal the underlying regulatory mechanisms that drive gene expression in response to external cues and signals. Consequently, in this research the overall theme has been set for the development of bioinformatics tools and systems biology approaches towards the analysis and modeling of transcriptional dynamics and the understanding of gene regulatory network. Two *in vivo* models, namely corticosteroid pharmacogenomics in rat and human endotoxemia in human, have been investigated to gain insights into (1) adverse-effects, tissue-specificity, and circadian effects under corticosteroid treatment, (2) temporal regulatory programs in acute inflammation, and (3) cellular variability and synchronization as well as time-dependent systemic responses under acute stress.

In order to pursue these goals, we first identify characteristic genome-wide transcriptional signatures by exploring the hypothesis that informative components of the genome-wide transcriptional dynamics are composed of genes which are either co-expressed and co-functional or co-expressed across multi-conditions. Subsequently, relevant transcriptional regulators and putative regulatory structures relevant to the regulation of corresponding transcription dynamics are explored using two main strategies, namely phylogenetic footprinting and context-specific CRM search. Finally, we embedded those transcriptional dynamics to an integrated dynamics model to gain insights into cellular behaviors.

A couple of novel statistical methods and improvements on existing algorithms have been developed to better extract critical transcriptional modules given a high-dimensional



transcriptional profiling dataset. First, we proposed a statistical model which can integrate the error information from repeated measurements to expression profiles, generating the so-called ‘true’ expression profiles. The output can be utilized for a variety of computational models that take expression profiles as the required input without any modification while still taking into account the advantage of using replicated data. We next explored concepts in consensus clustering and the hypothesis that the more clusterable the data is the more biologically relevant it is to identify, within a set of differentially expressed genes, a subset of genes that are either highly co-expressed or highly non-coexpressed with the hope of extracting a more biologically relevant subset of genes. Additionally, following the orientation of meta-analysis an extended computational approach was also proposed to identify gene clusters that share common expression patterns across multiple gene expression datasets as well as handling challenges in the analysis of microarray data from heterogeneous sources.

In order to predict relevant transcriptional regulators and putative regulatory structures relevant to the regulation of corresponding transcription dynamics, we explored two main strategies, namely phylogenetic foot-printing and context-specific CRM search, to identify relevant transcriptional regulators and examine the putative transcriptional regulatory program. Our analysis also allows for the reconstruction of a dynamic temporal regulatory network, making it a critical enabler for improving our understanding of how the transcriptional machinery ‘program’ effectively regulates key cellular processes. To examine cellular behaviors and detailed regulatory mechanisms, more specifically the interplay between circadian control and endotoxin challenge, we construct an *in silico* multi-level agent-based model of human endotoxemia model. The

model captures the stochasticity of transcriptional dynamics and critical aspects of the *in vivo* physiological human endotoxemia model. By defining novel hypothetical quantities such as the variability-based fitness and the synchronization level, we provided a step forward to the exploration of cell-to-cell variability and stochastic dynamics of cellular behaviors as well as predictive implications inferred from cellular variability. Ultimately, our work aims at (1) identification of transcriptional signatures and regulatory controls to provide a better understanding of the system behaviors and (2) simulation to gain insights into cellular behaviors and circadian effects within specific contexts.

### ***Developing tools***

1. *ExPatt*: explore the concepts of ‘clusterable’ data and consensus clustering to identify critical transcriptional responses given a high-dimensional gene expression dataset.
2. *MP-Clustering*: a multi-plus clustering framework that extends from ‘ExPatt’ to identify co-expressed gene clusters across multiple conditions/tissues as well as handling challenges in the analysis of microarray data from heterogeneous sources, e.g. different platforms, different time-grids, different lab-protocols.
3. *‘True’ expression profiles*: integrate the error information from repeated measurements to provide a better type of gene expression profiles compared to simple average expression profiles, supporting the usage of previous computational methods without changing anything but still taking the advantage of replicate data.
4. *TF-Explorer*: predict transcriptional factors relevant to the regulation of transcriptional responses using context-specific CRM search and a novel heuristic to handle the computational complexities.
5. *Agent-based human endotoxemia model*: construct an *in silico* human endotoxemia that can mimic important characteristics of the *in vivo* human endotoxemia.

## Chapter 2 – Materials and data

### 2.1. Expression data

#### 2.1.1. Corticosteroid pharmacogenomics

##### *Acute corticosteroid administration*

Forty-seven male ADX Wistar rats weighting from 225 to 250g underwent right jugular vein cannulation under light ether anesthesia 1 day before the study [99]. Forty-three rats were injected with a single intravenous bolus dose of methylprednisolone (MPL) of 50mg/kg. Animals were sacrificed by exsanguinations under anesthesia and liver samples were harvested at 0.25, 0.5, 0.75, 1, 2, 4, 5, 5.5, 6, 7, 8, 12, 18, 30, 48, and 72 after dosing. The sampling time points were selected based on preliminary studies describing GR dynamics and enzyme induction in liver. Four untreated rats were sacrificed at random times and nominally considered as 0h controls. The gene expression was obtained via the Affymetrix RG-U34A array which consists of 8,799 probesets. The data are publicly available through the GEO Omnibus Database under the accession number GDS253. After filtering by ANOVA ( $p$ -value = 0.05) [100, 101], 2,920 probesets considered as differential expression are used for further analysis.

##### *Chronic corticosteroid administration*

In a similar experiment model, forty rats were administered a low level of 0.3 mg/kg/hr infusions of MPL over 168h via an Azlet pump [11]. The pump drug solutions were prepared for each rat based on its predose body weight. Animals were sacrificed at various times up to 7 days; specifically the time-points included are 6, 10, 13, 18, 24, 36,

48, 72, 96, and 168h. A control group of four animals was implanted with a saline-filled pump and killed at various times throughout the 7-day study period. Unlike the previous experiment, the microarray platform for this dataset is the RAE230A which consists of 15,923 probesets. The data are publicly available through the GEO Omnibus Database under the accession number GDS972. After filtering by ANOVA ( $p$ -value = 0.05), 4,361 probesets are selected as significantly differentially expressed probesets for further analysis.

### ***Circadian data***

To examine the fluctuations of gene expression patterns in liver within the 24 hour circadian cycle in normal animals, fifty four normal male Wistar rats (body weights ~ 225-275g) were housed and allowed to acclimatize in a constant-temperature environments (22<sup>0</sup>C) equipped with 12h light/dark cycle [102]. Twenty-seven rats (Group I) were acclimatized for 2 weeks prior to study to a normal light/dark cycle where lights went on at 8AM and off at 8PM whereas the other 27 rats (Group II) were acclimatized a reserved light/dark cycle where lights went on at 8PM and off at 8AM. Rats in Group I were killed in three successive days at 0.25, 1, 2, 4, 6, 8, 10, 11, 11.75hr after lights on to capture the light period. Rats in Group II were killed on three successive days at 12.25, 13, 14, 16, 18, 20, 22, 23, 23.75h after lights on to capture the dark period. Animals sacrificed at the same time on successive days were treated as triplicate measurements. The gene expression was obtained via the Affymetrix RAE230A array which consists of 15,923 probesets. The data are publicly available through the GEO Omnibus Database under the accession number GSE8988. After filtering by ANOVA ( $p$ -value = 0.05), 2,468 probesets considered as differential expression are used for further analysis.

All protocols followed the Principles of Laboratory Animal Care (Institute of Laboratory Animal Resources, 1996) and were approved by the University at Buffalo Institutional Animal Care and Use Committee.

### **2.1.2. Human endotoxemia model**

#### ***In vivo data***

The data used in this study were generated as part of the Inflammation and Host Response to Injury Large Scale Collaborative Project funded by the USPHS, U54 GM621119 [64, 103]. Human subjects were injected intravenously with endotoxin (CC-RE, lot 2) at a dose of 2-ng/kg body weight (endotoxin treated subjects) or 0.9% sodium chloride (placebo treated subjects). Following lysis of erythrocytes and isolation of total RNA from leukocyte pellets [64], biotin-labelled cRNA was hybridized to the Hu133A and Hu133B arrays containing a total of 44,924 probesets for measuring the expression level of genes that can be either activated or repressed in response to endotoxin at 0 (before treatment), 2, 4, 6, 9, and 24hr. Data are publicly available through the GEO Database (#GSE3284). ANOVA technique ( $p < 10^{-4}$ ) was then applied to filter significantly differentially expressed probesets, resulting in 3,269 selected probesets [104]. Average expression profiles of probesets over replicates for each time-point were used as the final input data for further analyses [105]. The data have been appropriately de-identified, and appropriate IRB approval and informed, written consent were obtained by the glue grant investigators [64].

#### ***In vitro data***

Isolated from peripheral blood mononuclear cells collected from three healthy humans, adherent monocytes were cultured for 10 days in RPMI medium 1640 (20% FBS/L-

glutamine/20mM Hepes/penicillin/streptomycin/50 ng/ml macrophage colony-stimulating factor) to generate peripheral-blood-derived mononuclear cells [106]. These mononuclear cells were stimulated by 100 ng/ml LPS (*Salmonella minnesota* R595 ultra pure LPS; List Biological Laboratories, Campbell, CA) and sampled at 0 (before stimulation), 2, 4, 8, and 24hr. Total RNA was isolated with TRIzol (Invitrogen, Carlsbad, CA) and two samples for each time-point were analyzed using HG-U133 Plus2 Affymetrix GeneChips producing mRNA expression profiles of 54,675 probesets (#GSE5504). Fold change (fold = 2.5) was then applied to filter significantly differentially expressed probesets, resulting in 2,892 selected probesets. Average expression profiles of probesets over replicates for each time-point were used as the final input data for further analyses [105].

### **2.1.3. Synthetic data**

#### ***2-dimensional datasets***

To provide a visual view of how the strategy works and the effect of the agreement threshold (or confidence level) on the result of selection and clustering, we utilized five two-dimension testing sets from the work of Pei and Zaiane [107]. Each set (2,000 points) was created with a density- and a noise-level corresponding to a difficulty-level, in which the data in each cluster can be uniformly or normally distributed; then some mathematical techniques such as linear transformation, linear equation and circle equation were applied to generate the final datasets. The difficulty levels spread from one to five corresponding to five testing sets (one with standard cluster shapes and well separated, two with transformed shapes and well separated, three with arbitrary shapes and clearly separated, four with arbitrary shapes with obvious or vague space inside

clusters but still clearly distinguishable, five with clusters within clusters, irregular shapes, and some clusters are bridged).

### ***High dimensional datasets***

#### *Downloaded synthetic data*

A number of synthetic datasets from the open literature are utilized to assess our approach for finding common sets of genes that are highly coexpressed across multiple conditions. Specifically, we used a series of four high-noise 20-timepoint sine-format synthetic datasets with different number of replicates at each time-point (1, 3, 4, and 20 respectively) from [105, 108]. Each dataset contain five separate sets with 400 genes allocated equally in 6 classes, each of which contains the same list of genes but has different patterns across five conditions. For each set, in the first step the data are generated according to an artificial pattern  $\Phi(i, t, l)$  which shows the values for gene  $i$  at time-point  $t$  in cluster  $l$ ; four of six clusters follow the sine function i.e.  $\Phi(i, t, l) = \sin(2\pi t/10 - w_l)$  ( $w_l$  is some random phase shift between 0 and  $2\pi$ ), and the other two follow the non-periodic linear function ( $\Phi(i, t, 5) = t/20$  and  $\Phi(i, t, 6) = -t/20$ ),  $i = 1, \dots, 400$ ,  $t = 1, \dots, 20$ ,  $l = 1, \dots, 4$ . In the second step, let  $x(i, t, r)$  be the error-added value for gene  $i$ , time-point  $t$  and repeat;  $x(i, t, r)$  is randomly drawn from a normal distribution  $N(\mu, \sigma)$  where  $\mu$  is the value of the synthetic pattern  $\Phi(i, t, l)$  and  $\sigma$  is equal to  $\lambda\sigma_{it}$  ( $\sigma_{it}$  is randomly extracted from measurement errors observed in the yeast galactose data [109] and  $\lambda$  is the multiplicative factor that controls the noise level). High-noise synthetic data are generated with  $\lambda = 6$  [105].

### *Generated synthetic data*

Following the convention of previous studies [105, 110], we generate synthetic data which contain 6 clusters of genes, each of which consists of 66 genes across  $T = 20$  time-points. Four of six clusters are generated using the sine function plus some noise

$$g_{itr} = \sin(t\omega_m/T + \varphi_m) + \alpha\sigma_i\sigma_{it}x_{itr}$$

and the other two are generated following a non-periodic linear function plus some noise

$$g_{itr} = \pm t/T + \alpha\sigma_i\sigma_{it}x_{itr}$$

Here the subscript  $m$  denotes the cluster number and  $i$ ,  $t$ ,  $r$  indicate the gene id, the time, and the replicate numbers respectively. Therefore,  $\{g_{itr}\}$  is a synthetic expression profile of a simulated gene with  $r$  replicates for each of  $T$  time-points. The parameters  $\omega_m$  and  $\varphi_m$  represent the random wavelength and random shift for cluster  $m$  ( $\omega_m \in [0.5\pi, 5\pi]$ ,  $\varphi_m \in [0, 2\pi]$ ).  $\alpha$  is the level of noise which is 1.0 for low noise and 2.5 for high noise in this study. The parameters  $\sigma_i$  and  $\sigma_{it}$  represent the error levels for gene  $i$  and for experiment at time-point  $t$  which are randomly drawn from a uniform distribution in the interval  $[0.2, 1.2]$ . Finally,  $x_{itr}$  is a random variable drawn from a standard normal distribution to create the variability for replicates.

## **2.2. Promoter data**

### **2.2.1. Promoter sequences**

Promoters of genes are extracted from a rich database of promoter information using Gene2Promoter – Genomatix [111]. Given a gene, a set of transcript-relevant promoters are extracted coupled with multiple alternative promoters and experimental information about the promoter length including those with either an experimentally defined length or



a default if there is no associated prior length information (500bp upstream plus 100bp downstream the transcription start site – TSSs).

Orthologous promoters are also extracted using Gene2Promoter tool. Each promoter is characterized by a set of promoters from orthologous genes of other vertebrate species, if available (e.g. *Homo sapiens*, *Mus musculus*, *Macaca mulatta*, *Pan troglodytes*, *Equus caballus*, *Bos Taurus*, *Gallus gallus*, etc.). To be consistent in the search for conserved regions on promoter sequences in order to identify putative transcription factor binding sites (TFBSs) we eliminate those that do not consist of more than two orthologous promoters.

### **2.2.2. TF profiles**

In order to identify putative transcriptional regulators, we utilize position weight matrices (PWMs) saved in MatBase [111] which contains about 867 matrices of vertebrate transcription factor profiles classified into 182 families (version 8.4). MatInspector [112] is then applied to scan for position weight matrix (PWM) matches on the promoter sequences with a specific optimal threshold of the matrix similarity for each PWM and a common core similarity 0.75. The core similarity is the similarity of four continuous bases at the most conserved region in the TF profiles and the optimal threshold of the matrix similarity is the one which is optimized so that only a maximum of three matches are allowed in 10,000bp of non-regulatory test sequences (supported from MatBase [111]).

## **Chapter 3 – Identification of critical transcriptional modules**

### **3.1. The ‘true’ expression profiles**

Microarray technology is a powerful and widely accepted experimental technique in molecular biology that allows studying genome wide transcriptional responses. However, experimental data usually contain potential sources of uncertainty and thus many experiments are now designed with repeated measurements to better assess such inherent variability. Many computational methods have been proposed to account for the variability in replicates. As yet, there is no model to output expression profiles accounting for replicate information so that a variety of computational models that take the expression profiles as the input data can explore this information without any modification. Thus we here propose a methodology which integrates replicate variability into expression profiles, to generate so-called ‘true’ expression profiles. The model utilizes a previously proposed error model and the concept of ‘relative difference’. The clustering effectiveness when using this ‘true’ profile coupled with clustering techniques is demonstrated through synthetic data where several methods are compared.

#### **3.1.1. Background**

Global gene expression analysis using microarrays has become an essential tool to study genome-wide transcriptional responses. Although this high-throughput technology produces a huge volume of useful data, enabling researchers to study the response of thousands of genes simultaneously, it faces many potential sources of uncertainties (e.g. technical noise, experimental treatments, biological sampling) [113, 114]. As such, a

number of statistical methods have demonstrated that the information contained in replicates is a valuable asset in order to assign proper confidence levels [115-118]. Rocke et al. [119] proposed a model accounting for measurement error to model gene expression profiles which has been used often in conjunction with variance-stabilizing transformation [120-123] and model-based clustering [124, 125]. Consequently, researchers are designing more experiments with repeated measurements per gene per chip even though it is significantly more costly and time consuming. However, properly incorporating the replicate information remains a challenge.

A typical step in analyzing gene microarray data involves filtering for differential expression [126]. A number of methods have been proposed in this direction demonstrating the extensive insight gained in utilizing the information from replicates for determining the change of gene expression values e.g. t-test [127-129], ANOVA [100, 130], SAM [131], EDGE [132]. An equally important part of the analysis is clustering which has been proven to be a powerful tool to rationalize transcriptional responses, identify possible functional relationships among them, and further elucidate important transcription factors as well as relevant biological pathways [125]. However, most clustering methods do not take into account the variability of gene expression profiles in the form of replicates. Variability is usually lumped into a mean effect and expression profiles are clustered based on average values of independently repeated measurements for each gene, thus missing, potentially, useful information [124].

Given that replicates can provide important insights into the nature of inherent variability among gene expression profiles [115], recent approaches have attempted to incorporate repeated measurements. There are two primary ways to handle replicated data: (i)

indirectly integrate the error information among replicates into a pairwise similarity metric between two expression profiles to produce a more robust distance metric, and (ii) directly integrate the replicate information into clustering models. The former offers a relative advantage since clustering methods that take the distance metric as input can be utilized without any modification e.g. standard deviation-weighted correlation coefficient [133], shrinkage correlation coefficient [110]. Meanwhile, various models have been proposed for (ii) including those whose design centers around a specific statistical model (e.g. Bayesian mixture model [108, 134], linear mixed model [124], random-effects model [125]) and those that are more general (e.g. CORE [135], trajectory clustering [136], mass distributed clustering [137]). Although such approaches produce more promising results, they are limited in that only a small number of computational methods can explore this information while many others requiring expression profiles as the input cannot.

Here we address a somewhat different question, namely *whether we can integrate the error information into the time-series expression profiles so that we can utilize a variety of computational models [14, 99, 138] that take the expression profiles as the required input without any modification while taking into account the advantage of using replicated data (especially for clustering methods e.g. mclust [139], som [140], micro-clustering [141], consensus clustering [101], etc.)*. In this aspect, the most straightforward approach to estimate time-series gene expression profiles is by computing the average expression levels over all replicates for each gene at each time-point (or condition in general). Of course, this approach does not properly take into account the variability in repeated measurements [105, 110]. Therefore, in an attempt to estimate

more robust expression profiles that integrate the error information from replicates, so-called ‘true’ expression profiles, we explore the error model [133] to estimate the ‘true’ mean expression value of a gene across all time-points and the concept of ‘relative difference’ driven by the theory of t-statistic [128, 131] to compute the difference between the ‘true’ mean expression value across all time-points and the mean expression value at each time-point. Those relative differences are then used to infer the ‘true’ expression profile of the gene.

### 3.1.2. The statistical model

In order to utilize a variety of computational models that take the expression profiles as the required input without any modification while taking into account the information of repeated measurements, we will estimate a more robust expression profile that integrate the error information from replicates. Let us assume that the ‘average’ time-series expression profile of gene  $i$  across  $T$  time-points with  $R_t$  replicates at each time-point can be generally represented as

$$g_i = \{\bar{g}_{it}\}_{t=1}^T, \bar{g}_{it} = \frac{1}{R_t} \sum_r g_{itr}$$

The subscripts  $i$ ,  $t$ ,  $r$  indicate the gene id, time, and replicate respectively. The procedure to estimate the ‘true’ expression profile consists of two main steps:

#### *i. Estimate the ‘true’ mean expression value of a gene across all time-points*

Utilizing the variance (error) of repeated measurements at each time-point  $\sigma_{it}$ , the error model weights the average expression values at each time-point when computing the mean expression value of the gene across all time-points [133]

$$\hat{g}_i = \frac{1}{\sum_t w_t} \sum_{t=1}^T w_t \bar{g}_{it} \text{ where } w_t = \begin{cases} \frac{1}{\sigma_{it}} & \text{if } \sigma_{it} \neq 0 \\ 1 & \text{if } \sigma_{it} = 0 \text{ or } R_t = 1 \end{cases}, \sigma_{it}^2 = \frac{1}{R_t - 1} \sum_r (g_{itr} - \bar{g}_{it})^2 \quad (1)$$

The variance of  $\hat{g}_i$  can be calculated in two ways: one is to propagate the errors  $\sigma_{it}$  and the other is from the scatter of  $\bar{g}_{it}$  around  $\hat{g}_i$

$$\sigma_p^2 = 1 / \sum_t w_t^2 \text{ or } \sigma_s^2 = \frac{1}{(T-1) \sum_t w_t} \sum_t w_t (\bar{g}_{it} - \hat{g}_i)^2 \quad (2)$$

The propagation of variance  $\sigma_p$  is based on the error estimation of each individual time-point, leading to bias and/or systematic uncertainties whereas the other  $\sigma_s$  has large fluctuation when the number of measurements is small although it is an unbiased measure. Statistically one can combine these two variances in estimation of the variance for  $\hat{g}_i$  [133]

$$\sigma_{\hat{g}_i} = \frac{\sigma_p + (T-1)\sigma_s}{T} \quad (3)$$

**ii. Estimate the relative difference between the ‘true’ mean expression value across all time-points and that at each time-point (one is replaced for the ‘true’ mean expression value)**

In order to infer the expression value at each time-point of a gene, we utilized the concept of ‘relative difference’ [128, 131] from the t-statistic to estimate its difference from the ‘true’ mean expression value of the gene. Let  $d_{it}$  represent the relative difference between the ‘true’ mean expression value across all time-points and the mean value at a specific time-point:

$$d_{it} = \frac{\bar{g}_{it} - \hat{g}_i}{s_t \sqrt{\frac{1}{R_t} + \frac{1}{T}}} \quad (4)$$

where  $s_t$  is the standard deviation of these two quantities

$$s_t = \sqrt{\frac{(R_t - 1)\sigma_{it}^2 + (T - 1)\sigma_{\hat{g}_i}^2}{R_t + T - 2}} \quad (5)$$

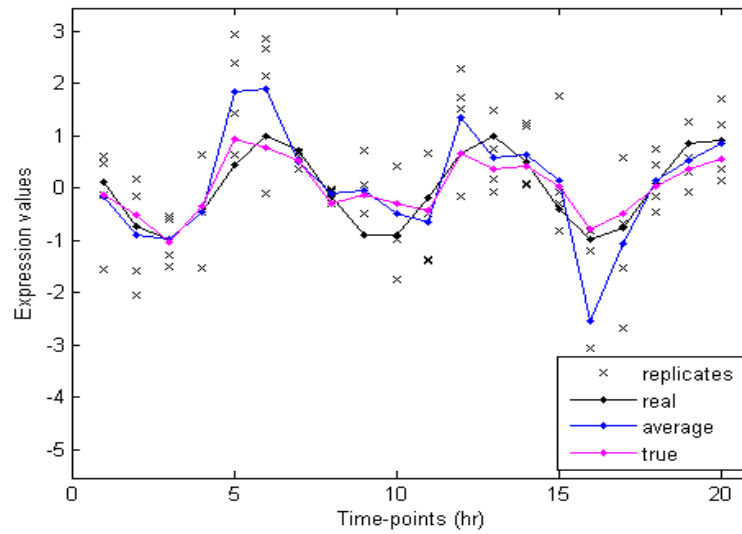
And thus, we propose a more accurate estimation of the average expression value at a specific time-point as follows

$$\bar{g}'_{it} = \hat{g}_i + d_{it} \quad (6)$$

As we rationalized the importance of microarray replicates in the background section, we hypothesize that the expression profiles would be more robust if there is some statistical approach that integrates the error information from replicates into the estimation. For average expression profiles, the expression value at a specific time-point is  $\bar{g}_{it} = \bar{g}_i + (\bar{g}_{it} - \bar{g}_i)$  where  $\bar{g}_i = \frac{1}{T} \sum_t \bar{g}_{it}$ . In a similar manner we obtain formula (6) in a

way that integrates the error information into two parts of the formula; the  $\hat{g}_i$  part is the ‘true’ mean expression value across all time-points and the latter part  $d_{it}$  is the relative difference between the ‘true’ mean expression across all time-points and the one at that specific time-point. Figure 3.1 compares the ‘true’ expression profile to the average one. Its effectiveness will be further demonstrated with the clustering performance on synthetic and real data.

**Figure 3.1:** The ‘true’ expression profiles are more robust than the average ones. ‘real’ is the actual profile from simulated data without noise. ‘replicates’ are obtained when noise is added to the actual value. The average profile is showed to be more deviated from the actual profile than the ‘true’ profile.



### 3.1.3. The clustering effectiveness of the ‘true’ expression profiles

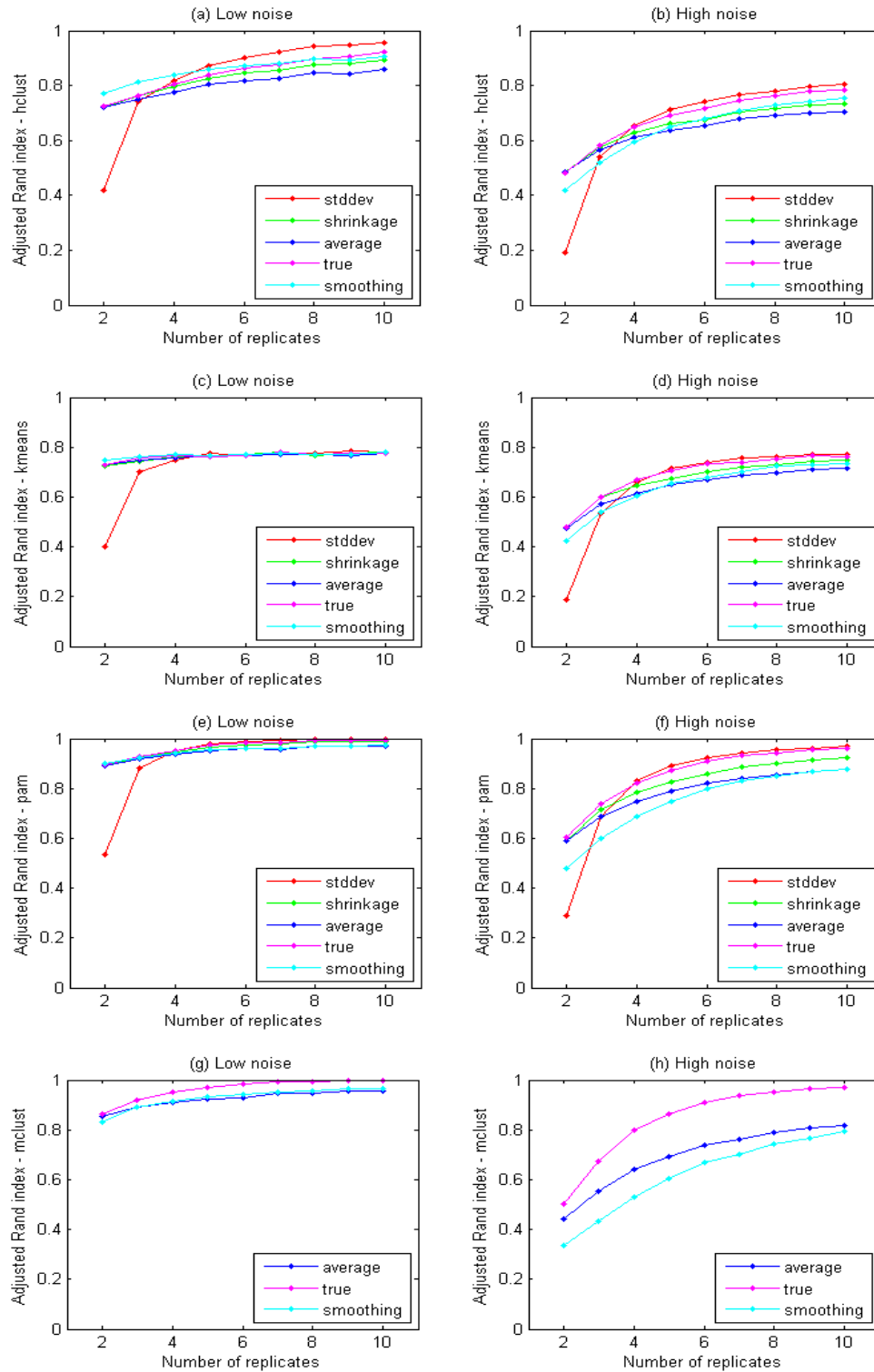
To evaluate the effectiveness of the ‘true’ expression profile compared to using the ‘average’ profile, we use the synthetic data with known class structure as described earlier. As in previous studies [110], we also assess the effect of the number of replicates on cluster quality. Each synthetic data contains 20 time-points with  $r$  replicates ( $r = 2, 3, 4, 5, 6, 7, 8, 9, 10$ ) at each time-point and two different levels of noise (low and high). In addition to comparing the clustering performance using the ‘true’ profiles with the average profile, we also compare with several other methods that take into account error information from replicated data. Specifically, we measure cluster quality when using two typical similarity distance metrics which include the error information, namely the standard deviation (SD)-weighted correlation coefficient [133] and the shrinkage correlation coefficient [110]. Since our model generates expression profiles which are applicable to any clustering method, we also tested an alternative method that uses a cubic spline to infer expression profiles which account for repeated measurements, so-called ‘smoothing’ profiles. For each gene, we establish two vectors – one consists of all



replicates and another contains corresponding time-points. They are then input into function ‘smooth.spline’ in stats R package [142]; other parameters (e.g. the degree of freedom, smoothing parameters) are optimized from an internal ‘generalized’ cross-validation process provided by the tool. After that, the expression value at each time-point is inferred to create the ‘smoothing’ profile for the gene (using function ‘predict’ in R). Subsequently, the Pearson correlation coefficient is applied to estimate the similarity distance between two genes with the average profiles, the ‘true’ ones, and the ‘smoothing’ ones. After obtaining the pairwise distance matrix, we apply three popular clustering methods: hierarchical clustering (with average linkage option, available in MATLAB), partitional clustering (k-means [143], pam [144]), and model-based clustering (mclust [139]) to cluster the data into six clusters. In order to assess the clustering performance, we use the adjusted Rand index [105, 145] which is a statistic that measures the extent of concurrence between the clustering results and the underlying known class structure. The larger the Rand index is, the higher the agreement between clustering results and prior knowledge of class structure i.e. better performance.

**Figure 3.2:** The performance of typical clustering methods on different error-measurement integrated approaches. ‘stddev’ represents for the clustering performance on synthetic data using the approach with the SD-weighted correlation coefficient metric; similarly, ‘shrinkage’ is for the approach with the shrinkage correlation coefficient metric; ‘average’ is for the clustering performance on average profiles; ‘true’ is for that on ‘true’ profiles; and ‘smoothing’ is for that when using method ‘spline’ to infer the expression profiles and then clustering. The horizontal axis shows the corresponding number of replicates in the dataset while the vertical axis demonstrates the clustering

performance of the corresponding approach (the higher the better). Results are the average of clustering accuracies over 1000 randomly generated synthetic datasets.



**Figure 3.2** depicts the clustering performance when using our proposed model compared to other approaches. We evaluate the average of 1000 randomly generated synthetic data sets. Figure 3.2a and 3.2b show the comparisons using hierarchical clustering. For the low-noise level (Figure 3.2a), the clustering performance using the ‘true’ profiles is slightly worse than that when using the SD-weighted correlation coefficient metric or ‘smoothing’ profiles. However, it is still much better than that when using the average profiles. For the high-noise level, it is comparable to the best achievable by any other method (Figure 3.2b). When other clustering methods are used (e.g. kmeans – Figure 3.2c & 3.2d, pam – Figures 3.2e & 3.2f, mclust – Figures 3.2g & 3.2h), the clustering performance on the ‘true’ expression profiles is always superior, or comparable, to any other approach on both low and high noise data, and far better than that of the average profiles in high noise data. Additionally, when datasets are sampled with few time-points and/or few replicates, the alternative method that uses spline to infer expression profiles may be less advantageous than the proposed approach since it may be failed in detecting proper parameters for ‘spline’ profiles to recover the actual expression profiles.

### 3.2. Consensus clustering

Instead of clustering the entire dataset, we explore the hypothesis that the more clusterable the data is the more biologically relevant it is and utilize the concepts of consensus clustering to identify, within a set of differentially expressed genes, a subset of genes that are either highly co-expressed or highly non-coexpressed with the hope of extracting a more biologically relevant subset of genes. The main problem to be addressed can be defined as follows. Given a set of  $n$  objects  $G = \{g_i\}_{i=1}^n$ , with each

described by a list of  $d$  numerical attributes  $g_i = \{g_{i1}, g_{i2}, \dots, g_{id}\}$ ,  $g_{ij} \in R$ ,  $i = \overline{1..n}$ ,  $j = \overline{1..d}$ ,

we wish to pick out a ‘clusterable’ subset of objects  $G' \subset G$  with a confidence level  $\delta\%$ .

The term ‘clusterable’ subset is used in the sense that

$$\forall g_i, g_j \in G' \Rightarrow \left\{ \exists C_q \ni P\left[(g_i \wedge g_j) \in C_q\right] \geq \delta \vee \forall C_q, P\left[(g_i \wedge g_j) \in C_q\right] \leq 1 - \delta \right\}$$

where  $C_q$  denotes a, yet to be determined, cluster and  $P$  is the probability that the two objects belong to the same cluster.

### 3.2.1. Background

The traditional way for performing clustering analyses is using one clustering method to group all genes in a dataset into a number of clusters given a pre-defined metric of similarity. Those genes that belong to one cluster can be considered as co-expressed and those that belong to different clusters are non-coexpressed. However, it is widely accepted that a number of critical problems associated with clustering remain open: (i) it is not immediately obvious what the optimal number of clusters is [146] and it has been recognized that it is difficult to develop a systematic and generic method for addressing such a question [147-154]. Approaches such as DBSCAN [155] showed great promise but the issue associated with high-dimensional data still remains [156]; (ii) every clustering method relies on the definition of an appropriate distance metric such as Euclidean, Pearson correlation, Manhattan etc. [157-159], however an algorithm’s performance is highly sensitive to the selection of the metric, particularly for objects lying at the boundaries between clusters; (iii) all clustering methods express their own bias and assumptions, and their performance depends highly on the properties of the input dataset [156]. Therefore, alternative methods have been proposed to attempt to reduce the

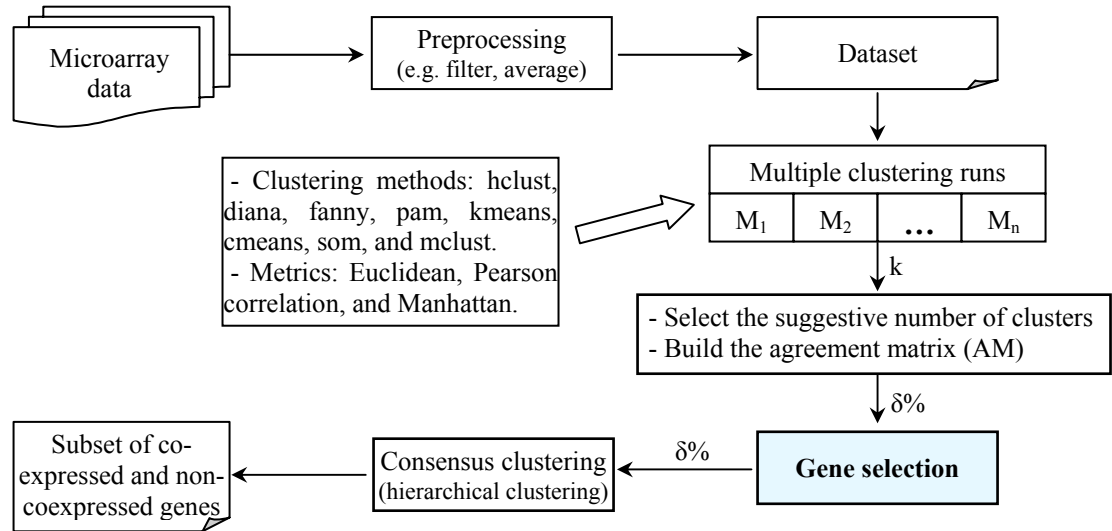
bias by combining two or more clustering algorithms [160, 161] or by incorporating with prior domain-specific knowledge to guide the clustering process. In the context of microarray analysis it may include gene ontology [162, 163], gene annotation [157], gene function [164] etc.; finally (iv) the significance of a cluster and/or the probability that two genes may belong to one cluster or two different clusters is also an issue [161, 165, 166].

In order to overcome some of the aforementioned complications simultaneously, the concept ‘ensemble’ or ‘consensus’ was introduced into the clustering literature [167]. By averaging, in some way, the results of multiple runs, one can estimate an ‘agreement matrix’ (AM) and infer a better proxy of what a more ‘correct’ result ought to look like. A number of approaches have been proposed; for example Monti *et al.* [149] applied one clustering algorithm on multiple sub-sampled datasets without replacement based on the original data set whereas Grotkjaer *et al.* [168] used different random initializations of a single clustering method to generate multiple results from which the agreement matrix was built to yield the final clustering assignment. Although such approaches offer definite advantages, they still express a strong bias for a given method and/or metric. Consequently, an alternative strategy with a meta-clustering step is applied on the agreement matrix as the distance matrix to reach the final result. Different studies chose different clustering methods based on those frequently used in the literature for the first level. In the meta-level, although it is based on a single method, it is still not evident which clustering method should be used and different studies selected different methods, e.g. simulated annealing [169, 170], mapping by Jaccard index [171], expectation maximization algorithm [172].

However, we here do not mainly focus on solving the problems of clustering. Instead we will explore the concept of consensus clustering to identify, within a set of differentially expressed genes, a subset of genes that are either highly co-expressed or highly non-coexpressed with the hypothesis that this subset would serve as a better starting point for further analysis, such as coregulation. A number of core clustering methods, supported by R packages e.g. hierarchical clustering (hclust), divisive analysis clustering (diana), fuzzy analysis clustering (fanny), partitioning around medoid (pam), k-means (kmeans), fuzzy c-means (cmeans), self-organizing map (som), and model-based clustering (mclust) will be employed in the first-level [139, 140, 144, 173-175]. Additionally, in order to overcome the limitations of using a single distance metric, we explore different metrics (Euclidean, Pearson correlation, and Manhattan) that have already been established [176]. The sensitivity of the AM was also examined as a function of the input number of clusters to find a suggestive number of clusters that best describes a particular dataset. The result of the first-level analysis is a systematic framework for eliminating all genes that cannot be clearly characterized as either coexpressed or non-coexpressed with others in the ongoing selected subset. Subsequently, an agglomerative hierarchical clustering approach is applied to cluster the selected subset using the agreement metric information as the similarity measure (Figure 3.3).

The problem is quite general and applicable to a large family of problems. However, to be consistent with the specific problem at hand (microarray data), objects will refer to genes, or better yet probesets, with expression levels measured in different experimental conditions or time-points, and clusters are groups of genes sharing similar expression profiles. Thus, genes that belong to one cluster are considered to be coexpressed and

genes that belong to different clusters are considered as non-coexpressed with a confidence level.



**Figure 3.3:** Schematic overview of microarray data analysis using multiple clustering runs to select a ‘clusterable’ subset – the subset which contains genes that are either highly coexpressed or non-coexpressed with a confidence level  $\delta\%$ . The preprocessing step (filtered by fold-change, ANOVA [100, 177], SAM [131], EDGE[132]) removes as many as possible genes that are not significantly differentially expressed across conditions or time-points. Data with repeated measurements can be averaged before clustering [105]. Each clustering method needs an input number of clusters  $k$  as the required input parameter; therefore we examine the agreement matrix (AM) for a number of different  $k$  and try to select one as a suggestive number of clusters for the dataset. Then, the final AM is built and pass through the process of gene selection which eliminates all genes that have at least one ‘inconsistent’ value with some other gene.  $\delta$  is the threshold to say whether the agreement level of two genes belong to one cluster ( $\geq \delta$ )

or two clusters ( $\leq 1-\delta$ ) is consistent or not. The last step is dividing the selected subset into a number of patterns with the agreement threshold  $\delta$  based on the remainder of the AM as the input distance matrix.

### 3.2.2. The agreement matrix

The *agreement matrix*  $M$  quantifies the frequency with which two genes belong to the same cluster (Figure 3.4). If  $N$  clustering runs are performed on the data, each entry  $M_{ij}$  (termed ‘agreement level’) shows the fraction of clustering times two genes are assigned to the same cluster. The AM entries are defined as:

$$M_{ij} = \frac{1}{N} \sum_{h=1}^N M^{(h)}(g_i, g_j)$$

$$\text{where } M^{(h)}(g_i, g_j) = \begin{cases} 1 & \text{if } g_i \text{ and } g_j \text{ are clustered together when running method } M^{(h)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and  $N$  is the number of clustering runs performed with either different methods or distance metrics. In our work, we are using hclust, diana, fanny, and pam with Pearson correlation and Manhattan metric, kmeans, cmeans, som, and mclust with Euclidean metric as the core clustering methods [139, 140, 144, 173-175].

	$M^1$	$M^2$	$M^3$	$M^4$		$g_1$	$g_2$	$g_3$	$g_4$	$g_5$	$g_6$	$g_7$
$g_1$	1	2	1	3	$g_1$	1.0	0.0	1.0	0.0	0.0	1.0	0.5
$g_2$	2	1	3	2	$g_2$		1.0	0.0	0.5	0.75	0.0	0.25
$g_3$	1	2	1	3	$g_3$			1.0	0.0	0.0	1.0	0.5
$g_4$	3	1	3	1	$g_4$				1.0	0.25	0.0	0.0
$g_5$	2	1	2	2	$g_5$					1.0	0.0	0.25
$g_6$	1	2	1	3	$g_6$						1.0	0.5
$g_7$	1	3	1	2	$g_7$							1.0



**Figure 3.4:** An example of the agreement matrix (right). The left is the results from  $N$  clustering runs ( $N = 4$  in this example, represented by  $M^1 \dots M^4$ ) with  $k = 3$  as the input number of clusters on  $n$  genes ( $n = 7$ , represented by  $g_1 \dots g_7$ ). The right shows the corresponding agreement matrix that each entry  $M_{ij}$  is the frequency of gene  $i$  and gene  $j$  grouped into the same cluster by  $M^1 \dots M^4$ .

### 3.2.3. The optimal suggestive number of clusters

The evaluation of the AM entries requires the identification of a ‘suggestive’ number of clusters since, as mentioned earlier, clustering results are highly dependent upon this input value. Motivated by the work of Monti et al. [149], in order to identify a robust estimate for the suggestive number of clusters of a given dataset, we examined the distribution of the agreement matrix entries as a function of the number of clusters ( $k$ ). By definition the AM entries vary from zero to one whereas the number of entries falling into the zero-end region always increases as the input number of clusters  $k$  increases. Ideally, and assuming that the data in question do possess a definite underlying structure there should exist an ‘optimal’ number of clusters ( $k^*$ ). Thus, one would expect that as  $k$  varies from 2 to  $k^*$ , the rate at which the AM entries shift to the zero-end is faster than that when  $k > k^*$ . The rationale behind this hypothesis is that when the optimal number of clusters is reached, each clustering method individually makes a more appropriate cluster assignment to objects in the dataset and thus the cluster assignments from various clustering methods are more common. After that, the reassignment rate is reduced, making the agreement levels between objects change lesser and lesser. As a result, we would expect the distribution of the AM entries to change rapidly early on and eventually the rate of change would drop as  $k > k^*$ .

We tested the hypothesis by observing the histogram of the AM entries [146] as the number of putative clusters  $k$  changes i.e.  $k$  is varied from 2 to some number  $K$  and successive AM matrices are built. The corresponding distribution of the AM entries is represented by an empirical cumulative distribution function [149] (Figure 3.5)

$$CDF_k(x) = \frac{\sum_{i < j} 1 \text{ if } M_{ij} < x}{\frac{1}{2}n(n-1)}, i, j \in 1..n \quad (2)$$

The histogram-based area under the CDF curve (AUC) corresponding to each value of  $k$  is evaluated by

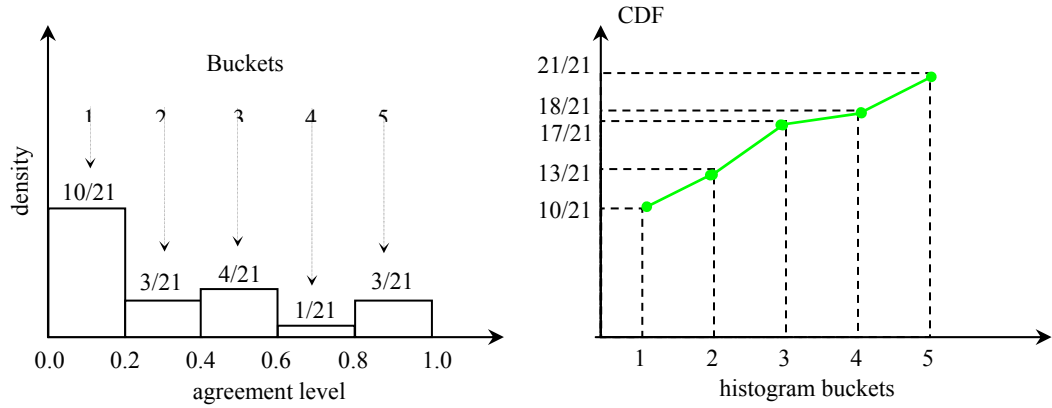
$$AUC_k = \sum_l (x_l - x_{l-1}) CDF_k(x_l), x_l = l / B, l \in 1..B \quad (3)$$

where  $B$  is the number of buckets used to construct the histogram or numerically define the CDF. As a result, the change of the distribution of AM entries is reflected by the changes of the  $AUC_k$ . The hypothesis earlier stated effectively is to look at the rate at which the successive distributions change when  $k$  increases in order to identify a putative number of clusters. Therefore, and in order to evaluate a more unbiased metric for determining the rate of change of the successive CDFs, we made use of the gap statistic metric [152, 153] and redefined it as:

$$Gap_k = mean \left\{ \Delta_{k_i} \right\}_{k_i=3}^K - \Delta_k, \Delta_k = |AUC_k - AUC_{k-1}| \quad (4)$$

Due to the high computational requirement, we used the mean of all  $\Delta_k$  (excluding  $\Delta_2$ ) instead of calculating the expected value for each  $\Delta_k$  in the first part of (4) from uniform data as originally suggested. Because of the faster shift to the zero-end region of AM entries early on, the rate of change of  $\Delta_k$  based on the AUCs is larger at the beginning and decreases gradually. As a result, the gap quantity ‘Gap<sub>k</sub>’ varies from negative to positive. We select the  $k$  value at which Gap<sub>k+1</sub> becomes positive to be the suggestive number of

clusters for the dataset since the distribution of the AM entries seems to be stabilized from that value. Besides, with the above definition the mean value of all  $\Delta_k$  will be highly dependent on the selection of value  $K$ . However, when  $k$  is over some value, the change of the AM distribution is trivial just because of the nature of the clustering methods. Consequently, value  $K$  must be selected to be appropriate with the changing amount  $\Delta_k$  of the AUCs. The key point here is to select the right ‘elbow’ of the curve of  $AUC_k$ . Therefore, we suggest an empirical default value  $K = \left\lceil \sqrt[4]{nd^2} \right\rceil$  which can be considered as a balance between the number of objects, object attributes and the significant change in the distribution of AM entries as well as the expected number of clusters in the dataset.



**Figure 3.5:** Histogram of AM entries (left) and the corresponding CDF curve (right) from the AM in Figure 3.4. Assume that five buckets ( $B=5$ ) are used to build the histogram; each represents the proportion of AM entries that fall into segment  $[(l-1)/B, l/B)$ ,  $l=1..5$ ; the last segment also includes all entries with value one. The CDF curve is constructed based on the histogram and thus the horizontal axis is the axis of the agreement level as well as the histogram buckets.

### 3.2.4. Clusterable data

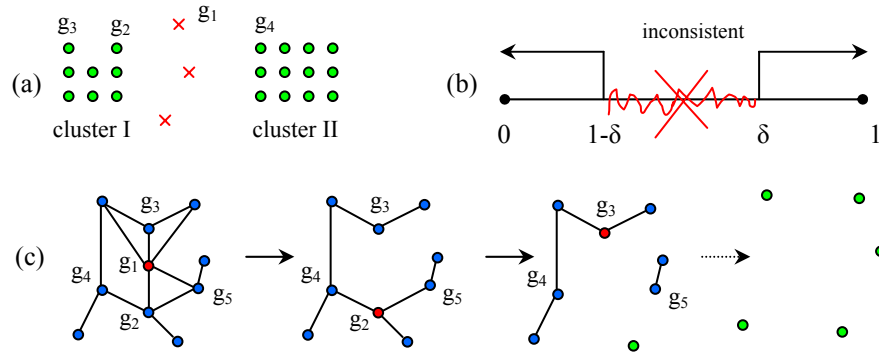
The analysis of the agreement matrix results reflects the expected relationship between two genes, i.e., the probability of belonging or not to the same cluster. As such entries associated with genes at the ‘hypothetical core of a cluster structure will be consistently grouped together over multiple runs. This should be manifested by high corresponding values in the AM, whereas genes belonging to the ‘hypothetical’ core of clearly distinct clusters should be associated with consistently low AM entries. On the contrary, genes around the ‘hypothetical’ boundary between two clusters would be very sensitive to changes in the clustering method. As a result, a gene at the cluster boundary should be characterized by relatively moderate agreement levels in relation to other genes (Figure 3.6a). Thus, our hypothesis is that eliminating genes associated with moderate AM entries would create a more ‘clusterable’ subset. It also should be emphasized that this approach is not aimed at identifying and eliminating ‘outliers’ and thus this is not an outlier detection procedure. We simply hypothesize on the potential properties of a more clusterable subset of objects.

In order to quantify the aforementioned observation, we define an AM entry as an ‘inconsistent’ entry if its value is within the interval  $1 - \delta < M_{ij} < \delta$ , where  $\delta$  expresses a user-defined confidence level (Figure 3.6b). The AM is now transformed into an adjacency matrix where consistent pairs of genes i.e. genes that are frequently assigned to the same or different cluster(s) receive a value of ‘0’ and inconsistent entries are assigned value ‘1’. The adjacency matrix is then converted to an ‘inconsistent’ graph with nodes indicating genes and edges connecting two nodes (genes) representing the cluster assignments between those two genes over multiple clustering runs are unclear. The

problem now is removing a number of vertices so that the resulting graph is completely disconnected [178]. We called an inconsistent rank of a vertex is the order of that vertex, i.e. the number of edges at that vertex

$$i\_rank_i = \sum_j 1 \text{ if } (1-\delta) < M_{ij} \text{ and } M_{ij} < \delta, j = 1..n \quad (5)$$

Therefore, vertices with many edges or genes with many inconsistent AM entries will get high inconsistent ranks; the ones with highest inconsistent rank and all of its edges will be removed first. The inconsistent rank for each vertex is then recalculated and the step is repeated (Figure 3.6c). If there are some equivalent inconsistent ranks, the removed vertex can be chosen to be the one with the highest original inconsistent rank or randomly (e.g. vertex with the smallest index in our implementation, creating the consistency of removed genes over different running times). The routine is repeated until the ‘inconsistent’ graph becomes completely disconnected i.e. the selected subset contains no gene with an ambiguous cluster assignment with other exiting genes with a given confidence level.



**Figure 3.6:** The gene selection process. (a) Genes at boundaries or outliers between clusters will have many moderate agreement levels; g<sub>2</sub> and g<sub>3</sub> in cluster I will have a high agreement level whereas g<sub>2</sub> and g<sub>4</sub> have a low agreement level; g<sub>1</sub> can belong to either

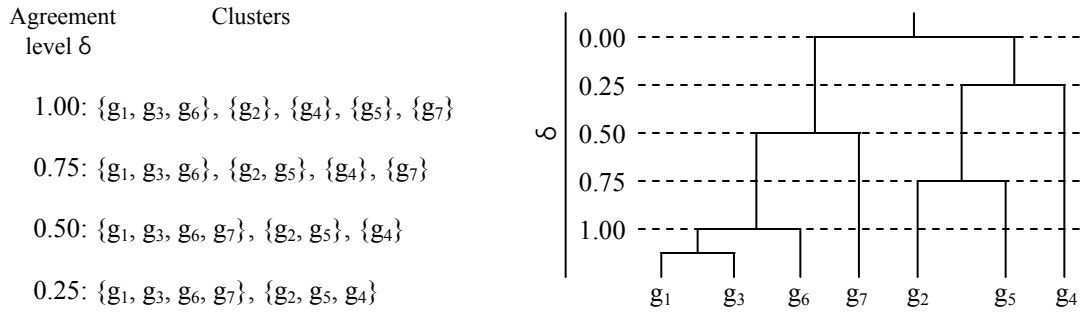
cluster I or cluster II among different clustering running times, causing agreement levels between  $g_1$  and other genes e.g.  $g_2, g_3, g_4$  are moderate. (b) The inconsistent region of agreement levels. (c) The process of disconnecting the inconsistent graph;  $g_1$  is selected to remove since it has the highest inconsistent rank;  $g_2$  has the same inconsistent rank with  $g_4$  but it is still removed next since it has a higher original inconsistent rank than  $g_4$ ; then  $g_3, g_4$ , and  $g_5$  are eliminated respectively (genes with green color will be remained; red ones are removed; blue ones are being examined).

### 3.2.5. Consensus clustering

Without dependence on any other parameter besides an agreement threshold to form clusters, hierarchical clustering was selected to perform the final clustering task. The algorithm starts with every gene as a cluster and tries to group two clusters into a new one at each iteration. Any pair of genes belonging to that new cluster needs to have an agreement level more than or equal to  $\delta$  ( $\delta$ -rule). A new cluster is formed by joining two clusters  $C_p$  and  $C_q$  whose total agreement of all pairs of genes in these two clusters

$$total\_agreement(C_p \wedge C_q) = \sum_{\substack{k=\{genes\ in\ C_p\} \\ l=\{genes\ in\ C_q\}}} M_{kl} \quad (6)$$

is maximal. This selection assures that large clusters are given priority to join together since the total agreement between cluster  $C$  and a large one will be greater than that between  $C$  and a smaller one (Figure 3.7). Besides, although new clusters can be formed with the  $\delta$ -rule, we still favor those which contain genes more likely to be clustered together. Therefore, we introduce a cooling rate to replace the role for  $\delta$ . As a result, instead of satisfying the  $\delta$ -rule, any pair of two genes in a new cluster now needs to satisfy the  $\theta$ -rule (i.e. their agreement level will be greater than or equal to  $\theta$ ) and  $\theta$  decreases slowly from 1.0 to  $\delta$ .



**Figure 3.7:** Illustration of the consensus clustering on the agreement matrix in Figure 3.4.

(a) List of clusters corresponding to different agreement thresholds. (b) A hierarchical dendrogram to visually show the way of forming clusters corresponding to  $\delta$ . This example also demonstrates the effects of the total agreement and/or the cooling rate  $\theta$ : the algorithm always guarantees that large clusters are taken priority and/or that clusters with more pairs of high agreement genes are joined together first, e.g. in the case of  $\delta = 0.50$ ,  $g_2$  and  $g_5$  (0.75) are joined first and  $g_4$  cannot join the group although the agreement level between  $g_2$  and  $g_4$  (0.5) satisfies the  $\delta$ -rule; this reduces the effect of breaking down the pattern or non-optimal patterns are formed e.g.  $(g_2 \wedge g_5)$  compared to  $(g_2 \wedge g_4)$ .

The algorithm produces a list of clusters in which any two genes belonging to one cluster always have an agreement level greater than or equal to  $\delta$ . Although  $\delta$  is a measure of the frequency with which two genes can be found in the same cluster over a variety of clustering runs, it can be also be considered as the confidence that the two genes are coexpressed since  $\delta$ , by construction, aims at eliminating method-specific biases and assumptions. Furthermore, the gene selection step assures that the inconsistencies in the AM are minimized since the relationship between any two genes is evaluated with a

confidence level. Therefore, genes that belong to different clusters can also be considered as highly non-coexpressed with a confidence level  $\delta$ .

Additionally, we also provide an optional procedure to exclude trivial clusters formed due to the nature of clustering methods. Each cluster  $C$  is assigned with a simple hypothetical quantity called ‘cluster significance’ which represents how large the cluster is and how coexpressed the genes in the cluster are. To select significant clusters, we then estimate the distribution of cluster significance on random data and compute the p-value for each cluster  $C$  above. The dataset is randomly resampled (permutation plus convex-hull [165]) a number of times ( $n_r$ ), for each of which the entire process starting from building the AM with  $k^*$  selected above to the consensus clustering step is done and random-resulting clusters are returned. Subsequently, the procedure estimates the cluster significance for these random clusters and builds up a distribution of cluster significance. The cluster significance of a cluster is defined as its size times its homogeneity as mentioned above; random clusters can be in large-size depending on the input number of clusters but these clusters contain arbitrarily objects (genes) and thus their homogeneity will not be large, and thus the cluster significance remains trivial. As a result, the number of random clusters with greater values of cluster significance than that of a selected cluster  $C$  over the total number of random-resulting clusters in  $n_r$  times resampling will be considered as the p-value of cluster  $C$  for selection.

### **3.2.6. Method evaluation**

To assess our approach for finding highly coexpressed and non-coexpressed genes, we analyzed a number of data sets from the open literature. Specifically, we used the synthetic data to evaluate fundamental concepts of the algorithm since the structure is



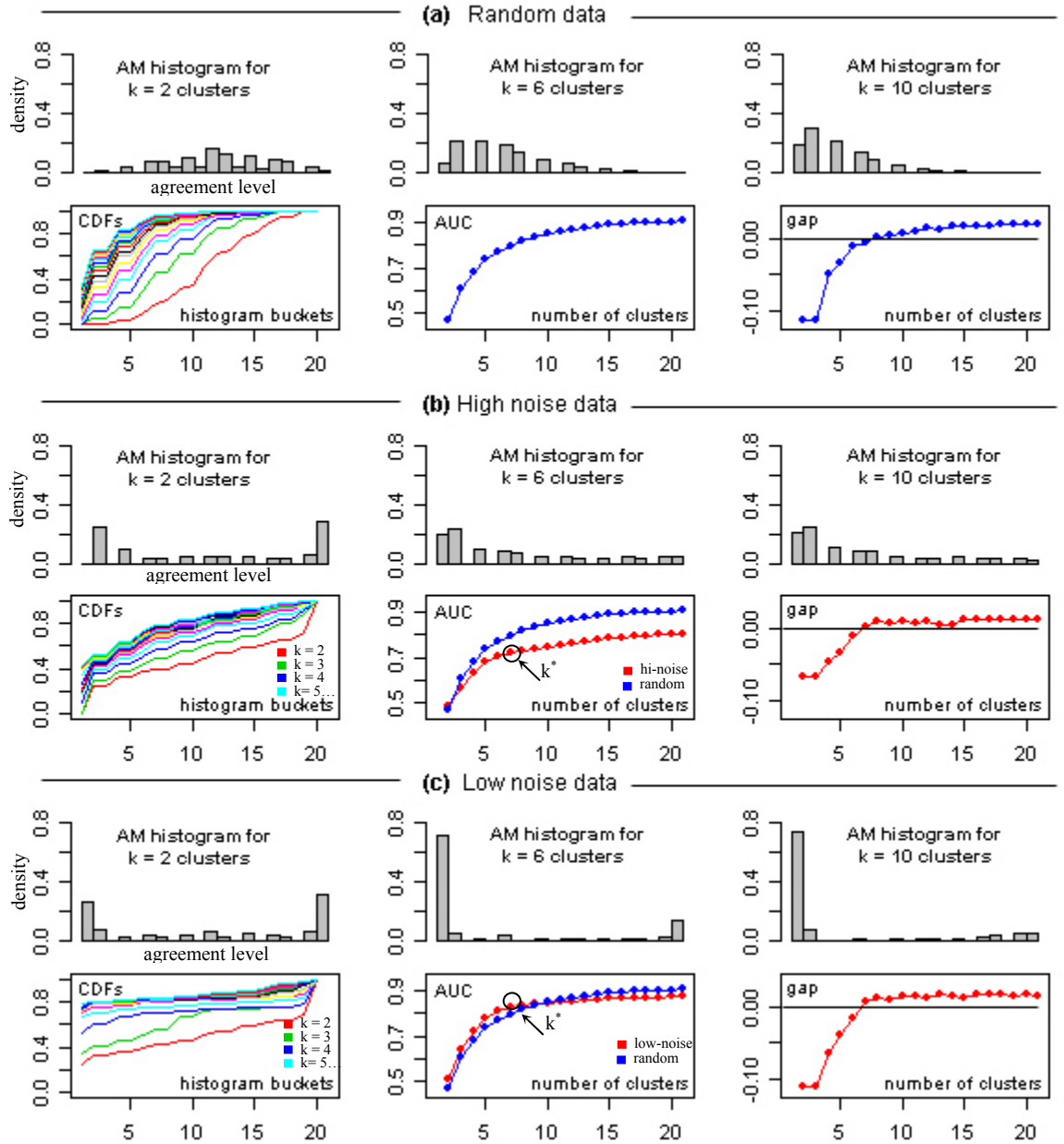
precisely known. We therefore utilized five low-noise and five high-noise 20-attribute sine-format synthetic datasets from [105]. To demonstrate the effectiveness of the approach, we illustrated the accuracy and the clusterability on the selected subset as well as the properties on the removed domain from the synthetic datasets using Rand index [179, 180] and Friedman-Rafsky test [181, 182]. Besides that, in order to visualize the effect of different cut-off agreement levels on the selection results, we used five two-dimension (2D) testing sets from [107]. The capability to find out a suggestive number of clusters for a dataset is also demonstrated using these synthetic datasets.

### ***Distribution of the AM entries***

In order to examine the properties of the AM, we made use of the available synthetic datasets where we could obtain random, and structured, high- and low-noise data. Random data are generated through resampling (permutation plus convex-hull [165]) synthetic datasets, each with 10 times. The AM histogram, CDF, AUC and gap curves were built independently for each dataset and then the average ones are made for each data type to have a consensus view (Figure 3.8).

**Figure 3.8:** Examining the agreement matrix. (a) Average histograms of the AM on 100 random datasets (random resampling 10 times from 10 synthetic datasets) for several input numbers of clusters. For each given input number of clusters  $k$ , we have a corresponding histogram of the AM entries for a specific dataset with agreement level index  $l = \{1..B = \sqrt{\|dataset\|}\}$  ( $B = 20$  buckets in this case); the height of each bucket  $l$  is proportional to the number of  $M_{ij}$  falling into the segment  $[(l-1)/B, l/B]$ ; repeat on 100 random datasets and take the average histogram for each corresponding  $k$ . (b) Histogram of the AM (top) and the CDF lines, AUC curve, as well as Gap curve on high-noise

synthetic sets; repeat on five synthetic high-noise sets and take the corresponding average. (c) Histogram of the AM (top) and the CDF lines, AUC curve, as well as Gap curve on low-noise synthetic sets; repeat on five synthetic low-noise sets and take the corresponding average.  $k = 6$  is the right number of clusters for these datasets. On the AUC graphs, the blue one is the AUC curve on random datasets obtained from (a).



The expected change of the histograms as  $k$  increases is manifested by the shift of the distribution of AM entries observed with both high- and low noise data (Figure 3.8b & 3.8c panels – top row). In both cases, AM entries shift faster to the zero-end when  $k$  is less than  $k^*$  ( $k^* = 6$  in this case). Therefore, compared to the random data, we observed that the synthetic data (i) produce distributions that are not normal, and (ii) beyond the hypothetical optimal  $k^*$  the distributions changes at a slower rate. Besides, the random AUC curve is also drawn to be compared to the non-random AUC curves.

### *Accuracy in predicting a suggestive number of clusters*

A most critical parameter characterizing the performance of this, and any clustering, approach is related to the selection of an appropriate suggestive number of clusters  $k$ . The results on the synthetic data here provide strong evidence for the method, suggesting that this could be used as a reasonable starting point (Table 3.1). However, one could attempt to interpret the AUC curve to suggest alternatives but for consistency purposes, in all our studies here, we made use of the Gap-based heuristic for estimating the putative value for  $k^*$  as showed below.

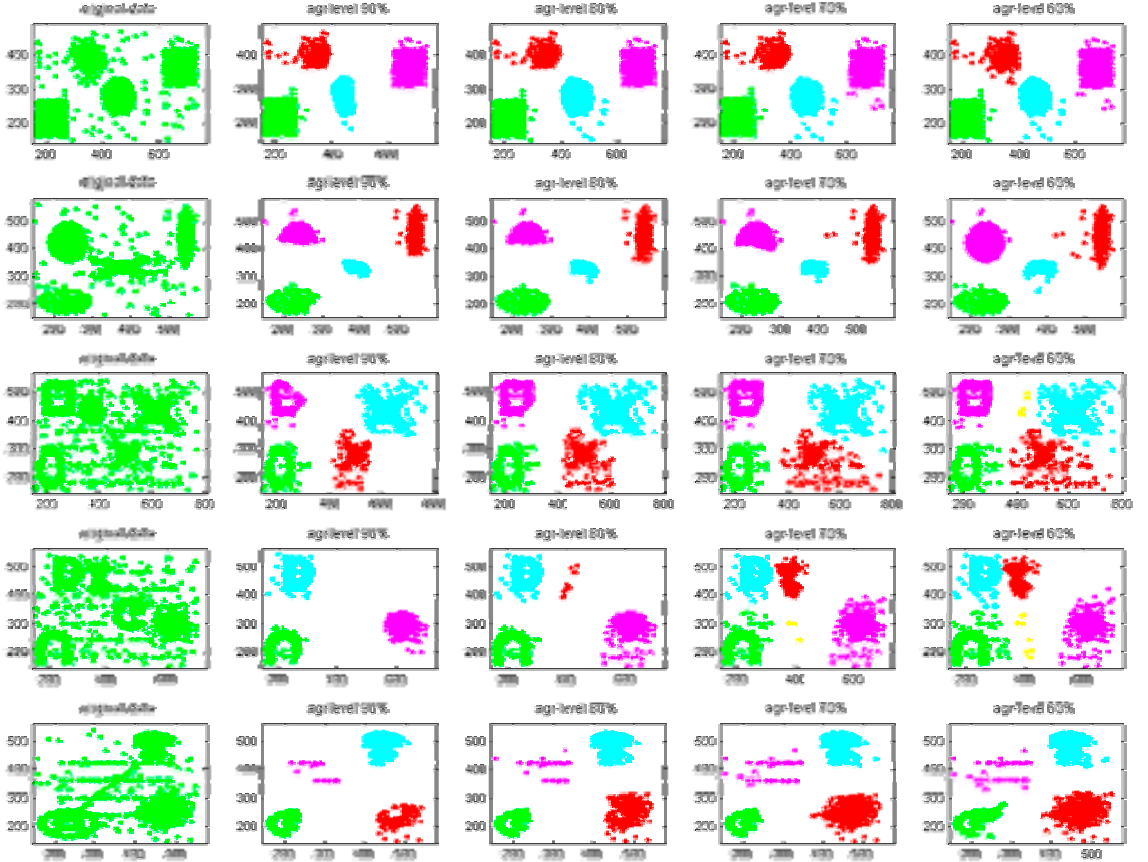
**Table 3.1:** Prediction the number of clusters by the process automatically

Datasets	2D synthetic		synthetic data				Real data	
	true	suggestive	low-noise		high-noise		Sporulation suggestive	LPS suggestive
			true	sugg.	true	sugg.		
Set 1	4	4	6	6	6	6		
Set 2	4	5	6	6	6	7*		
Set 3	5	5	6	6	6	6	7	6
Set 4	5	5	6	6	6	6		
Set 5	5	5	6	6	6	6		

\* implies that the suggestive numbers of clusters are suitable even though the true ones are different.

### *The impact of the confidence level $\delta$*

The next critical user-defined parameter in the overall scheme is the agreement threshold or confidence level  $\delta$  as it affects the selection and clustering steps. We examined five 2D instances of the synthetic data sets [107] with varying levels of noise and object selection is performed for different agreement thresholds  $\delta$ . The selected clusters are kept intact without applying the trivial-cluster removal procedure. At high confidence levels, the majority of points lying in the boundaries between core clusters are eliminated and possibly core clusters are partitioned further due to the strict requirements for membership to the same cluster. As  $\delta$  decreases, clusters get bigger but their boundaries appear to become fuzzier.



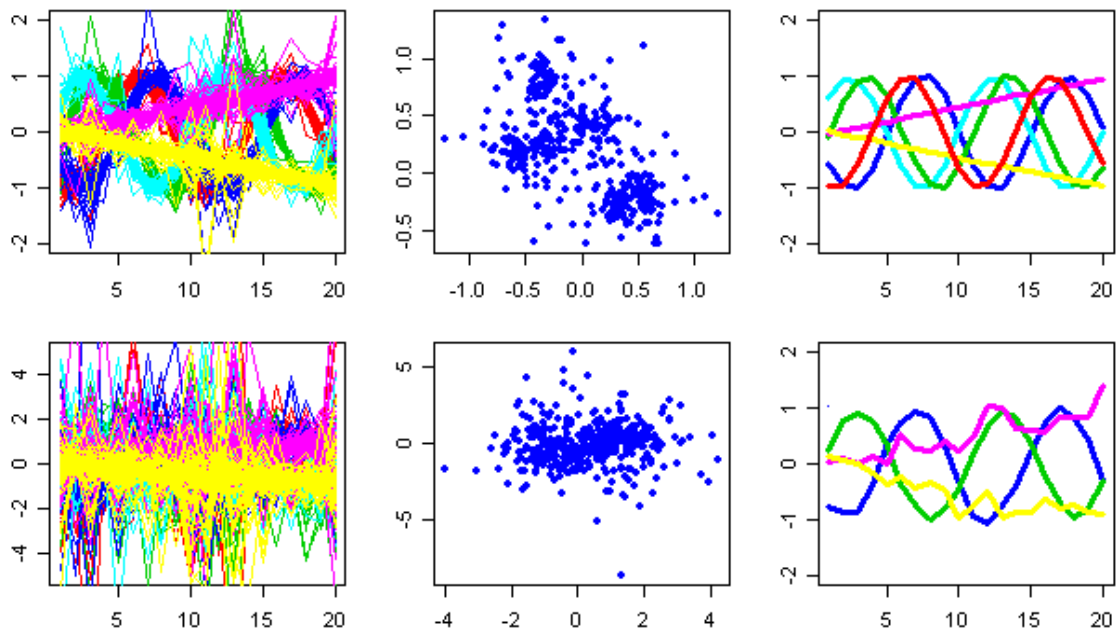
**Figure 3.9:** Illustration the selection and clustering as well as the effect of different confidence levels  $\delta$ . The objects at the boundaries and the outliers between clusters are eliminated under different viewpoints of clustering methods and differently used metrics. As the agreement threshold  $\delta$  decreases, selected clusters become bigger, some more ‘noise’ is added, and thus the confidence of coexpressed and non-coexpressed also reduces in the case of gene expression data.

### ***Consistency and accuracy of clustering and selection results***

To assess the accuracy of the selection and the quality of clustering, we applied the approach on synthetic datasets with a known class-structure distribution of each object. Ten downloaded, high-dimensional, synthetic datasets, 5 low- and 5 high-noise with log-transformation [121] were used for this purpose. Since the question we originally posed was whether selected objects are either highly similar or non-similar to each other (or highly coexpressed and non-coexpressed in the context of gene expression data), we do not need to classify all objects into their correct class structure. However, we need to identify a smaller set of objects for which we would be confident that the correct assignment can be made. A brief look on how the data look like and what our approach picked out is presented in Figure 3.10. To evaluate the accuracy we used the original Rand index [179, 180] to estimate the correctness of the selection and clustering on the selected subset (Table 3.2).

**Figure 3.10:** A brief look on the synthetic data and selected genes from low-noise set 1 (top) and high-noise set 1 (bottom). Left is the original data; middle is the projection of the original data on its two first eigenvectors; right is the selected genes and corresponding patterns ( $\delta=70\%$ ). The class structure or patterns of the datasets can be

viewed visually from the low-noise set. The red and the blue pattern (or the green and the cyan pattern) are close together, resulting that one of them is removed if the confidence level is high. In high-noise datasets, the closeness is more difficult to distinguish, leading to the removal since genes belong to these patterns are unclear about their status i.e. highly coexpressed or non-coexpressed (the horizontal axis is 20 time-points and the vertical is expression values).



**Table 3.2:** Accuracy of the selection and clustering on the synthetic class structure\*

Confidence $\delta$	Datasets									
	low-noise					high-noise				
	set 1	set 2	set 3	set 4	set 5	set 1	set 2	set 3	set 4	set 5
0.9	262 4 100	317 5 99.69	234 4 99.79	333 5 100	331 5 100	76 2 100	76 2 92.63	98 2 78.29	98 2 100	90 2 95.63
0.8	266 4 100	324 5 99.37	247 4 98.53	400 6 100	332 5 100	108 4 98.65	106 3 90.3	188 4 83.5	159 3 100	101 3 90.73
0.7	399 6 100	329 5 99.36	261 4 98.18	400 6 100	399 6 100	228 4 89.32	134 4 92.04	264 4 85.27	196 4 100	161 4 87.28
0.6	399 6 100	336 5 98.83	316 5 98.34	400 6 100	400 6 100	316 4 86.98	185 5 93.03	316 4 86.73	229 4 99.53	232 4 89.88

\*: the format of each cell is 'number of selected genes | corresponding number of patterns | accuracy'

We further evaluate the accuracy when a single method, a single metric and the entire dataset is used (Table 3.3). Even though some clustering methods/metrics can be highly accurate, the average accuracies still fluctuate around 80% on high-noise datasets whereas the accuracies our selection and clustering are around 90% (even with the moderate agreement value of  $\delta=70\%$ ). Overall the accuracy is very high in all cases, further confirming the efficacy of the selection.

**Table 3.3:** Accuracy of running one clustering element on the entire dataset

clustering methods	Datasets									
	low-noise					high-noise				
	set 1	set 2	set 3	set 4	set 5	set 1	set 2	set 3	set 4	set 5
hclust – Pear	88.43	94.21	88.00	94.38	94.29	74.16	86.04	75.44	83.69	81.88
diana – Pear	88.43	88.51	87.84	94.38	94.14	80.52	83.17	79.80	89.06	82.92
fanny – Pear	99.83	96.64	96.42	100.0	99.83	83.05	63.95	82.32	74.60	64.47
pam – Pear	100.0	96.75	96.42	100.0	100.0	88.36	87.41	84.40	92.37	88.97
hclust–Manh	100.0	94.29	88.67	100.0	100.0	67.37	21.41	66.58	22.76	19.78
diana–Manh	100.0	94.21	93.57	100.0	94.21	87.71	90.70	87.31	87.39	83.84
fanny–Manh	100.0	96.09	98.71	100.0	100.0	66.19	64.28	65.49	64.83	63.68
pam–Manh	100.0	98.45	99.02	100.0	100.0	98.68	92.77	89.02	98.70	96.98
kmeans–Euc	99.83	94.14	95.91	100.0	100.0	95.62	90.75	87.11	96.79	94.65
cmean –Euc	99.83	97.26	92.23	100.0	100.0	86.08	60.79	86.77	75.00	62.80
som–Euc	88.75	93.72	88.75	94.46	88.75	86.13	84.66	85.88	85.73	86.55
mclust –Euc	100.0	94.37	91.42	100.0	100.0	83.82	84.69	82.33	84.25	84.51
average	97.09	94.89	93.08	98.60	97.60	83.14	75.88	81.04	79.60	75.92

(Pear: Pearson correlation metric; Manh: Manhattan metric; Euc: Euclidean metric)

### *Evaluating the ‘clusterability’ of the selected subset*

‘Noisy’ data tend to lack class structure and as a result different clustering methods with different metrics produce very inconsistent class assignment results. Consequently, the process of gene selection tries to remove the noise’ from the data and pick out a more clusterable subset which contains distinguishable patterns. To evaluate this property of the selected subset we applied the uniformity testing suggested in [182] by using Friedman-Rafsky’s minimum spanning tree test [181, 182] to estimate the clusterability of a dataset. Table 3.4 quantifies the ‘clusterability’ of the original, the selected and

removed data for each of the synthetic datasets. The selected subsets have consistently superior clusterability characteristics compared to both the entire set and removed subset. Furthermore, the removed subset is consistently the less clusterable, compared even to the entire dataset

**Table 3.4:** Friedman-Rafsky test\* for clusterability on high-noise synthetic sets ( $\delta=70\%$ )

	set 1	set 2	set 3	set 4	set 5
Original data	0.54	0.49	0.56	0.56	0.58
Selected domain	0.49	0.35	0.41	0.45	0.56
Removed domain	0.74	0.58	0.86	0.65	0.65

\*: the smaller the better

In conclusion, the purpose of this approach is to enable a systematic identification of smaller, clusterable, subsets of gene expression data exploring the concept of consensus clustering. The fundamental assumption of our approach is that an appropriate weighting of multiple alternative methods would eliminate the biases associated with specific clustering methods. Also, it must be emphasized that the proposed framework is not designed, or proposed, in order to replace more refined clustering analysis, but is advocated as a critical preliminary steps in order to identify putatively informative subsets of genes given a high-dimensional expression dataset.

### 3.3. Multi-plus clustering

Hypothetically, transcriptional modules that are significantly coexpressed under different conditions/tissues will be more important gene clusters for further analysis. For example, genes with similar temporal expression profiles in response to different conditions are hypothesized to be more likely to share some common regulatory mechanisms.



### 3.3.1. Background

Rich *in vivo* datasets of pharmacological time-series across multiple dosing regimens are often obtained from different microarray platforms and time-sets [11, 99], leading to a problematical issue for computational analysis [183-185]. As an example, in a study comparing normal and chronic lymphocytic leukemia B-cells, Wang et al. [186] identified only 9 differentially expressed genes across all three studies, when combining results from three different platforms, while there are at least 1,172 differentially expressed genes in each individual platform. In general, there are two important issues relevant to the analysis of data derived from different platforms: (i) genes may be present in one platform but not in the other, and (ii) genes present on both platforms may not be represented by the same probes. Since different microarray platforms do not contain the same probesets, and even do not have a similar hardware design and sample processing protocols, standard analyses may not yield comparable expression level quantifications across platforms, leading to many challenges for computational models aiming at the analysis of microarray data from heterogeneous sources [184, 187, 188].

A number of approaches have been proposed and are generally classified into two main categories: (1) integrate raw expression profiles from different studies into one dataset so that available computational models can be directly applied, and (2) develop and/or utilize a unitless statistic as a primary analysis and then combine the result through a meta-level analysis. The former category can be further divided into two sub-classes, namely combining raw data through a normalization and/or transformation procedure [189-192] and pooling raw information from common probes that can be mapped to the same Unigene clusters or full-length mRNA transcripts [193-196]. However, these

approaches are not general enough to make data from different platforms fully compatible [184, 197]. Since combining data across different platforms remains a serious challenge, meta-analysis – the second category - has been identified as a more popular technique in order to combine results, and thus data, from a number of independent studies [198, 199]. The assumption here is that while the raw expression levels from different platforms may not be comparable, the results of the primary analysis should be. However, almost all prior studies has focused on the discovery of genes that are differentially expressed in conjunction with standard models such as effect size models [200-202], Bayesian models [203, 204].

Consequently, in order to identify significant clusters of genes that share common expression patterns across multiple dosing regimens, we extend our previous proposed method [104] in the aspect of (i) producing an agreement matrix (AM) that describes the agreement levels of co-expression of genes across multiple conditions and (ii) successively searching clusterable subsets to infer all such gene clusters. The approach follows the concept of meta-analysis to avoid the limitation of incompatible data across multiple datasets from different platforms (also different tissues, time-grids, as well as lab-protocols when applicable). The unitless statistic, expressing the confidence level of co-expression is the agreement level of cluster assignments drawn from multiple clustering runs. There remain a number of open critical issues associated with a single clustering run (e.g. the input number of clusters [149, 153], the biases and assumptions of distance metrics and/or clustering methods [156], cluster significance [165]), and thus consensus clustering coupled with the examination of AM distribution has been designed with the aims of reducing aforementioned limitations [104, 167]. Once the AM is

obtained for each condition independently (e.g. each dosing regimen in this case), an average agreement matrix is calculated to estimate the confidence levels of coexpression between genes across multiple conditions, thus combining data from different datasets into a single input for the next analysis. For the analysis at the meta-level, we extend the selection and clustering processes (also proposed in [104]) to identify all possible clusters of genes that are highly coexpressed with the average AM above as the input. As such these clusters of genes will share common patterns of expression across multiple dosing regimens. Additionally, due to the selection of all possible patterns of expression several clusters may have similar expression patterns and thus we also provide a heuristic as an optional procedure to merge such similar clusters based on a criterion of maximizing the total homogeneity and separation of selected clusters.

### 3.3.2. Problem definition

The general computational problem can be briefly defined as follows. We are given a set of  $N$  genes  $G = \{g_i\}_{i=1}^N$  and  $K$  conditions. For each condition  $k$ , every gene is characterized by one or more time-series expression profiles with  $R_{ki}$  corresponding probesets over  $T_k$  time-points

$G_k = \{g_{ki}\}_{i=1}^N, g_{ki} = \{g_{ki}^r, r \in R_{ki}\}, g_{ki}^r = \{g_{kit}^r\}_{t=1}^{T_k}, k = 1, \dots, K$ . The question then becomes

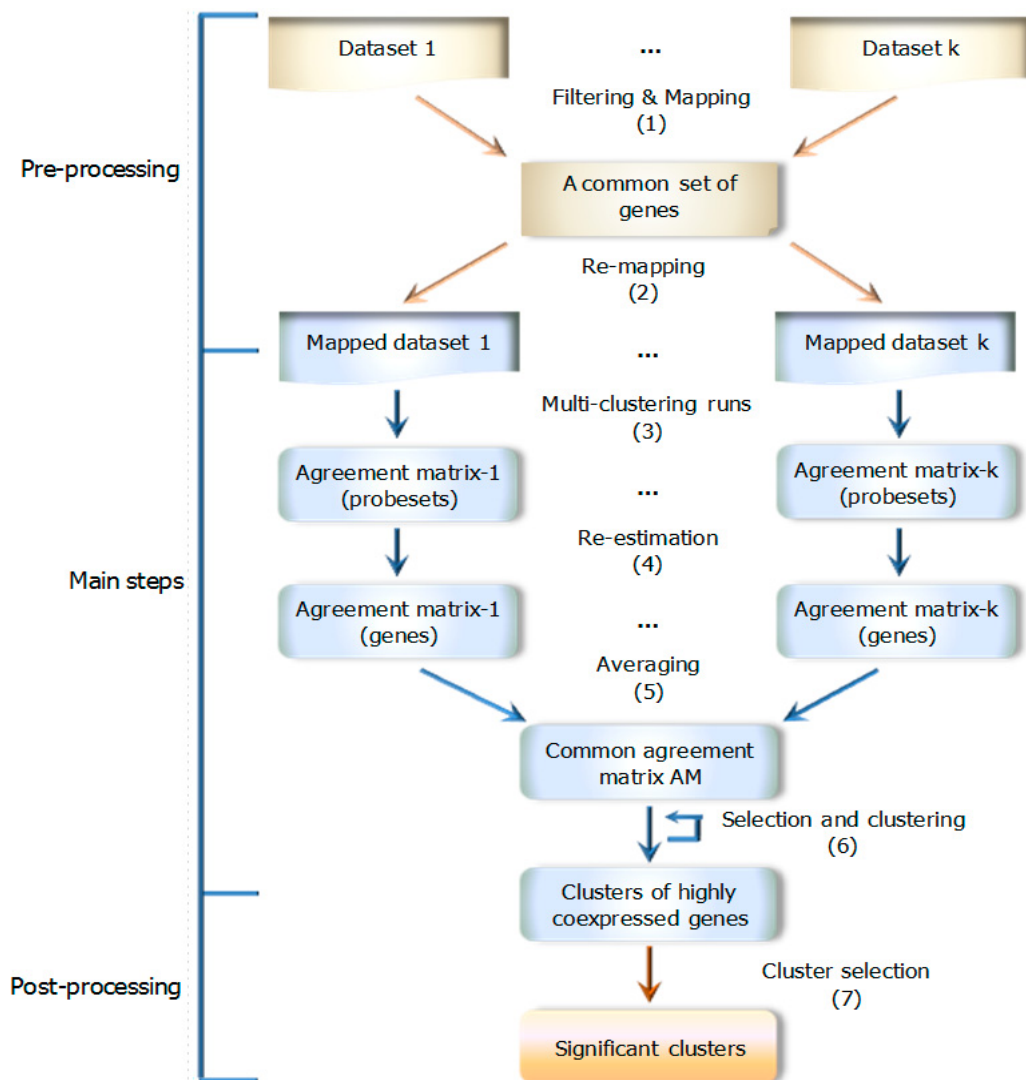
to search for clusters of genes that are highly coexpressed across all  $K$  conditions with a confidence level  $\delta$ . The term ‘highly coexpressed’ is used in the sense that

$\forall g_i, g_j \in C, \frac{1}{K} \sum_{k=1}^K P_k(g_i \wedge g_j) \geq \delta$  where  $C$  denotes a, yet to be determined, cluster

and  $P_k(g_i \wedge g_j)$  is the confidence level that two gene profiles  $i$  and  $j$  are clustered

together in condition  $k$ ; a gene profile includes sets of corresponding probesets  $R_{ki}$  of

gene  $i$  in condition  $k$ ,  $k = 1, \dots, K$ . The subscripts  $\{i, j\}$ ,  $t$ ,  $k$ ,  $r$  indicate the {gene id}, time, condition, and probesets respectively. It should be also noted that in this work, we used three different terms to refer to the same object (e.g. a set of genes that are coexpressed across multiple conditions): ‘cluster’ when designing the algorithm, ‘pattern’ when exhibiting the expression changes, and ‘module’ when charactering the biological function. The framework contains several step displayed in Figure 3.11.



**Figure 3.11:** The flowchart of the approach. The pre-processing section refers to filtering for differentially expressed probesets in each dataset, mapping to gene symbols to extract a set of common genes that are present across all datasets, and then re-mapping to corresponding probesets in each particular dataset. The main steps include establishing the AM to characterize how much confidence two probesets (and two genes) are co-expressed in each condition (and then across all conditions) and searching for all possible clusters of co-expressed genes based on the common AM. The post-processing step will select those clusters that are significant and optionally merge those with similar expression patterns if indicated.

### 3.3.3. The pre-processing step

Each dataset is pre-filtered to identify differentially expressed probesets. Since we would like to identify gene clusters with common expression patterns across multiple conditions, input datasets must contain the same set of genes. Thus using the respective platform information, probesets in each dataset are mapped to a list of genes and then the intersection across those gene lists is evaluated to extract a common set of genes which are differentially expressed across multiple conditions (i.e. datasets). However, genes are sometimes characterized by multiple probesets whose expression profiles may be similar or sometimes different, but not identical. These probesets can be considered as replicates of expression profiles for a gene and thus taking an average expression profiles across all these probesets to characterize for the expression profile of the gene may lose useful potential information. Therefore, from the common set of genes we re-map genes to corresponding probesets in each dataset with the respective platform before starting the analysis.

### 3.3.4. Construction of the meta-agreement matrix

The *agreement matrix* (AM) quantifies the likelihood that two objects (x, y) are assigned to the same cluster. If m clustering runs are performed on the data, each entry (termed ‘agreement level’) will show the frequency with which two objects are assigned to the same cluster over ‘m’ clustering runs. The AM entries are defined as follows:

$$M_{xy} = \frac{1}{m} \sum_{h=1}^m M^{(h)}(x, y)$$

$$\text{where } M^{(h)}(x, y) = \begin{cases} 1 & \text{if } x \text{ and } y \text{ are clustered together when running method } M^{(h)} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

In addition to the various clustering methods that were utilized, different distance metrics (Euclidean, Pearson correlation, and Manhattan [176]) are also explored in order to attenuate the biases associated with individual distance metrics. In our implementation, we are using hierarchical clustering (hclust), divisive analysis clustering (diana), fuzzy analysis clustering (fanny), partitioning around medoid (pam) with Pearson correlation and Manhattan metric, k-means (kmeans), fuzzy c-means (cmeans), self-organizing map (som), and model-based clustering (mclust) with Euclidean metric as the core clustering methods (supported by R packages) [139, 140, 144, 173-175]. Since clustering results are highly dependent on the input number of clusters (nc), the sensitivity of the AM as a function of nc was examined to find a ‘suggestive’ number of clusters (nc\*) for each particular dataset. After identifying nc\* based on the procedure in our prior work [104], all clustering runs are repeated with nc\* to produce the final AM for further analysis (see more details in [104]).

If two probesets (x, y) are clustered together, it is implied that their expression profiles are similar under a specific condition k. Therefore, the fraction of times (M<sub>xy</sub>) they are

clustered together over multiple clustering runs can be considered as the confidence level that they are coexpressed since  $M_{xy}$ , by construction, aims at eliminating method-specific biases and assumptions. Subsequently, we calculate the average agreement levels between sets of corresponding probesets of any two genes to estimate the confidence level that those two genes are coexpressed in a specific condition. The AM entries in condition  $k$  is re-estimated as follows

$$AM_{ij}^{(k)} = \frac{1}{|R_{ki}| |R_{kj}|} \sum_{x \in R_{ki}} \sum_{y \in R_{kj}} M_{xy}, i, j = 1, \dots, N \quad (2)$$

With the assumption that the unitless statistics, i.e. the confidence level of co-expression, is comparable across multiple conditions and different platforms [196], we estimate the confidence level of co-expression between two genes across multiple conditions by taking the average. While combining raw data remain challenges, the estimation of a unitless statistics provides a simple but efficient combination of heterogeneous data for further analyses.

$$AM_{ij} = \frac{1}{K} \sum_{k=1}^K AM_{ij}^{(k)}, i, j = 1, \dots, N \quad (3)$$

As a result, we obtain an agreement matrix whose entries exhibit a quantity that shows how confident genes are coexpressed. This will be the input for the selection and clustering process.

### 3.3.5. Selection and clustering

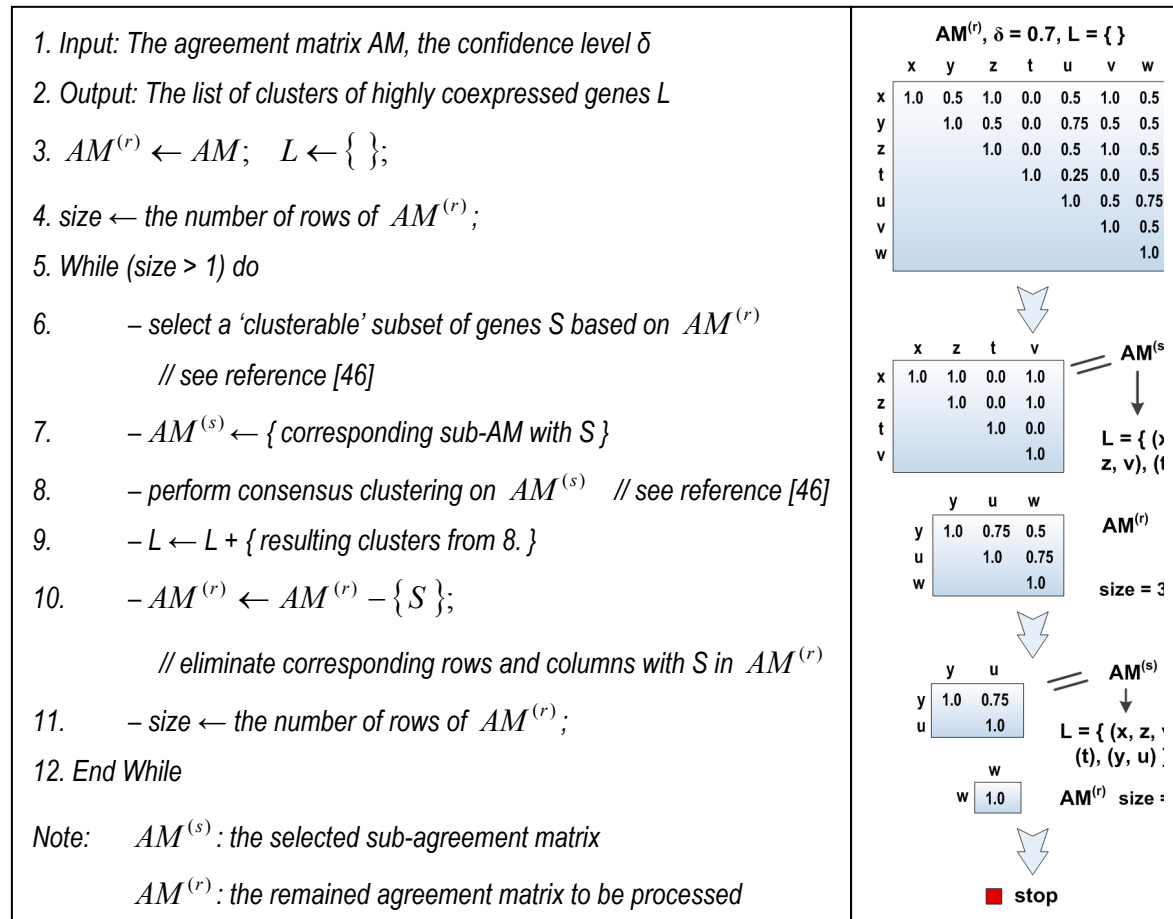
With the hypothesis that the more clusterable the data is the more biologically relevant it is, we applied our previously proposed procedure to select a more ‘hypothetically clusterable’ subset from the entire set of genes [104]. The main hypothesis underlying the selection is that AM entries associated with genes at the ‘hypothetical’ core of an

expression pattern (or a cluster) will be consistently grouped together over multiple clustering runs. This should be manifested by high corresponding values in the AM, whereas genes belonging to the ‘hypothetical’ core of two clearly distinct clusters are associated with consistently low AM entries. On the contrary, cluster assignments associated with genes at cluster boundaries or between clusters would be very sensitive to the method used and thus they would have relatively moderate agreement levels with other genes. As a result, with a user-defined confidence level  $\delta$  genes associated with moderate AM entries ( $1 - \delta < AM_{ij}^{(k)} < \delta$ ) are eliminated to produce a more ‘clusterable’ subset of genes ( $\delta = 70\%$  in this study). The process starts removing genes associated with the highest number of moderate AM entries and then updates the AM for the next loop until no moderate AM entry exists. The corresponding subset of genes is now considered as a ‘hypothetically clusterable’ subset since any two genes are highly coexpressed or non-coexpressed with the confidence level at least  $\delta$ . Subsequently, we used the concept of consensus clustering [167, 169, 171] to divide the subset of genes into a number of clusters by applying the hierarchical clustering with the selected AM as input data. The algorithm starts with every gene filling a cluster and then grouping two clusters into a new one for each loop so that any pair of genes belonging to a new cluster always has an agreement level greater than or equal to  $\delta$ . The iteration is stopped when no more new cluster is formed (see more details in [104]).

However, since there should be existed clusters of genes located closely to other clusters in the data and the input number of clusters for the core analysis is only a suggestive one, those clusters may not be completely separated. As a result, although genes that belong to those clusters are identified as highly coexpressed, their relationship to genes in other



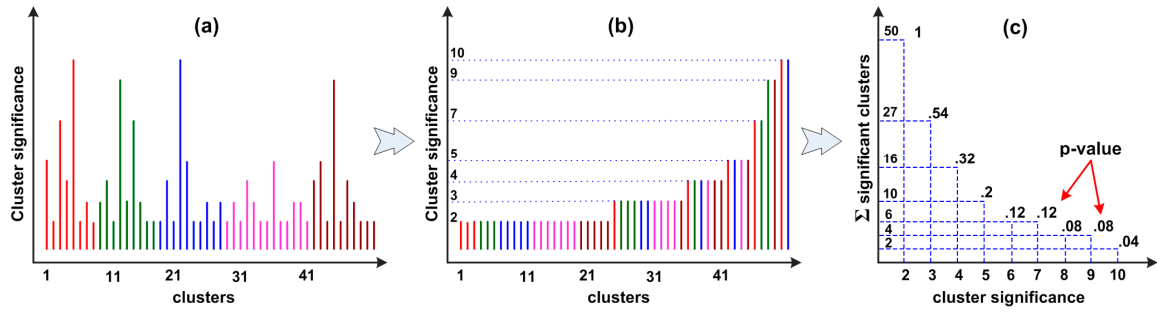
clusters cannot be uniquely determined. Therefore, some significant clusters may be not included in the selected subset due to the constraint of ‘clusterable’ selection. Since we would like to obtain all significant patterns of expression, the procedure of selection and clustering is repeated on the removed domain. The removed domain consists of a set of unselected genes whose co-expression levels are still high as quantified agreement levels in the original AM. After extracting the sub-agreement matrix corresponding to the set of unselected genes, the entire process of selection and clustering is applied again with the same confidence level  $\delta$  as before. The procedure is reiterated until no more clusters of genes are recognized. Figure 3.12 presents the pseudo-code of the procedure and an example to illustrate the process.



**Figure 3.12:** The procedure of selection and clustering. The left is the pseudo-code algorithm of the procedure. The right is an example to illustrate the process with a specific AM. At iteration 1, the process selects a ‘clusterable’ subset of genes including (x, z, t, v) that results in two clusters (x, z, v) and (t). The remained AM consists of corresponding rows and columns of genes (y, u, w) from the original AM. At iteration 2, the procedure selects (y, u) and the remained AM now contains only one gene (w); at that point, the process terminates.

Furthermore, due to the nature of clustering, trivial clusters may be identified in the final results. In order to exclude such trivial clusters, each cluster  $C$  is assigned with a simple hypothetical quantity called ‘cluster significance’ which represents how large the cluster is in this study. We then create the distribution of cluster significance on random data to estimate the cluster significance threshold corresponding to a user-defined threshold  $p$ -value for cluster selection. Without loss of generality we select  $K = 1$  for the random data and assume that each probeset in the mapped dataset  $D$  correspond to a gene; thus  $D$  is cut-off to have the number of probesets equal to the number of genes  $N$ . The suggestive number of clusters  $nc^*$  for  $D$  is searched with the process in [104]. Subsequently,  $D$  is randomly resampled (permutation plus convex-hull [165]) a number of times ( $n_r$ ), for each of which the entire process starting from building the AM with the same  $nc^*$  to extracting clusters of highly coexpressed genes is repeated and the resulting random clusters are returned. After that, the procedure estimates the cluster significance, which is simply the cluster-size in this study, for these random clusters and constructs a distribution of cluster significance. The corresponding  $p$ -value of cluster significance  $cs_0$  is defined as the number of random clusters whose significance is greater than  $cs_0$  over

the total number of random-resulting clusters in  $n_r$  times resampling and repeating the process. As a result, given a threshold p-value (p-value = 0.05 in this study), the corresponding cluster significance threshold is inferred (Figure 3.13) and only clusters with significance greater than this threshold are selected.



**Figure 3.13:** Estimating the cluster significance threshold given a user-defined p-value.

An illustrating example is shown in which  $n_r = 5$  random data are generated, the data are subsequently clustered according the proposed clustering/selection procedure and cluster significance distribution are depicted in (a) and (b) following sorting. The corresponding p-value for each cluster significance  $cs$  is estimated and depicted in (c). Thus, given a p-value, we can infer the corresponding cluster significance threshold. For example, for a p-value = 0.05, all clusters with cluster significance  $\geq 10$  are selected and if p-value = 0.1, all clusters with cluster significance  $\geq 8$  are considered as significant clusters.

Let assume that after  $n_r = 5$  times of repeating the procedure of selection and clustering from Figure 3.12 on random data, we obtain 50 random clusters whose cluster significances (cluster size in this study) are distributed as in (a) and (b) after sorted. The maximum level of cluster significance in this example is ten where only random clusters whose cluster significance is at least two are selected for the process. The corresponding p-value for each cluster significance  $cs_0$  is defined as

$pvalue(cs_0) = \frac{\sum \text{clusters with cluster significance} \geq cs_0}{\sum \text{random clusters}}$ . Therefore, given a p-value, we

can infer the corresponding cluster significance threshold. For example, if p-value = 0.05, all clusters with cluster significance  $\geq 10$  are selected and if p-value = 0.1, all clusters with cluster significance  $\geq 8$  are considered as significant clusters.

### 3.3.6. Merging similar patterns

Because of the nature of the approach, it is quite reasonable to expect that the clustering process can break out patterns of expression into several sub-patterns. Thus, we repeat the process on the eliminated domains to extract all possible significant clusters, resulting in that several clusters may have a similar expression pattern but are separated into two or more clusters. Because cluster homogeneity reflects the similarity of expression profiles within a given cluster and cluster separation quantifies how effectively expression profiles are discriminated, we provide an optional heuristic in order to merge similar patterns together according to the criterion of maximizing sum of homogeneity and separation of all final output clusters. Starting with all significantly selected clusters, the procedure searches for a combination of two similar patterns so that their combination will generate a maximal increase of the sum of homogeneity and separation of all current clusters after merging those two patterns. The process is repeated until no more combinations are found i.e. any new combination always reduces the sum of homogeneity and separation. Eventually, a list of significant expression patterns that characterize the underlying transcriptional responses is generated. The metric used during the evaluation of the heuristic is quantified as follows:

$$\begin{aligned}
\max_C \{homogeneity + separation\} &= \max_C \left\{ \frac{1}{K n} \sum_k \sum_p H_k(C_p) + \frac{2}{K n(n-1)} \sum_k \sum_{p < q} S_k(C_p, C_q) \right\}; \\
H_k(C_p) &= \frac{\sum_{g_i, g_j \in C_p; g_i \neq g_j} sim(g_{ki}, g_{kj})}{\|C_p\|(\|C_p\| - 1)}; S_k(C_p, C_q) = \frac{\sum_{g_i \in C_k; g_j \in C_l} dis(g_{ki}, g_{kj})}{\|C_p\| \|C_q\|}; \\
sim(g_{ki}, g_{kj}) &= \frac{1}{\|R_{ki}\| \|R_{kj}\|} similarity(g_{ki}^r, g_{kj}^{r'}); dis(g_{ki}, g_{kj}) = \frac{1}{\|R_{ki}\| \|R_{kj}\|} dissimilarity(g_{ki}^r, g_{kj}^{r'}); \\
p &= 1, \dots, n, q = 1, \dots, n, k = 1, \dots, K, r \in R_{ki}, r' \in R_{kj}
\end{aligned}$$

where  $C$  is the current set of selected clusters  $C = \{C_p\}_{p=1}^n$  and  $n$  is the current number of clusters;  $H_k(C_p)$  is the homogeneity of cluster  $C_p$  in condition  $k$  and  $S_k(C_p, C_q)$  is the separation between cluster  $C_p$  and  $C_q$  in condition  $k$ ;  $sim(g_{ki}, g_{kj})$  and  $dis(g_{ki}, g_{kj})$  are the average similarity and dissimilarity (or distance) respectively between all probesets of gene  $i$  and gene  $j$  in condition  $k$ . Similarity is measured by the Pearson correlation coefficient and dissimilarity is estimated by the Pearson correlation distance.

### 3.3.7. Method evaluation on synthetic data

In order to evaluate the effectiveness of the proposed approach, we use synthetic data with known class structure as described earlier. The process of evaluation is repeated four times with four different datasets that are created with different number of replicates for each time-point (1, 3, 4, and 20 respectively). In each time, we use five high-noise sets as the data for five conditions ( $K=5$ ), each of which has 400 genes distributed across 6 clusters; each cluster has different patterns across five conditions but has the same set of gene ids. We set the same parameters for all evaluation in this study and also for the analysis on real time-series datasets, specifically the confidence level of co-expression  $\delta = 70\%$  and  $p\text{-value} = 0.05$  for the selection of significant clusters. Furthermore, the

testing process on synthetic data is done without the merging option. Without loss of generality, we assume that each gene has only one probeset in this evaluation. The performance of the approach is assessed through its ability to recover the number of cluster structures and the list of gene ids identified in each cluster. We use the adjusted Rand index [105, 145] which is a statistic that measures the extent of concurrence between the clustering results and the underlying known class structure to evaluate the approach's performance in identifying gene clusters that are coexpressed across multiple conditions. The larger the Rand index is, the higher the agreement between the results and prior knowledge of cluster structure. The number of selected genes, recovered cluster structures and the accuracy on the selected domain are listed in Table 3.5.

**Table 3.5:** The clustering effectiveness of the approach

Synthetic data	Number of selected genes	Number of clusters	Accuracy* (Adjusted Rand-index)
Dataset 1	174/400	4	100.0%
Dataset 2	368/400	6	100.0%
Dataset 3	395/400	6	100.0%
Dataset 4	378/400	6	100.0%

\*: The accuracy is only estimated on the selected domain

Due to the fact that these datasets are high-noise synthetic data, some cluster structures may be missed when there is only one measurement at each time-point. However, when the number of replicates is increased, the number of cluster structures is recovered. As discussed in our previous study [205], this is a reasonable observation due to the effect of gene expression replicates on clustering performance. Additionally, we also examine an alternative approach which is more intuitive in identifying gene clusters that are coexpressed across multiple conditions. Instead of performing a meta-analysis to avoid the limitation of incompatible data across different platforms, we can separately identify

significant clusters of genes that are coexpressed in each condition (set of data) and then obtain the intersection among these gene clusters across all conditions. In this experiment, we used pam [144], mclust [139], and consensus clustering [104] as standard single clustering methods to identify clusters in each set of data, for which  $nc^* = 6$  is the input number of clusters. We then simply took the intersection between clusters from set to set and only keep those clusters that contain more than 5 genes as significant clusters for the final estimation of accuracy. The number of selected genes, number of clusters, and accuracy on the selected domain are listed in Table 3.6.

**Table 3.6:** Effectiveness of the approach on synthetic data

	pam			mclust			consclust		
	# of sel. genes	# of clusters <sup>+</sup>	Accuracy (%)	# of sel. genes	# of clusters	Accuracy (%)	# of sel. genes	# of clusters	Accuracy (%)
Dataset 1	122	7 6	74.8 82.8	155	9 7	87.8 84.1	68	4 4	100 100
Dataset 2	337	6 6	100 100	343	6 6	100 100	330	6 6	100 100
Dataset 3	374	6 6	100 100	380	6 6	100 100	374	6 6	100 100
Dataset 4	376	15 7	68.8 94.6	375	13 7	80.1 94.2	302	11 7	80.9 92.4

<sup>+</sup>: only clusters with more than 5 genes; \*: the accuracy is estimated on the selected domain; the number of clusters and the accuracy are formatted as ‘before merging’ | ‘after merging’

In general, this approach selects a smaller number of genes with an equal or greater number of cluster structures, resulting in a less accuracy. As an example, in each set of dataset 4 there are two cluster structures that are not clearly distinct. As a result, single clustering methods (even consensus clustering) may fail to properly separate them in each set, leading to the case that the intersection between clusters from set to set divides those cluster structures into many sub-clusters with small number of genes. On the contrast, by taking the average of the co-expression levels across multiple sets, the relationship of whether two genes are coexpressed across multiple conditions can be recovered.

Consequently, our proposed approach is more advantage, resulting in a final highly correct classification as illustrated in Table 3.5. Furthermore, since this simpler alternative approach produces many resulting clusters, we also attempted to apply the proposed merging process to reduce the number of clusters as well as improve the accuracy if applicable. However, its testing performance is still not as high as that of our proposed approach although we do not apply the merging process for the proposed approach in this test. Additionally, the alternative approach is highly sensitive with the initial number of clusters. For instance, when we constantly set  $nc^* = 7$  and test on dataset 3, without the merging option our approach still recovers the correct number of cluster structures with high accuracy: (number of selected genes, number of clusters, accuracy) = (386, 6, 100%) whereas ‘pam’ approach yields (366, 13, 87.8%), ‘mclust’ provides (360, 11, 82.3%), and ‘consclust’ does (351, 7, 98.3%). Since this information is almost not existed for all real datasets, the more sensitive with it the less robust the approach is. Therefore, by taking the average of the co-expression levels between two genes across multiple datasets, our proposed approach provides more robust results.

### **3.4. Results from corticosteroids pharmacogenomics model**

It has been noticed that long-term treatment with this kind of drugs causes a lot of side-effects and thus we ask that whether we can explore the complexity of gene expression changes to understand how the drug alter systemic physiology and contribute to adverse-effects. In this study we explore the hypothesis that genes that are coexpressed across multiple dosing regimens of corticosteroids may provide important implications for further analyses. Consequently, we applied proposed model to estimate ‘true’ expression profiles of genes in acute and chronic CS administration and then used the multi-plus



































































































clustering to find such transcriptional modules i.e. set of genes that are coexpressed across multiple conditions.

We first pre-processed datasets following the pre-processing step in Figure 3.11. The datasets are first filtered for differentially expressed probesets using ANOVA technique ( $p\text{-value} < 0.05$ ) implemented in R [100] and also customized by our previous work for easy uses [104]. 2,920 probesets in the acute and 4,361 probesets in the chronic are selected for further analysis. To obtain the common set of genes across two conditions, these probesets are mapped into sets of genes based on the corresponding platform information. 2,920 differentially expressed probesets in the acute are mapped into a set of 2,340 genes and 4,361 probesets in the chronic are mapped into another set of 4,076 genes. The intersection of these two gene-sets yields 967 genes in common for both dosing regimens. From this common gene set, the re-mapping process subsequently returns a corresponding set of 1,314 probesets for the acute and a set of 1,112 probesets for the chronic data. All datasets (including synthetic data) are pre-processed with the model in our previous study to estimate the ‘true’ expression profiles that are integrated with potential information in replicates instead of simply taking the average expression profiles [205]. The suggestive number of clusters  $nc^*$  for both datasets is 7.

Subsequently, we apply the proposed approach with the merging option to the intersection set of 967 genes that are affected by corticosteroid administration across the two dosing regimens. We obtain 6 significant clusters with 315 genes in total. These clusters are hypothesized to be transcriptional modules which share common regulatory mechanisms since they consist of genes that exhibit similar expression patterns in both acute and chronic dosing regimen. Table 3.7 shows the distribution of these 315 genes

over six modules and also briefly describes how the pattern of expression changes. Although genes may exhibit simple or complex patterns of expression during corticosteroid administration, we crudely classify those patterns into up- or down- with one or two phases of regulation.

**Table 3.7:** Characterization of significant transcriptional modules

Transcriptional modules	1	2	3	4	5	6
Number of genes	97	45	34	71	14	54
Expression pattern in acute*	            	          	        	      	    	  
Expression pattern in chronic*	            	          	        	      	    	  

\*: Patterns consist of one-phase regulation (up-down/down-up), two-phase regulation (up-down-up/down-up-down) or simply up- (red) or down-regulation (green).

### 3.4.1. Critical transcriptional modules

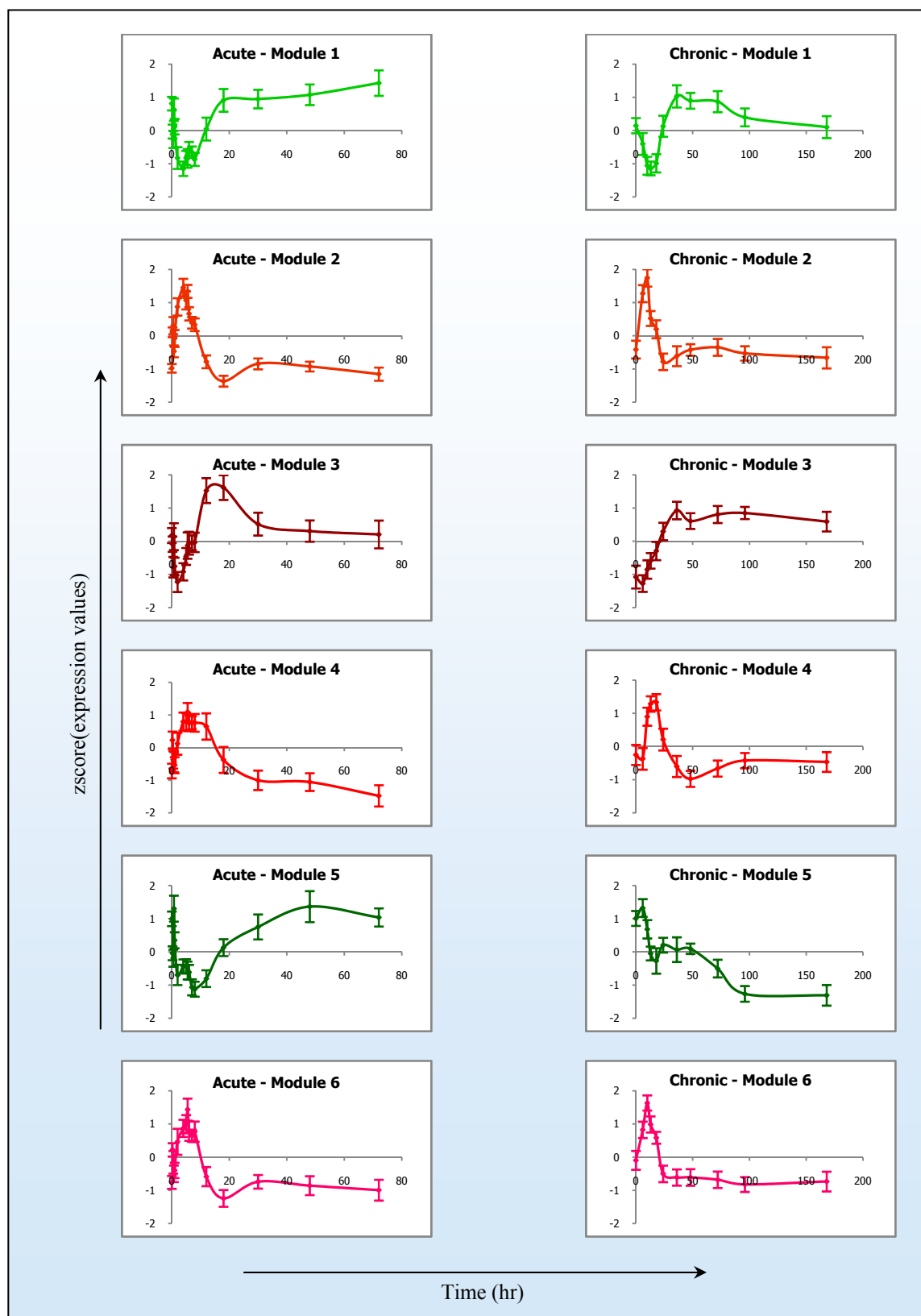
A detailed description for patterns of these transcriptional modules is shown in Figure 3.14 with the average expression patterns of all probesets clustered in each module following acute and chronic dosing. In brief, transcriptional module 1 (97 genes) is characterized by one-phase regulation in acute but two phases in chronic dosing. Genes in this module exhibited a fast and robust decline in mRNA, which reached its peak between 4h and 8h, and returned to the baseline after about 18h. However, when MPL is infused (chronic dosing) this set of genes shows a more complex pattern involving both enhanced and suppressed regulation. Although a strong down regulation is observed at the beginning, it is subsequently followed by a sharp induction with the maximum around 36h and then gradually returned to the baseline indicating some kind of possible tolerance. The second transcriptional module (45 genes) shows a similar pattern of expression in both acute and chronic regimen with two phases of regulation. Genes in this module exhibit an early up-regulation and reached their corresponding peaks at around 4h

in the acute and 10h in the chronic. Subsequently, both profiles denote a clear down-regulation (around 18h in acute and 24h in chronic) and possible slight fluctuation before returning to base line. An interesting dynamics is observed in the 34 genes of transcriptional module 3. In the acute dosing, the genes in this module clearly exhibit an expression pattern with two phases of regulation (down-up-down). Yet, in chronic administration they exhibited an early transient decline in mRNA followed by robust, sustained, up-regulation.

Similar to module 2 is the transcriptional dynamics exhibited by transcriptional module 4 (71 genes) characterized by an early induction with a maximum at 5.5h in the acute and 18h in the chronic. A typical pattern with down regulation for both acute and chronic administration is illustrated by transcriptional module 5 (14 genes). However, genes in the acute regimen exhibited a fluctuated repression with a maximum at around 8h and then followed by an induction to return to the baseline as late as 72h.

Meanwhile, genes in the chronic regimen characterized a pattern with a slightly transient up-regulation followed by a sustained down-regulation and eventual convergence to a new steady state in the presence of the drug. The last transcriptional module (54 genes) has a similar acute pattern of expression with two phases of regulation as that of module 2. However, in the chronic regimen after falling to a value below the baseline (~24h) this set of genes was further sustained a slight suppression.

**Figure 3.14:** Critical transcriptional modules of CS pharmacogenomic effects. Each module is characterized by the average gene expression profile of the corresponding cluster in the acute and the chronic data. The error bar shows the standard deviation of all probeset transcript levels at each time-point in each corresponding pattern.



While comparing these expression patterns, we observe that modules 2, 4, and 6 have similar expression patterns in acute (2 & 6) or chronic (2 & 4) with a slight difference in the other dosing regimen (e.g. 2 & 6 in chronic, 2 & 4 in acute). Although the difference is not large enough to be intuitively recognized, the merging process could not merge them together, implying that the difference is significant. Furthermore, the separation of these expression patterns is also reinforced with different functional characteristics which will be illustrated below. In summary, selected transcriptional modules exhibit a number of typical expression patterns under corticosteroid administration. The pattern can be simply expressed as an up- or down- regulation or as a more complex one with two phases of regulation plus some fluctuation.

### **3.4.2. Functional characterization of critical transcriptional modules**

Since selected transcriptional modules consist of sets of genes that are coexpressed across all dosing regimens, we hypothesize that these genes are more likely involved in critical functions following the drug treatment. Consequently, we search for enriched functions in these modules to explore the functional effects of corticosteroids on target genes as well as evaluate the importance of the selected modules. Using ArrayTrack [206], we first identify the gene ontology terms (GO) that are significant in each transcriptional module ( $p\text{-value} < 0.0001$ , at least 5 genes). We then classify them into super-categories (so-called main functions) based on the branch of molecular function and biological process in the GO tree. Table 3.8 lists the distribution of main functions across selected transcriptional modules.

**Table 3.8:** Connecting CS transcriptional modules to enriched gene ontology terms (p-value<0.0001)

No.	Gene Ontology Terms*		Module 1	Module 2	Module 3	Module 4	Module 5	Module 6
1	Metabolic process	Amino acid, compound, organic acid						
		mRNA						
		Nucleotide, nucleoside						
		Protein, macromolecule						
2	Binding	Cofactor, coenzyme, vitamin, heme, ion						
		Nucleotide, nucleic acid binding						
		RNA binding						
		Protein binding						
3	Cellular catabolic process							
4	Catalytic, oxidoreductase activity							
5	Oxidative phosphorylation							
6	Transmembrane transporter activity							
7	Protein-RNA complex assembly							
8	RNA splicing, processing							
9	Gene expression							
10	Translation activity							
11	Biosynthetic process							
12	Structural molecule activity							

In general, all modules are involved in metabolic processes and binding category (except module 5 since it is too small to include significant GO terms). Some modules seem to share almost all main functions e.g. module 2, 4 and 6 whereas others seem to share less e.g. module 2 and 3, 3 and 4, or 3 and 6. However, they are shown to have different roles with specific functions in those main categories. For example although module 2 and 4 are involved in metabolic processes and binding, module 2 is associated with RNAs and nucleotides whereas module 4 is specialized in proteins and macro-molecules. These functional differences (coupled with pathway analysis in Table 3.9) can be linked to the

similarities/differences in their corresponding expression patterns, strengthening the phenomenon that they are classified as distinct transcriptional modules although their expression patterns are not intuitively separated. However, the most important conclusion drawn from this analysis is that all these transcriptional modules consist of components that participate in metabolic processes, implying that they include genes that experience metabolic effects under corticosteroid administration.

Using ArrayTrack, we also searched for enriched pathways in these transcriptional modules ( $p\text{-value} < 0.01$ ). A large proportion of significant pathways selected in each module are metabolic pathways of amino acid metabolism or biosynthesis, providing another support that selected transcriptional modules are critical and able to capture metabolic side effects for further analysis. Table 3.9 shows significant pathways in each transcription module.

It is generally accepted that expression levels of many CS-affected genes are mediated through the binding motifs, called GREs – glucocorticoid response elements, on their control regions. We thus examine the presence of this binding site on the promoter of genes in each of the enriched pathways in order to assess the possible effect of GRE of metabolic functions. However, such GREs are short (5–9 bp) and fairly degenerate, leading to matches occurring by chance alone thus not implying any kind of functionality. In order to address this issue, after extracting gene promoters from the Genomatix database we identified conserved regions across sets of orthologous promoters. As a result, those matches located on these conserved regions would be more reliable estimates of functional binding sites.

**Table 3.9:** Connecting CS transcriptional modules to enriched biological pathways

Transcriptional modules	Enriched biological pathways	p-value (<0.01)	GRE <sup>+</sup>
1	Nitrogen metabolism(rno00910)	0.0000313	X
	Glycine, serine and threonine metabolism(rno00260)	0.0006195	x
	Bisphenol A degradation(rno00363)	0.0009858	√
	Tryptophan metabolism(rno00380)	0.0013596	x
	Histidine metabolism(rno00340)	0.0017470	√
	beta-Alanine metabolism(rno00410)	0.0020365	√
	Bile acid biosynthesis(rno00120)	0.0027013	√
	Arachidonic acid metabolism(rno00590)	0.0053445	
	Pantothenate and CoA biosynthesis(rno00770)	0.0056735	√
	Butanoate metabolism(rno00650)	0.0072639	√
	Tyrosine metabolism(rno00350)	0.0079428	√
	Valine, leucine and isoleucine degradation(rno00280)	0.0094101	√
2	Tyrosine metabolism(rno00350)	0.0000590	√
	Aminophosphonate metabolism(rno00440)	0.0001267	x
	Selenoamino acid metabolism(rno00450)	0.0004668	x
	Histidine metabolism(rno00340)	0.0010152	x
	Alanine and aspartate metabolism(rno00252)	0.0013658	√
	Arginine and proline metabolism(rno00330)	0.0019112	√
	Tryptophan metabolism(rno00380)	0.0040672	x
	Androgen and estrogen metabolism(rno00150)	0.0042813	X
3	Oxidative phosphorylation(rno00190)	9.000E-08	x
	Androgen and estrogen metabolism(rno00150)	0.0000888	
	Starch and sucrose metabolism(rno00500)	0.0020632	
	Urea cycle and metabolism of amino groups(rno00220)	0.0069082	
	Pentose and glucuronate interconversions(rno00040)	0.0076060	
4	Ribosome(rno03010)	0.000E+00	
	Proteasome(rno03050)	0.0000037	X
5	None		
6	Proteasome(rno03050)	2.570E-04	x
	Tight junction(rno04530)	3.410E-04	x
	Long-term depression(rno04730)	4.090E-04	x
	TGF-beta signaling pathway(rno04350)	5.040E-04	x
	Wnt signaling pathway(rno04310)	0.0032164	X

<sup>+</sup>: Glucocorticoid Receptor Element - GRE binding sites; √: GRE binding sites are present on the promoters of almost all genes in the corresponding function group; x: possibly because of not enough promoter information to be considered.

Although it is currently believed that GREs are composed of two hexamers with a three-nucleotide random-hinge region in between, the general consensus is that towards one hexamer, namely TGTTCT [10]. We therefore search for this motif on conserved

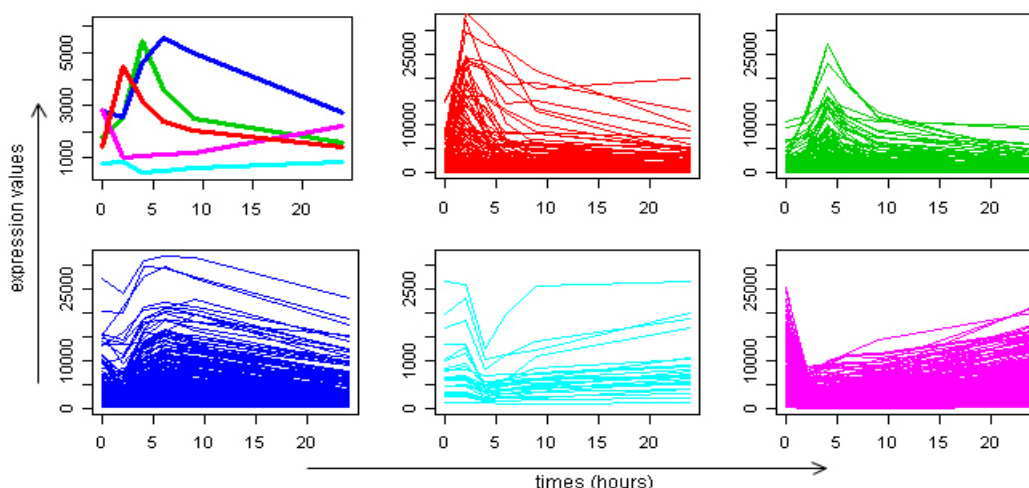


promoter regions across orthologous promoters of the selected genes. The results are shown in Table 3.9 and detailed information is provided in additional files – ‘functional\_characterization.xls’. In general, almost all metabolic pathways contain genes with the GRE binding sites, implying that these genes are more likely to be directly regulated by the complex between corticosteroids and glucocorticoid receptors. Additionally, we also examine how frequently the GRE binding sites are present on the control regions of all selected genes (315 genes). Furthermore, we determined that given a background set of 2,000 randomly selected genes, the frequency of GREs in a set of genes is similar to that in the random set (~20%), implying that not all genes in those modules are directly regulated by the drug and that the presence of GRE binding sites on the control regions of genes in enriched pathways is very significant and not random.

### **3.5. Results from human endotoxemia model**

The analysis identifies a reduced subset of genes which form, initially, five distinct responses whereas four are significant clusters (Figure 3.15). These include two clusters that exhibit an early and middle up-regulation event (182 and 199 genes respectively), one cluster that is characterized by later up-regulation (284 genes) and two clusters that exhibit a down-regulation response (1118 and 27 genes respectively). The smallest of the down-regulated clusters can be eliminated using our cluster elimination procedure as a non-significant statistic cluster ( $p\text{-value} = 0.05$ ). It must be emphasized that the design of the study was to evaluate a self-limited inflammatory response in humans injected with endotoxin. As such, once the infection is cleared the system is expected to return to homeostasis. It is important therefore to realize that all clusters to show deviations from homeostasis and eventual return to base line. In order to further evaluate the significance

of our selection we characterized functionally the populations making up the identified clusters using ArrayTrack [206] (Table 3.10).



**Figure 3.15:** Selected genes and patterns from LPS dataset. The initial number of clusters is six and we picked out five distinguished patterns (3 up- and 2 down- regulation) in which cyan pattern can be omitted since it is not significant; early up – 182 genes (red), middle up – 119 genes (green), late up – 284 genes (blue), cyan – 27 genes, down – 1118 genes (magenta), and totally 1730 selected genes over 3269. Top-left is the average expression profiles of those six patterns and the rest are expression profiles of selected genes in five patterns (the horizontal axis is six time-points (0, 2, 4, 6, 9, 24) and the vertical is the intensity of mRNA levels).

We will discuss a brief of biological implications from significant pathways of those selected patterns (Table 3.10) to show how the strategy captures the biological event. First, the ‘early-up’ pattern contains genes whose expression levels increase during the first 2hrs after the administration of endotoxin and then return to the baseline within the first 24hrs. Such an ‘early-peak’ response consists of genes that are involved in critical pro-inflammatory signaling pathways (e.g. Toll-like receptor signaling (TNF, CCL4,

IL1B, NFkBIA) and Cytokine-cytokine receptor interaction (C-X-C motifs, CXCL1, CXCL2, CCL20, IL1A) which play an integral role in the progression of systemic inflammation. For example, endotoxin when binds to its signaling receptor triggers a signal transduction cascade that converges to the activation of transcription factors (NF-kB) essential for the transcriptional synthesis of various pro-inflammatory genes (IL1, TNF, IL8) [207]. Therefore, the expression level of NFkBIA which encodes for the primary inhibitor of NF-kB [208] goes up, coupling with the co-expression of the pro-inflammatory cytokines (TNF, IL1A, IL1B).

**Table 3.10:** Pathway enrichment in four selected patterns (p-value < 0.05)

Patterns	Map Title	P-value	Patterns	Map Title	P-value
Early-up	<i>Toll-like receptor signaling pathway</i> *	0.00039	Late-up	<i>Apoptosis</i> *	0.00042
	Type I diabetes mellitus	0.00126		<i>Toll-like receptor signaling pathway</i>	0.00650
	<i>Cytokine-cytokine receptor interaction</i> *	0.00155		<i>Cytokine-cytokine receptor interaction</i> *	0.00968
	Coumarine and phenylpropanoid biosynthesis	0.00241		Limonene and pinene degradation	0.01177
	<i>Apoptosis</i>	0.01309		<i>Jak-STAT signaling pathway</i> *	0.01277
	Alzheimer's disease	0.03749		Hematopoietic cell lineage	0.01478
	Epithelial cell signal. in Heli. pylori infection	0.03816		Epithelial cell signal. in Heli. pylori infection	0.02561
	Glycan structures – degradation	0.03999		Alkaloid biosynthesis II	0.04661
	Adipocytokine signaling pathway	0.04406	Magenta	<i>Oxidative phosphorylation</i> *	0.00000
	Fc epsilon RI signaling pathway	0.04877		<i>Ribosome</i> *	0.00000
Middle-up	<i>Apoptosis</i> *	0.00000		Caprolactam degradation	0.00130
	Adipocytokine signaling pathway	0.00334		Lysine degradation	0.00147
	<i>Toll-like receptor signaling pathway</i> *	0.00743		Fatty acid elongation in mitochondria	0.00191
	B cell receptor signaling pathway	0.01715		Reductive carboxylate cycle (CO2 fixation)	0.00287
	Epithelial cell signal. in Heli. pylori infection	0.02101		<i>Citrate cycle (TCA cycle)</i> *	0.00514
	Pancreatic cancer	0.02531		Folate biosynthesis	0.00716
	Chronic myeloid leukemia	0.02810		N-Glycan biosynthesis	0.00825
	Prostate cancer	0.03856		Butanoate metabolism	0.01386
	Small cell lung cancer	0.03856		Type I diabetes mellitus	0.01386
	Sphingolipid metabolism	0.03910		T cell receptor signaling pathway	0.02075
	Folate biosynthesis	0.04525		Antigen processing and presentation	0.02215
	T cell receptor signaling pathway	0.04691		Aminoacyl-tRNA biosynthesis	0.02295
				Amyotrophic lateral sclerosis (ALS)	0.02321
				<i>Pyrimidine metabolism</i> *	0.03201
				<i>Pyruvate metabolism</i> *	0.03584
				Valine, leucine and isoleucine degradation	0.03966
				Galactose metabolism	0.04307
				Purine metabolism	0.04504

\*: selected pathways for discussing biological functions

Next, the ‘middle-up’ pattern is characterized by an increased expression of genes that peak at 4hrs post-endotoxin administration and participate in inflammatory relevant signaling pathways such as Apoptosis (CASP10, CFLAR, FAS) and Toll-like receptor signaling (NFkB1, NFkB2, RELA). The Toll-like receptor signaling is repeatedly appeared as an enriched pathway in this pattern compared with the ‘early-up’ one since some inflammatory genes (e.g. members of NF-kB/RELA family) show increased expression levels during the first 2-4hrs which were already reported in [64]. In the other hand, recent insight [209] indicated that there was an excessive death of immune effector cells (apoptotic cells) during the progression of an aberrant inflammatory response. This fact shows how the apoptosis is important and thus how efficient the approach captured the biology function with the fact that the most enriched pathway in this class of genes is Apoptosis (p-value  $\sim 10^{-7}$ ).

Subsequently, the ‘late-up’ pattern composes of genes with late expression level during the 4-6hrs post-endotoxin administration and subsequent resolution at 24hrs. Such a temporal pattern is enriched with genes involved in inflammatory relevant biological pathways as it previously stated e.g. Apoptosis (CASP8, IRAK4, PIK3G) and Cytokine-cytokine receptor interaction (IL10RB, IL13RA1, IL8RB). However, herein, JAK-STAT cascade (IL10RB, STAT5B, JAK3, and IL13RA) is an additional inflammatory relevant pathway that discriminates this pattern from the aforementioned. From a biological point of view, JAK-STAT cascade is essential to regulate the expression of target genes that counteract the inflammatory response. In addition to this, research evidence [210] suggest that a STAT pathway from a receptor signaling system is a major determinant of key regulatory systems including feedback loops such as SOCS induction which subsequently

suppresses the early induced cytokine signaling. Therefore, genes that are co-expressed in this pattern participate in anti-inflammatory processes that aim to restore homeostasis.

Finally, the ‘down’ pattern is the most populated expression motif characterized by a decreased gene expression level during the time course of the experiment. These genes are involved in cellular bio-energetic processes with a large array of genes to participate in pathways (p-value  $\sim 10^{-7}$ ) such as Oxidative phosphorylation (ATP5A, COX and NDUF members) and Ribosome biogenesis and assembly (RPL/RPS family). Other suppressed genes that involve Purine (PDE4A, PDE8A, PRPS1) and Pyruvate metabolism (GLO1, PDHB, LDHB) participate in TCA cycle (MDH1, MDH2, ACLY) as well as in metabolic pathways. Endotoxin-induced inflammation causes the dysregulation of leukocyte bioenergetics and persistent decrease in mitochondrial activity can lead to reduced cellular metabolism [211]. That is to say, co-expressed genes in this down-regulated pattern indicate the shut-down in cellular energetic of human blood leukocytes when exposed to an inflammatory stress.

Altogether, the computational analysis of the genome-wide transcriptional profiling of peripheral human blood leukocytes identifies the emergence of four distinct expression patterns that play an integral role in the progression of an endotoxin-induced inflammatory response.

### **3.6. Summary**

We have proposed a statistical model that accounts for the variability in repeated measurements to estimate more robust expression profiles, so-called ‘true’ expression profiles. The effectiveness of the model has been demonstrated on synthetic data as the method that achieves superior and/or comparable clustering performance to that of other

related approaches, especially much better to that when using the average expression profiles. Our results on synthetic data demonstrate that the clustering performance using ‘true’ expression profiles is superior to that when using average expression profiles and also to other methods with integrated error information. The output of this representation can be used as a powerful input to a variety of computational models that require gene expression profiles as the input without any modification while still taking into account the information content in replicated data.

We next explore the hypothesis that the more clusterable the data is the more biologically relevant it is and utilize the concepts of consensus clustering to identify, within a set of differentially expressed genes, a subset of genes that are either highly co-expressed or highly non-coexpressed with the hope of extracting a more biologically relevant subset of genes. The purpose of this approach is to enable a systematic identification of smaller, clusterable, subsets of gene expression data exploring the concept of consensus clustering. The fundamental assumption of our approach is that an appropriate weighting of multiple alternative methods would eliminate the biases associated with specific clustering methods. Also, it must be emphasized that the proposed framework is not designed, or proposed, in order to replace more refined clustering analysis, but is advocated as a critical preliminary steps in order to identify putatively informative subsets of genes given a high-dimensional expression dataset.

Eventually, we have proposed a framework to identify significant coexpressed clusters of genes across multiple datasets. Following the orientation of meta-analysis, an extended computational approach that explores the concept of agreement matrix from consensus clustering has been proposed with the aims of identifying gene clusters that share

common expression patterns across multiple dosing regimens as well as handling challenges in the analysis of microarray data from heterogeneous sources, e.g. different platforms and time-grids in this study. Analysis on rich *in vivo* datasets of corticosteroid time-series yielded significant insights into the pharmacogenomic effects of corticosteroids, especially the relevance to metabolic side-effects. This has been illustrated through enriched metabolic functions in those transcriptional modules and the presence of GRE binding motifs in those enriched pathways, providing significant modules for further analysis on pharmacogenomic corticosteroid effects.

## **Chapter 4 – Reconstruction of the transcriptional regulatory program**

### **4.1. Introduction to transcriptional regulation**

The gene is the fundamental unit on the genomic DNA which contains the required information to carry out the biological functions of cells. The expression of genes i.e. mRNA synthesis can be measured efficiently in a high-throughput fashion and such expression patterns are characteristic of cellular responses to external stimuli [212]. It is widely accepted that these responses are mainly driven by the interactions between transcription factors (TFs) and their corresponding transcription factor binding sites (TFBSs) on the proximal promoters of their target genes [96, 97]. However, with a large number of genes in eukaryotic genomes, deciphering how these interactions evolve to control the expression of tens of thousands of genes (~ 35,000 genes in human) remains an open question. Recent studies [213] have shown that the underlying regulatory mechanisms are complex, dynamic (especially in higher organisms) and can be arranged in multiple hierarchical levels such as the sequence, the chromatin, and the nuclear level. The sequence level, also the best-studied level of gene regulation, is characterized by the linear organization of transcription units and cis-regulatory elements considered as the regulatory code which governs gene expression. These cis-regulatory elements i.e. binding sites which are more important when found on the proximal promoters form a highly flexible and context-dependent structure [214] for each gene [98, 215, 216]. Furthermore, in eukaryotic cells genomic DNA is ‘packed’ into an efficient structure,



called chromatin, composed of nucleosomes that consist of approximately 147bp of DNA wrapped around a protein octamer [217, 218]. This structure not only packs DNA but also creates an added layer of gene regulation which ensures correct gene expression and accessibility to DNA-dependent processes e.g. gene transcription, DNA repair, and DNA replication. The overall process of the transcription process encompassing the nuclear architecture and/or the complex spatial arrangement of genes, gene clusters, chromatin, and regulatory DNA elements [219, 220] is beyond the scope of this research and hence we only focus on the sequence level aiming at discovering cis-regulatory elements on the proximal promoters.

Two of the most important functional elements in gene regulation are transcription factors and their binding sites on the promoters of their target genes. A TF is a protein which binds to specific DNA binding motifs that can be present multiple times on the same promoter of a gene or on different promoters of different genes. The transcription factor binding sites where a TF binds are usually short (5-15bp) and degenerate but highly selective through evolution [221]. A gene can have multiple alternative promoters [222, 223] and each promoter frequently contains a large number of binding sites (10 – 50 binding sites) for 5 – 15 different TFs [224]. Therefore, a more comprehensive understanding of these elements and their interactions will provide a deeper understanding of the regulatory pathways within cells and potential functions of individual genes and/or gene clusters [225].

## 4.2. Gene promoter structure

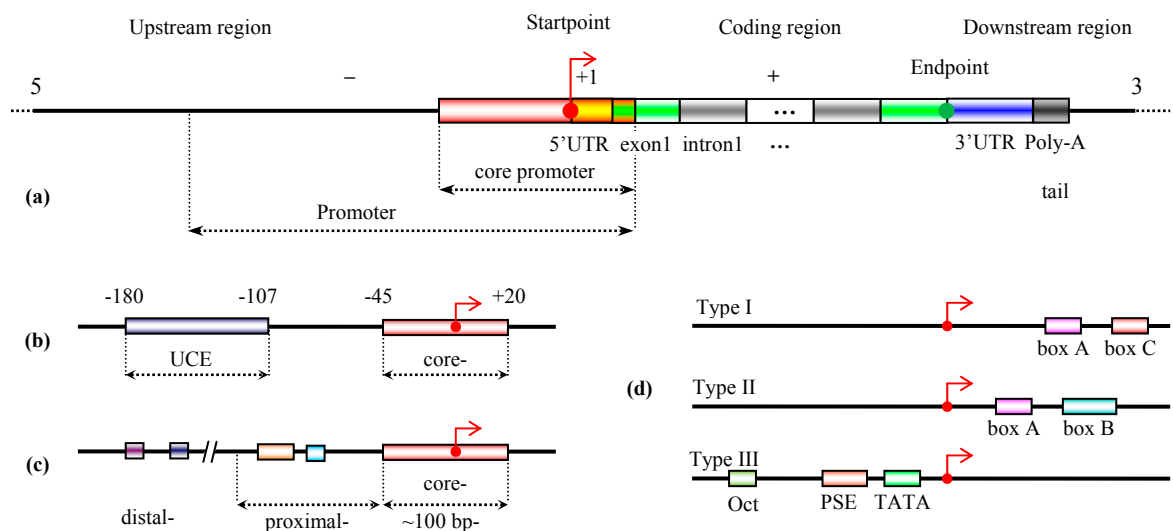
### 4.2.1. Gene structure

Promoters are DNA sequences located upstream the coding region of each gene towards the 5' endpoint. Combined with other regulatory elements in the upstream region of a gene, these elements in the promoter region interact with transcription factors, recruit RNA polymerases, and then initiate the transcription of a gene. There are three classes of promoters that are recognized by three corresponding RNA polymerases (Figure 4.1):

- Class I promoters are made up of two regions, an upstream control element and a core promoter. They serve for the regulation of ribosomal RNAs synthesis (5.8S, 18S, and 28S rRNAs).
- Class II promoters are mainly involved in transcribing protein-coding genes which generate pre-mRNAs and almost all small nuclear RNAs (snRNAs). Each member of this class consists of a core promoter, proximal promoter elements, and distal regulatory elements.
- Class III promoters have three types: type I and II are internal promoters that regulate the synthesis of 5S rRNAs and tRNAs and interact with sites in the RNA polymerase. Type III promoters are upstream promoters similar to class II promoters and regulates the synthesis of some snRNAs or viral-associated RNAs [226].

**Figure 4.1:** Basic structure of promoter classes [226]. (a) A general structure of an eukaryote gene; the promoter region contains crucial regulatory elements to control the transcription of the gene; the gene is copied to a pre-mRNA from which the RNA Pol-II transcribes into an mRNA; the coding region contains alternatively exons and introns where introns are removed in the transcription process; a gene is marked by an integer 1D-coordinate system without zero point, i.e. TSS is +1 and before is negative; the un-

translated regions (UTRs) are particular sections of mRNA; the 5' UTR starts from the TSS and ends just before the start codon (usually AUG), the 3' UTR follows the coding region and ends before the poly-A tail – the sign to stop the transcription. (b)(c) Typical structures of class I promoters and class II promoters, respectively. (d) The typical structure of class III promoters; box A, B, C as well as TATA, PSE, Oct are conserved sequences which are bound by TFs to initialize the transcription process; internal promoters (Type I, II) have short conserved sequences located within the coding region; upstream promoters (Type III) contain short conserved sequences upstream of the start point.



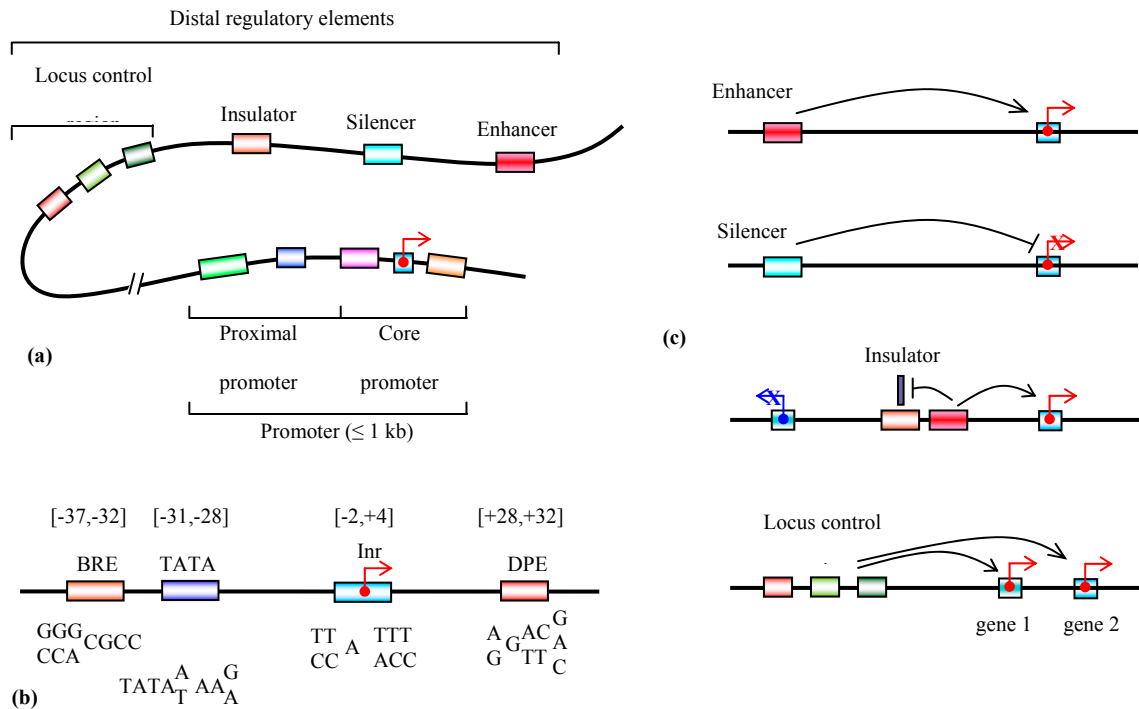
#### 4.2.2. Promoter elements

Although the process of gene expression is regulated at many levels e.g. genomic level, transcriptional level, RNA processing level, translational level, or post-translation level, promoter regions and regulatory elements are still considered as one of the most important factors [227]. Since proteins in eukaryotes are mostly transcribed by RNA polymerase II, computational promoter studies are mainly focused on protein-coding

genes, in this review we will concentrate on the structure of class II promoters (Figure 4.2a) which are characterized by the core promoter, proximal- and distal- promoter elements [228].

**Core promoter** is a small stretch sequence about 100bp flanking the transcription start site (TSS) which incorporate a combination of four common components consisting of the TATA box, initiator (Inr), TFIIB recognition element (BRE), and the downstream promoter element (DPE) [229, 230]. This serves for the initiation of the transcription process (Figure 4.2b). The TATA box, the binding site for TATA-binding protein (TBP), is a TA-rich site at 26-31bp upstream in higher eukaryotes and 40-120bp upstream in yeast [231]. Inr, also called the Transcription Start Site (TSS), is the start position located in the core promoter and functions similarly to the TATA box [230]. A comprehensive statistical analysis on a dataset with more than 10,000 human promoter from EPD [232, 233] and DBTSS [234, 235] demonstrated that it is not necessary for all these components to be simultaneously present in the core promoter [236]. Specifically, Inr elements are present in nearly half of the promoters whereas TATA boxes are present in only around 10% of the promoters in the dataset and seem to simultaneously present with the Inr elements. BRE and DPE elements are present about 25% of the time. Furthermore, the presence of DPE is independent of the presence of TATA-box and Inr elements whereas BRE-containing promoters are present in TATA-less promoters. Besides these elements, a number of other motifs in this region e.g. YY1, CAAT, CREB, etc. were also discovered in an analysis on a set of high-quality human core promoters [237].

**Proximal promoter elements** are located on the proximal promoter which is defined as the region up to 1Kbp upstream of the core promoter. The presence and importance of these cis-regulatory elements were characterized via a technique called linker-scanning mutagenesis [238] which showed that any mutation at one site in a regulatory element in this region can cause a significant change in transcription levels. Elements in the region between -350 and -40 have positive effects on the promoter activity whereas those in the region from -350 up to -1000 appear to have a negative regulation on the expression of genes [227].



**Figure 4.2:** Class II promoter structure and relevant regulatory elements; these are directly redrawn from ([228, 230]). (a) Typical regulatory elements of a gene including a core promoter, proximal promoter elements and distal regulatory elements; the promoter region which contains a core promoter and proximal promoter elements is usually no

longer than 1kb. (b) A detailed structure of a core promoter; the top is the positions of the conserved elements in the core promoter within the gene coordinate system; the bottom is the corresponding consensus sequences (c) Four typical types of distal regulatory elements and their corresponding effects; enhancers activate whereas silencers repress the transcription; insulators block the gene from being affected by other regulatory elements; a locus control region can affect the transcription of a number of genes.

***Distal regulatory elements*** are characterized by four regulatory groups (Figure 4.2c). Enhancers work as cis-regulatory elements near the TSS with the positive effects on promoter activity and in many cases, they both share the same activators [239]. Silencers are bound by repressors to negatively regulate the expression. The third group is insulators which are similar to a wall, preventing the mutual transcriptional effects of regulatory elements between neighbor genes. The last is a combination of different regulatory elements (known as locus control regions (LCRs) which regulate an entire locus or a number of genes [240]. These trans-regulatory elements function in the same way as cis-regulatory elements although they are located far from the TSS and work under the control of trans-acting factors[228].

#### **4.2.3. Promoters identification**

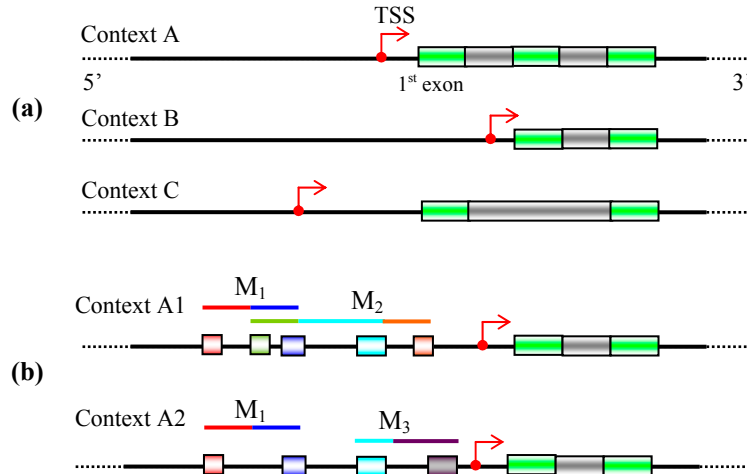
The first step towards discovering TFBSs is identifying the set of promoters. In principle, they are defined as the upstream regions proximal to the transcription start sites (TSSs) of genes; however, their length is still not clearly defined among different studies although it is one of the most important factors affecting to the computational predictions. Numerous activities have been proposed such as the recent experiment known as genome-wide open chromatin map that integrates high-throughput sequencing and genome-wide tiled array

technologies has been performed to identify DNase I hypersensitive sites within human primary CD4<sup>+</sup> T cells [241]. Such activities aim at better defining proximal promoter lengths which are subsequently incorporated in commercial tools, such as [242].

Besides experimentally identified promoters, a number of computational methods have been proposed to predict promoter regions. Available tools include PromoterInspector [243], DragonGSF [244], EnSemPro [245], and have all been thoroughly reviewed [246, 247]. Prediction tools can be classified into two main categories, signal-based approaches which rely on conserved signals relevant to promoters, e.g. TATA box, CAAT box, CpG islands, and content-based approaches that utilize conserved motifs to distinguish between promoters and non-promoter regions [248]. Several models have been shown to be promising but due to the complexities of the genome structure, large-scale predictions are still difficult [249].

The structure of promoters, especially in mammals, is a complex which can be considered as a mini-structure of a gene where regulatory elements are interspersed within a large number of regions non-conserved and unknown function [249]. Traditionally, it has been assumed that the combinatorial interaction of multiple transcription factors with the gene promoter is sufficient to explain the process of transcription. However, recent studies provided results to show that a large proportion of mammalian genes possess multiple transcription start sites (TSSs) and thus multiple promoters driving gene expression in a context-specific manner [250-252]. Specifically, in a recent study Singer et al. [223] developed and employed a custom microarray platform to show that nearly 35,000 alternative putative promoters are present on around 7,000 human genes. Furthermore, each set of unique combination of TFBSs in the

promoter will determine its temporal and spatial expression in a specific context [249] (Figure 4.3). These observations significantly increase the complexity of understanding gene regulation and the transcription process in general, and create a huge challenge for computational TFBS identification.



**Figure 4.3:** Data complexities in TFBS prediction. (a) Alternative promoters usually occur for genes in higher eukaryotes e.g. nearly 35,000 alternative putative promoters are present on around 7,000 human genes [223]. For a specific gene, different promoters are activated to drive the gene expression in different corresponding contexts. (b) Alternative sets of combinatorial TFs regulate the transcription process even though only one promoter is activated in these contexts. M<sub>1</sub>, M<sub>2</sub>, M<sub>3</sub> are three example transcriptional modules (a set of TFs or corresponding TFBSs) activated to regulate the transcription process; module M<sub>1</sub> is present on two cases whereas only a part of M<sub>2</sub> is functional in the other case e.g. human RANTES/CCL5 gene consists of different set of functional TFBSs in different cell types [249]. These complexities create a huge challenge for both computational and experimental in detecting functional binding sites for a specific context.



### 4.3. Binding site representation

Assuming a list of DNA binding sites for some TF is available, one of the very first questions is how to best represent and characterize the information contained in these sites for further analysis. The goal is to find a representation that matches as closely as possible all the binding sites in the collection and is clearly distinguished from the background. From the point of view of string processing, a simple and widely-used concept is the of consensus sequence in which the most frequent character at each position is chosen to represent the binding motif at that position. However, some positions might consist of characters of equivalent frequency and thus a more complex pattern, based on the IUPAC sequence [253, 254] was used to characterize the diversity of those binding sites (**Figure 4.4a**). Although this representation works well for highly conserved and short binding motifs, it is defined somewhat arbitrarily and removes much of the information in the original set of binding sites. In a case for yeast TF ABF1, for instance, two IUPAC sequences (RTCRYNNNNACG or RTCRYNNNNNACG) have been published and used as a relatively precise description of ABF1 binding sites [255]. However, these representations failed to recognize the binding site SCPK01 on PYK1 promoter from position -610 to -598 which was showed to be bound by TF ABF1 experimentally [256]. Consequently, a more precise representation was proposed to utilize almost all binding site information, known as the nucleotide distribution matrix or position weight matrix (PWM) [255, 257, 258], which has been proven very successful in various problems in DNA and protein sequence analysis [255, 259]. The PWM is a matrix of scores (e.g. occurrences, frequencies) with four rows corresponding to four DNA bases and m columns, each of which is a position in the binding motif. The basic

assumption of the PWM is that the base-pairs at different positions are statistically independent and thus the fitness score of a matched oligonucleotide 'p' with this profile is the sum of the fitness at each position. This representation reflects the extent to which a position is conserved within the binding motifs and thus the higher the similarity the higher the fitness.

The main weakness of the PWM approach stems from the assumption that the positions contribute independently and additively to the total activity of the binding site. However, position dependence may exist on the binding sites and has been experimentally and/or statistically verified in some cases [260]. For example, using a new quantitative multiple fluorescence relative affinity assay Man et al. [261] showed that position 16 and 17 on the operator DNA were not independent in the interactions with its TF, *Salmonella* bacteriophage repressor Mnt; or in another case, when Ellrott et al. [262] applied  $\chi^2$  test on the 71 binding sites of TF *hepatocyte nuclear factor 4 $\alpha$*  HNF4 $\alpha$ , a significant dependence was found between several pairs of positions e.g. position 4 and 8, 4 and 11. Therefore, more comprehensive representations were introduced to capture the potential dependence between positions in binding sites, such as maximal dependence decomposition [263], hidden Markov model [264, 265], Markov chain optimization [262], as well as a more flexible approach based on variable-order Bayesian network which combines PWM, Markov models and Bayesian model to fit with each particular subset of binding sites of a TF [266].

However, despite the limitations of the basic PWM approach, it is still the leading model in the search for discovering potential TFBSs. In fact, besides its intuitive representation and fast computation, it has been shown to be comparable at least, and in some case

outperforms, other more complicated models e.g. fixed-order Markov models that are usually over-fitted due to a limited training data [266]. Therefore, emphasis has been given to strategies that optimize the PWM instead of building more powerful models. For example, the scores in the cells of the matrix can be transformed to improve the specificity of the binding motif model (e.g. convert frequencies to probabilities, adding pseudo-count, taking logarithms, etc. [257, 267]) and the binding sites can be aligned before creating the PWM [255]. In some cases, the information content (IC) of the PWM, or some similar form, is be made use to select a suitable number of binding sites for

creating the binding motif model [112, 267, 268];  $IC = \sum_i \sum_{b \in \{A,C,G,T\}} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$  where  $f_{b,i}$

is the observed frequency of base  $b$  at position  $i$  and  $p_b$  is the background frequency of base  $b$  (usually 25% as neutral distribution across the genome is assumed).

In Figure 4.4, the top-left window is the collection of binding sites, each of which is called an oligo or conserved sequence; oligos can be aligned with gaps to maximize the motif content but in this case, it is a gap-free alignment. Several models have been displayed and lastly an advance model of PWM [112] is presented; the normalized formula is inferred from the original equation to ensure the rule that the fitness score of a matched oligo can be estimated by taking the sum of the fitness at each position. Thus, there are different formulae to normalize the raw PWM up to different studies. The ‘bold’ part is the core region of the binding sites i.e. the most conserved region in the binding motif model. Bottom-right is the sequence logo that can quickly visualize the specificity of the conserved information in each column.

**Figure 4.4:** Binding site representation. (a) Illustration of several motif models for human factor ETS1. (b) A brief look on the history of binding motif models.

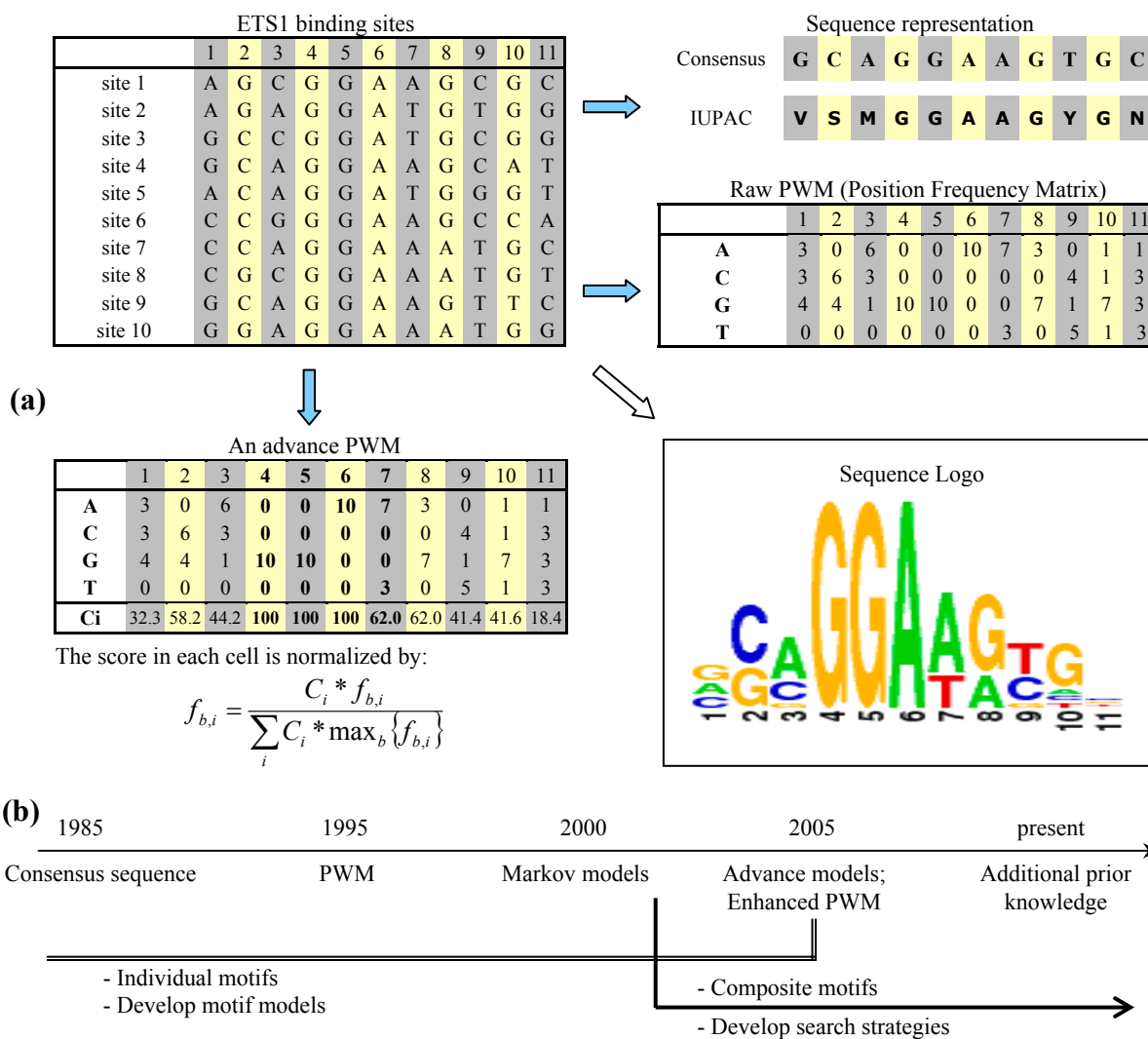


Figure 4.4b shows a brief look on the history of binding motif models. Starting from the first simple representation, consensus sequences, one has developed more advance models to characterize the binding motifs of TFs. However, due to the nature of the binding sites e.g. short, degenerate, etc., the problem has become a challenge and the developing strategies have been changed when applying to higher eukaryotes e.g. search

for composite motifs (a set of TFBSs) instead of single motifs, combine additional lines of biological evidence in detecting TFBSs (phylogeny, co-expression, and/or co-function). Additionally, other significant efforts have been devoted towards enhancing the power of the PWM in order to better discriminate between real binding sites and the background e.g. random data or non-regulatory regions. In this direction, Gershenzon et al. [269] proposed 16-row matrices to replace the 4-row PWMs; Sandelin et al. [270] tried to classify TFBSs into TF families based on the constrained binding sequence diversity for groups of structurally related TFs to create familial binding profiles; Hannenhalli et al. [271] computationally divided the binding site collection of a TF into two subsets corresponding to two-child PWMs to increase the binding specificity of TF profiles. As earlier noted, however, the short length of the binding sites makes them appear fairly redundant and predictive methods are often replete with false positives. Therefore, given that the main question concerns the actual identification of TFBSs and effective the location of the promoter, searching becomes a more critical issue than simply optimizing the representation.

#### 4.4. Discovery of ‘physical’ transcription factor binding sites

One of the first questions related to TFBS identification would be how to detect a conserved motif in a given set of sequences. The problem can be simply stated as follows: given a set of  $N$  sequences  $S = \{s_i\}_{i=1}^N$ ,  $s_i = \{s_{il}\}_{l=1}^{\|s_i\|}$ ,  $s_{il} \in A$ ,  $A = \{A, C, G, T\}$ , identify conserved motifs  $p = \{p_k\}_{k=1}^K$ ,  $p_k \in A$  that are overrepresented, i.e. motifs present in  $S$  at a statistical significant rate. The fundamental assumption is that if the sequences

are promoters of genes, then conserved motifs can be assumed to be potential binding sites for TFs.

There have been a wide range of possible applications for such *in silico* motif discovery methods. First, they greatly assist experimental studies aiming towards detection of the collection of binding sites for a given TF [272]. ChIP-chip assays, for example, identify genomic regions to which a TF of interest binds. However, locating exact sites where the TF binds might be very difficult due to the limitations of the assays. As a result, once the DNA sequences to which the TF binds have been collected motif discovery algorithms, e.g. consensus [273], Gibbs sampling [274], MEME [275], are then applied to locate the exact binding sites. Secondly, if one identifies a set of genes that can be considered as regulated by some common TF(s), then one can begin to search computationally for conserved motifs in the corresponding promoter to infer regulating TFs. The underlying assumption of such a computation is that the common patterns are the likely functional ones. Furthermore, motif discovery algorithms can also assist in cross-species extrapolation to improve the specificity of finding TFBSs on a gene promoter. Once a set of corresponding promoters of a gene across multiple species have been extracted, motif discovery algorithms are used to detect conserved sub-sequences in this promoter across species in an attempt to identify all potential cis-regulatory elements (discussed more details in the next section).

Because of the importance of this problem, a variety of algorithms as well as computational tools have been developed for those problems above for the past twenty years (Table 4.1). However, generally speaking the core algorithms can be classified into two categories: combinatorial and probabilistic [267, 276, 277]. Exhaustive search with

pattern-based scoring (combinatorial category) is the starting point of discovering conserved motifs in a set of promoter sequences [277]. Due to magnitude of the search space, methods were further improved by exploring sequence-based exhaustive search [278] and also consensus search [279]. The probabilistic-based methods employ two main algorithms e.g. Gibbs sampling [274] and MEME [275] and have also been used extensively for motif discovery tools. The basic idea is to continuously reduce the search space and the false positive matches by more accurately representing the motif models.

**Table 4.1:** Selected resources and relevant tools for *in silico* TFBS identification.

<b>Genome Browsers</b>			
UCSC	<a href="http://genome.ucsc.edu">genome.ucsc.edu</a>	VISTA	<a href="http://genome.lbl.gov/vista">http://genome.lbl.gov/vista</a>
<b>Promoter resources</b>			
<b>Databases</b>		<b>Prediction Tools</b>	
Genomatix	<a href="http://genomatix.de/products/Gene2Promoter">genomatix.de/products/Gene2Promoter</a>	PromoterInspector	<a href="http://genomatix.de/promoterinspector.html">genomatix.de/promoterinspector.html</a>
CSHL	<a href="http://rulai.cshl.edu/CSHLmpd2">rulai.cshl.edu/CSHLmpd2</a>	DragonGSF	<a href="http://research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm">research.i2r.a-star.edu.sg/promoter/dragonGSF1_0/genestart.htm</a>
DBTSS	<a href="http://dbtss.hgc.jp">dbtss.hgc.jp</a>	Eponine	<a href="http://www.sanger.ac.uk/Users/td2/eponine">www.sanger.ac.uk/Users/td2/eponine</a>
EPD	<a href="http://www.epd.isb-sib.ch">www.epd.isb-sib.ch</a>	FirstEF	<a href="http://rulai.cshl.org/tools/FirstEF">rulai.cshl.org/tools/FirstEF</a>
<b>Transcription factor resources</b>			
<b>PWM databases</b>		<b>Phylogenetic footprinting tools</b>	
Genomatix	<a href="http://genomatix.de/products/MatBase">genomatix.de/products/MatBase</a>	FootPrinter	<a href="http://bio.cs.washington.edu/software.html#footprinter">bio.cs.washington.edu/software.html#footprinter</a>
TRANSFAC	<a href="http://www.gene-regulation.com/pub/databases.html">www.gene-regulation.com/pub/databases.html</a>	PhyloME	<a href="http://bio.cs.washington.edu/software.html#phyyme">bio.cs.washington.edu/software.html#phyyme</a>
JASPAR	<a href="http://jaspar.cgb.ki.se">jaspar.cgb.ki.se</a>	PhyloGibbs	<a href="http://www.phylogibbs.unibas.ch/cgi-bin/phylogibbs.pl">www.phylogibbs.unibas.ch/cgi-bin/phylogibbs.pl</a>
		PhyloGibbs-MP	<a href="http://www.imsc.res.in/~rsidd/phylogibbs-mp">www.imsc.res.in/~rsidd/phylogibbs-mp</a>
		MONKEY	<a href="http://rana.lbl.gov/monkey">rana.lbl.gov/monkey</a>
<b>Single-motif discovery tools</b>		<b>Cis-regulatory module discovery tools</b>	
MatInspector	<a href="http://genomatix.de/products/MatInspector">genomatix.de/products/MatInspector</a>	FrameWorker	<a href="http://genomatix.de/frameworker.html">genomatix.de/frameworker.html</a>
P-Match	<a href="http://www.gene-regulation.com/pub/programs.html">www.gene-regulation.com/pub/programs.html</a>	CMA	<a href="http://www.gene-regulation.com/pub/programs.html">www.gene-regulation.com/pub/programs.html</a>
AlignACE	<a href="http://atlas.med.harvard.edu">atlas.med.harvard.edu</a>	CisModule	<a href="http://www.stat.ucla.edu/~zhou/CisModule">www.stat.ucla.edu/~zhou/CisModule</a>
Consensus	<a href="http://bifrost.wustl.edu/consensus">bifrost.wustl.edu/consensus</a>	CisPlusFinder	<a href="http://jakob.genetik.uni-koeln.de/bioinformatik/people/nora/nora.html">jakob.genetik.uni-koeln.de/bioinformatik/people/nora/nora.html</a>
MEME	<a href="http://meme.sdsc.edu">meme.sdsc.edu</a>	DiRE	<a href="http://dire.dcode.org">dire.dcode.org</a>

However, it is important to realize that although a large number of TFs has already been identified, and more are being identified, through numerous high-throughput activities emanating from the decoding of the human, *in silico* analysis is further hindered by the fact that only a fraction of those can currently be mapped to known and well characterized profiles [242, 280, 281] (around 600 human TFs in [www.genomatix.de](http://www.genomatix.de) vs. approximately 1,850 found TFs in human [282]). When conserved motifs are predicted computationally that are not present in available collections, these are then considered as novel binding sites and/or regulatory regions but they are set aside for further investigation. Therefore, besides such motif discovery methods, another approach to detect potential TFBSs is directly scanning known TF profiles and scoring to determine whether or not the matches are potential binding sites.

Given that the scoring metric would assign relative importance to alternative binding sites in motif discovery methods [283-285], it is of equal importance to score directly the subsequences of interest in terms of their potential of being binding sites compared to known TF profiles. Despite the large number of alternative representation models and their associated scoring function, the most widely-used approach is still the one based on the PWM model and the sum fitness function, as discussed above. Given, therefore, that the sum fitness is used, which based on the relative abundance of bases in a specific position based on scanning the TF profiles, the strategy to predict whether or not a site is a binding site is among the most critical factors. Therefore, major emphasis is placed on developing strategies that score a candidate oligo and identify the thresholds for the prediction. A typical approach is based on core similarity matches (Figure 4.4a) to reduce the number of false positive matches [112]. Furthermore, the threshold for each



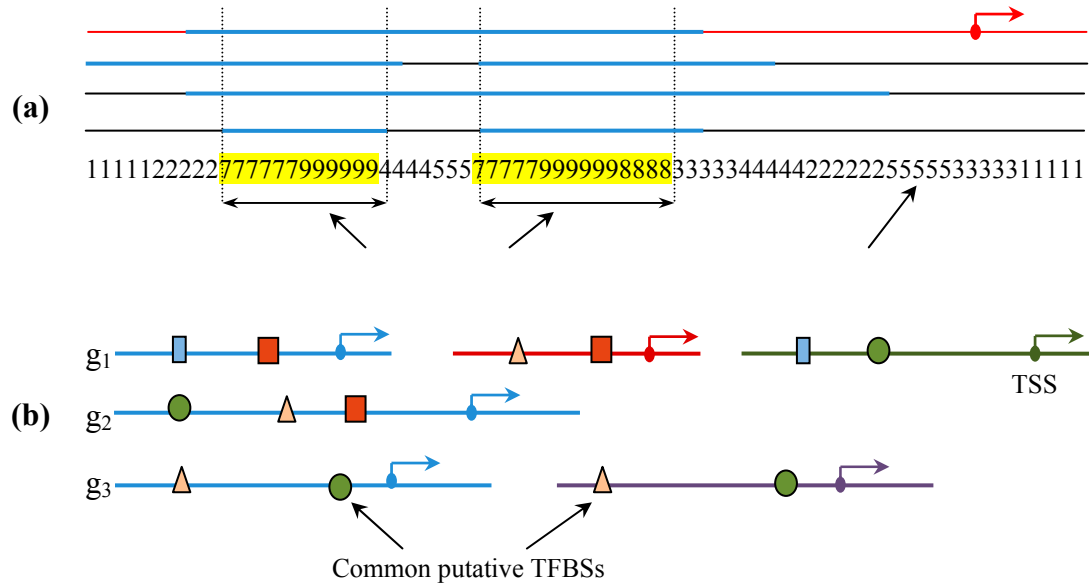
PWM is optimized so that a maximum of three matches are allowed in 10,000bp of non-regulatory test sequences (coding sequences excluding first exons and genomic repeats). This is the approach used in tool MatInspector in Genomatix [112]. As an alternative strategy, [268] implemented P-Match in TRANSFAC to select the optimized thresholds so that the false positive rate is minimum and/or the false negative rate reaches some user-defined threshold. The threshold for minimum false positive rate is the one at which no match is found on the background set of exon sequences; and the threshold for false negative rate  $\alpha$  is the rate at which  $\alpha\%$  of binding sites in the collection used to build the TF profile are not detected by that threshold using leave-one-out cross validation. Besides determining is the magnitude of a score threshold, both approaches also make use of the concept of TF family profiles [270, 271] with some variations to reduce the redundant matches in scanning TF profiles on a promoter sequence. Generally speaking, the key idea here is using prior knowledge such as known TF profiles to predict the most probable TFBSs on promoter sequences with a minimum false positive matches; for example, those PWMs that represent similar DNA patterns will be assigned into the same TF family [112].

#### **4.5. Phylogenetic footprinting**

The basic underlying assumption of comparative genomics, or phylogenetic footprinting, is that functional regions evolve under constraints and thus at a lower rate than non-functional regions. Therefore, it is hypothesized that well conserved regions in a set of orthologous sequences survived due to their special functional implications, making them become promising candidates for functional cis-regulatory elements [286]. Preliminary evidence seems to support the hypothesis that conservation does imply so kind of, yet to

be determined, significance. For instance, Cliften et al. [287] sequenced six *Saccharomyces* species and verified that many TFBSs are conserved across species and also located in conserved blocks although the blocks are often times much longer than the binding sites. Similarly, Gibbs et al. [288] demonstrated that regions with high-scoring PWM matches that are conserved across human-mouse-rat genomic alignment provided a 44-fold increase in the specificity of the predictions compared to those that are not conserved. Therefore, utilizing the information from orthologous genes across multiple species is becoming a useful paradigm in predicting putatively functional binding sites as well as reducing the false positive matches in motif discovering methods.

Given a set of genes, in order to identify conserved regions for each promoter DIALIGN [289] was used to perform multiple sequence alignments with the input sequences including each sequence as well as its orthologous promoters. DIALIGN was selected because it has many applications in comparative genomics [290]. Also, a recent benchmark study for the alignment of non-coding DNA sequences has concluded that it can produce alignments with high sensitivity as well as specificity to detect constrained sites [291]. Following the alignment among orthologous promoter sequences, we relied on the conserved scores returning from DIALIGN (with the similarity threshold of diagonals or corresponding segments involved at least 5 bases) to locate conserved regions which are defined as sub-sequences that are longer than 10bp and continuously scored greater than the average score of all the alignment's conserved scores (Figure 4.5a).



**Figure 4.5:** Identification of promoter conserved regions and common physical TFBSs.

(a) Estimation of conserved regions on a single promoter (the red one) based on Dialign's alignment scores from a set of orthologous promoters. (b) Finding common physical TFBSs accounting for the case that genes may have multiple alternative promoters. TFBSs present on the conserved regions of any alternative promoter of a gene are also considered as putative TFBSs for that gene.

We next apply MatInspector [112] to scan for all physical TFBSs and only those that overlap with the conserved regions selected above are kept for further analysis. We used a common core similarity 0.75 and utilized the optimal matrix similarity threshold for each position weight matrix (a corresponding profile of TFBSs) suggested from MatBase, Genomatix [111] which ensure that a minimum number of matches are found in non-regulatory sequences i.e. the false positive matches is minimized. However, a gene may have multiple alternative promoters [223] and virtually in all cases, it is not known which promoter of the gene is activated. Therefore, all putative TFBSs detected from all

alternative promoters of a gene are considered as candidates to infer putative transcriptional regulators for the gene. Subsequently, we estimate the common level of each candidate above in each corresponding module and select those TFBSs present more than a common threshold (70% in this study) (Figure 4.5b). Associated TF families with those selected TFBSs are inferred and considered as transcriptional regulators.

#### **4.6. Context-specific transcriptional regulators**

While it is recognized that not all binding sites found on a promoter will be functional elements, it is also recognized that functional sites are not activated simultaneously or independently of condition, or environment, since the cooperation of TFs is highly dependent on context [292-296]. Human RANTES/CCL5, a member of the CC- or  $\beta$ -subfamily chemotactic cytokines for instance, appears to have six functionally characterized short regulatory elements on its promoter that mediate its transcription initiation. However, not all six elements are activated simultaneously in any specific tissue in five cell types analyzed and the elements are also highly selective under different stimulating signals regulating gene expression [297]. Consequently, a critical question is to establish a relationship between binding sites and the context in which these sites become functional. The term ‘context’ here is used in a way that implicitly refers to a set of potentially co-regulated genes e.g. genes that appear either to exhibit correlation in their expression patterns and/or to be involved in similar functions in a specific condition and/or tissue [293, 298, 299].

The main idea in this direction is to use prior knowledge to identify the set of potentially co-regulated genes and then search the corresponding promoter set for common and/or significant cis-regulatory modules. Earlier studies assumed that a cluster of coexpressed

genes could be set under the same regulatory mechanism, e.g. co-regulation [300, 301] or co-function [302]. However, more recent evidence suggests that co-expression alone is not enough to infer the existence of common regulatory mechanisms and instead additional information is required [298, 303], especially in higher organisms. Specifically, recent studies have shown that genes sharing similar expression patterns can participate in a number of different biological functions and/or genes in the same pathway can exhibit different patterns of expression [304, 305]. Moreover, the underlying gene regulation is shown to be tissue and/or condition specific and the TFs that drive the gene expression are very flexible in function and activity under different conditions [293-296]. Therefore, defining the context in which a set of genes are more likely to be co-regulated poses a formidable challenge to researchers.

As such it is more appropriate to explore the concept of ‘gene battery’ originally proposed by Britten and Davidson [306] and has been further explored in the literature [307-310]. A gene battery refers to a group of genes that are coordinately expressed and/or functionally coupled since their regulatory regions respond to the same transcriptional signals [311, 312]. With the assumption that genes in a gene battery are involved in key biological processes, recognized CRMs will consist of putative functional binding sites that are associated with essential transcriptional regulators. Yet, in higher eukaryotes especially in humans the problem turns to be much more difficult. One of the most critical issues is to determine which genes belong to the same gene battery. Prior studies assume that either coexpressed genes [300, 301, 313] or genes that belong to the same biological process [299, 314] could be governed by some common regulatory mechanism. However, recent evidence suggests that co-expression or co-function alone is

not sufficient to infer the existence of common regulatory mechanisms [298, 315]. Oftentimes co-expressed genes can participate in a diverse array of biological functions while functionally-relevant genes can be characterized by different expression patterns [304, 305]. Predicated upon these, in this study we explore the possibility that genes that are both co-expressed and functionally-relevant may be more likely to be co-regulated. Since genes within the same pathway encode for a set of interacting proteins, they are more likely to be governed by some common regulatory mechanism [316]. Therefore, the unifying hypothesis of this study is that genes that participate in the same pathway are functional relevant.

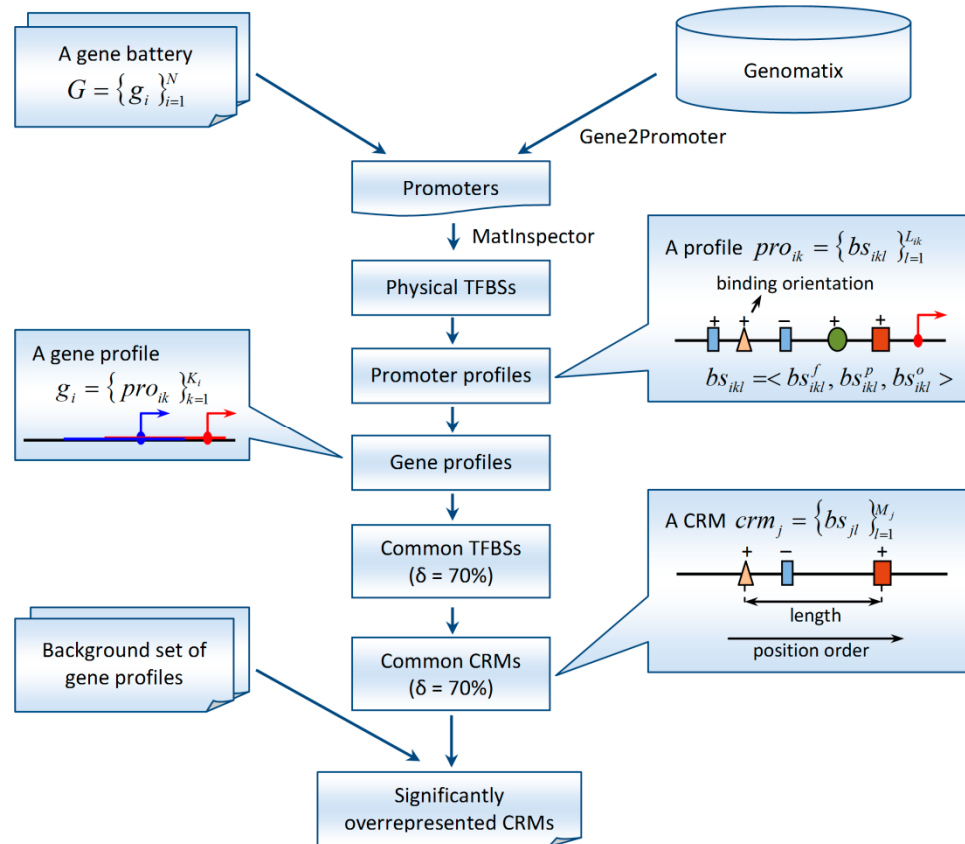
#### 4.6.1. *cis*-regulatory modules (CRMs)

Due to the fact that TFs in higher organisms regulate gene expression in a combinatorial manner rather than in isolation [292, 297] and that TFBSs tend to form clusters of binding sites, known as *cis*-regulatory modules (CRMs) [317, 318], computational methods have shifted towards discovering CRMs instead of a single TFBS. A *cis*-regulatory module is generally considered as the smallest functional regulatory unit [111]. From a computational standpoint, such module is mainly characterized by two factors: (i) composition which consists of a set of non-overlapping binding sites of TFs on the control regions of a gene and (ii) structural constraints that take into account the strand orientation to which TFs bind, the order and the distance between successive binding sites [319, 320].

The problem of CRM searching can be formalized as follows: given a set of  $N$  putatively coregulated genes  $G = \{g_i\}_{i=1}^N$ , each of which contains  $K_i$  alternative promoters

$g_i = \{pro_{ik}\}_{k=1}^{K_i}$  whereas each promoter is represented by a list of  $L_{ik}$  binding sites

(‘promoter profiles’)  $pro_{ik} = \{bs_{ikl}\}_{l=1}^{L_{ik}}$  and each binding site is a 3-tuple of corresponding transcription factor name  $f$ , position  $p$  and binding orientation  $o$ :  $bs_{ikl} = \langle bs_{ikl}^f, bs_{ikl}^p, bs_{ikl}^o \rangle$ , find a set of  $M$  *cis*-regulatory modules (CRMs)  $cCRM = \{crm_j\}_{j=1}^M$ ,  $crm_j = \{bs_{jl}\}_{l=1}^{M_j}$  that are present as common over a threshold  $\delta$  (70% in this study) on the set of gene promoters ( $M_j$  is the number of binding sites, yet to be determined, in CRM  $crm_j$ ). The statistical significance of each commonly recognized CRM vs. a background set of genes is then estimated selecting only significant CRMs. The subscripts  $i, k, l, j$  indicate the gene number, the promoter number, the binding site number, and the CRM number respectively. An illustration of the computational framework is presented in **Figure 4.6** while more details are discussed in the following section.



**Figure 4.6:** Flowchart of the CRM searching process. Each binding site is characterized by the TF name, position and binding strand orientation (+: forward and -: backward). Promoter sequences are converted into promoter profiles to speed up the calculation. A gene profile contains a set of promoter profiles that are corresponding to a set of alternative promoters of that gene. The background set contains 5,000 randomly selected human genes.

#### 4.6.2. Discovery of TFBSs and promoter profiles

Based on a comprehensive database of promoters – Genomatix [111], a set of transcript-relevant promoters are extracted coupled with multiple alternative promoters and experimental information about the promoter length including those with either an experimentally defined length or a default if there is no associated prior length information (500bp upstream plus 100bp downstream the TSSs). MatInspector [112] is then applied to scan for PWM matches on those promoter sequences using optimal parameters from MatBase [111]. In order to speed up the process of discovering CRMs as outlined in [321], each promoter is re-modelled with a list of  $L_{ik}$  TFBSs ordered by their local positions on the promoter sequences and represented by the corresponding TF name (e.g. NFkB, ETSF) along with the binding orientation  $pro_{ik} = \{bs_{ikl}\}_{l=1}^{L_{ik}}$ . The conversion aims to answer two basic questions: (i) given a promoter sequence, identify whether a TFBS or a CRM is present on this promoter or not, and (ii) given a gene with  $K_i$  alternative promoters, determine if a TFBS or a CRM is present on any promoter sequence of this gene. From a computational standpoint each promoter profile is loaded into a hash table whose field ‘key’ includes the TFBS name plus the binding orientation (e.g. +ETSF, -PAX6, ‘+’ as forward and ‘-’ as backward binding orientation) and field



‘value’ is the position list of the corresponding TFBS with the same binding orientation. For example, if the key is ‘+ETSF’ and the corresponding value is ‘373\_\_386’, we know that transcription factor ETSF is forward binding to the promoter at the local position -373 or -386 upstream. As a result, to decide the existence of a TFBS including the binding orientation on a promoter the process only makes a quick search in the hash keys. In a similar way, to determine the present of a CRM on a promoter the process will take into account the binding orientation from the keys and the positions from the values of corresponding keys to evaluate the structural constraints.

#### **4.6.3. Common cis-regulatory modules**

Computationally, a *cis*-regulatory module  $crm_j$  is a list of  $M_j$  non-overlapping TFBSs ordered by their positions on the promoter sequence and characterized with their corresponding binding strand orientation. For example, CRM ‘+NFKB\_\_CREB\_\_SP1F’ consists of three successive TFBSs of transcription factors NFKB, CREB, SP1F with the binding strand orientation forward, backward, and backward respectively. Besides the binding orientation and the position order of TFBSs, CRMs are also characterized by their length. If CRM A appears to be common in a gene battery of  $N$  genes, the average length of all instances of A on  $N$  genes is considered to be the length of this CRM. In the case that A presents more than one time on promoters of gene  $i$ , the length of instance A for this gene will be the minimum one. Subsequently, to estimate the common level of this CRM we only take into account those instances with the length approximate to the average one (e.g. from the half to the double). If the number of such instances over  $N$  is higher than a frequency threshold ( $\delta = 70\%$  in this study), CRM A is considered as a common CRM of the gene battery.

However, a gene can have multiple alternative promoters and virtually in all cases, it is not known which promoter of the gene is activated. To identify activation of putative promoters, one solution would be to search for all possible combinations of promoters in the gene set. Yet given a set of  $N$  genes, each gene with  $K$  alternative promoters in average, the total combinatorial number of promoter sets is  $K^N$  which is computationally intense and sometimes impossible to search for all promoter combinations. Consequently, we propose a novel heuristic *where if a TFBS or a CRM is present on any promoter sequence of a gene, it is considered as present on the control regions of that gene*. The heuristic results in one-time searching instead of  $K^N$  but still produces the same results as the brute-force search in all combinations of promoters (see Appendix S1 and Algorithms S1, procedure ‘IsPresent’). Using this heuristic, the main algorithm to search for common CRMs in a gene battery, similar to FrameWorker [320], can be simply described with two primary steps as follows: (1) identify all potential TFBSs that are common in a gene battery and (2) employ the breadth first search technique to search for all possible combination of all commonly found TFBSs in step 1, each of which is a potentially common CRM yet to be determined quickly by the heuristic above.

#### **4.6.4. Statistical significance of CRMs**

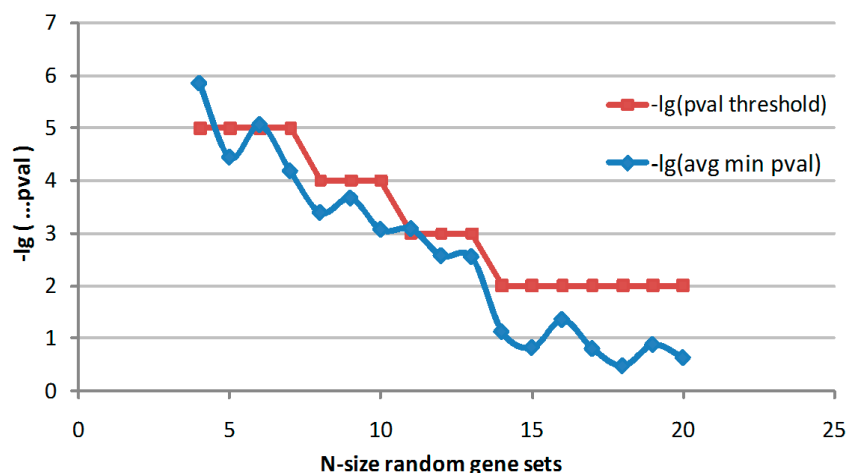
Within a gene battery, CRMs that are present on the control regions of corresponding genes above a frequency threshold (e.g.  $\delta = 70\%$  of the number of genes) are considered as common CRMs. However, such CRMs can also be overrepresented in random gene sets. Therefore, in order to restrict the false positive matches and increase the statistical power of our method, we estimate the hyper-geometric p-values of common CRMs vs. a background set and only select those CRMs whose p-values exceed a pre-defined

statistically significance threshold (e.g.  $10^{-4}$ ). However, this threshold is very sensitive to the size of the gene battery and thus a uniform significance threshold cannot be applied for all gene batteries. As a result, we developed a heuristic procedure for estimating the significance threshold of common CRMs with respect to the size of gene batteries. The procedure is repeated 100 times for each N-size gene-set ( $N = 4, 5, \dots, 20$ ). At each iteration, the algorithm randomly selects N genes from the background set, searches for common CRMs that are present on the promoters of these genes ( $\delta = 0.7$ ), estimates the statistical significance (p-values) for each CRM (see materials and methods), and records the minimum one. In this study, we choose the approximate values of the mean of these minimum p-values to set the criterion for the statistical significance of CRMs in a gene battery with size N (**Figure 4.7**). Consequently, for each gene battery only those CRMs that are identified with p-values less than the corresponding p-value thresholds are used to infer relevant transcription factors.

The hyper-geometric p-value defined as follows:

$$p\text{-value}(CRM_A) = \frac{\binom{b}{n} \binom{B-b}{N-n}}{\binom{B}{N}}$$

where B and b is the number of genes and the number of hits respectively in the background set which is made up of 5,000 randomly selected genes in genome of corresponding species; N and n is the number of genes and hits in the gene battery, respectively.



**Figure 4.7:** Statistical significance thresholds of CRMs. The procedure randomly picks a gene-set with  $N$  genes from the background and search for common CRMs ( $\delta = 0.7$ ) in that gene-set. The statistical significant p-value for each CRM is estimated and the minimum one is reported. Each point in the blue curve is a transformed value of the mean of minimum p-values of CRMs in 100 times repeat the procedure. Approximately, the red curve shows which thresholds should be used for the non-random cases. After  $N = 14$  genes, only one threshold is used to ensure the significance (p-value = 0.01).

#### 4.6.5. Other relevant issues

The most critical issue is that a large proportion of mammalian genes possess multiple transcription start sites (TSSs) and therefore multiple alternative promoters regulate gene expression in a context-specific manner [250-252]. For instance, in a recent study Singer et al. [223] developed and employed a custom microarray platform to show that there are nearly 35,000 alternative putative promoters present on around 7,000 human genes. As a result, the computational identification of CRMs becomes a combinatorial problem and oftentimes a daunting task due to the large number of alternative promoters of genes in the gene battery. For example, 7 genes that belong to Apoptosis pathway and late-up

expression pattern can produce totally 5,600 combinatorial promoter sets; or 10 genes that are in Cytokine-cytokine pathway and late-up expression pattern can create 13,440 combinatorial promoter sets; while complexity further increases in the oxidative phosphorylation group (down expression pattern) characterized by 40 genes and 1,274,019,840 combinatorial promoter sets. Consequently, searching for common CRMs in all promoter combinations is computationally intense. Yet, our novelty heuristic can reduce these complexities into only one running time but still preserve the same result (see appendix, lemma 1). In a similar manner, the strategy of converting promoter sequences into promoter profiles also makes the estimation of the significance of common CRMs vs. a large background set more computationally tractable [321].

Additionally, since it is not clear how long the promoter length should be, our computational analysis extracts highly qualitatively defined promoters from Genomatix databases [111] including those with either an experimentally defined length or a default if there is no associated prior length information. This default length (500bp upstream plus 100bp downstream the TSSs) is also supported from a recent experiment known as genome-wide open chromatin map [241]. Additionally, we also examined how the promoter length affects the *in silico* inference of CRMs. Specifically, we count the number of relevant TFs that can be considered as transcriptional regulators for the group of 8 genes that belong to the middle-up expression pattern and the apoptosis pathway. For each specific length of extracted promoters (27 promoters that are relevant transcripts; 100\*x upstream and 20\*x downstream bases, x from 4 to 10), we applied the same procedure to search for statistically significant CRMs and then infer the list of relevant TFs. The results show that the number of relevant TFs increases linearly with respect to

increasing promoter lengths (see Data S1, sheet 'Promoter length'). Thus, including prior information of the promoter lengths is very important to provide reliable computational predictions.

Another important challenge in computationally identifying TFs is associated with the fact that transcription factors can bind to regions far from the TSSs. For example, the P53 factor is a well established regulator for the programmed cell death (apoptosis) [322, 323]; however such regulator is not identified as putative TF in the gene batteries relevant to apoptosis pathway. However, if we increase the promoter length up to approximately 1,000bp P53 is identified within the statistically significant CRMs. This leads to the hypothesis that P53 might work in a cooperative manner with other TFs that bind to the distant promoter regions. Alternatively, it has been recognized that P53 can affect apoptosis via novel transcription-independent pathways although its role as a mediator of transcription is well established [324-326]. For instance, apoptosis can still occur when P53 mutants incapable of acting as transcription regulator are introduced [327, 328]. This indicates the possibility that P53 might not directly regulate the apoptotic gene batteries as identified from our analysis. Thus, computational missing P53 as a relevant TF may be a reasonable result rather than a limitation from our computational analysis; yet, it is still a question to us in this study. However, since our analysis only searches for CRMs on the proximal promoters of genes, it should be acknowledged that we may miss some relevant transcription factors that bind to the regions far from the TSSs as well as enhancers that regulate the transcriptional process.

## **4.7. Putative transcriptional regulatory program**

### **4.7.1. Results from corticosteroids pharmacogenomics model**

It has been widely accepted that after corticosteroids bind to cytosolic glucocorticoid receptors (GR), the activated steroid-receptor complex is rapidly translocated into the nucleus where it can alter the expression of target genes. However, the drug seems to be cleared within about 6h following a bolus injection, suggesting that the mRNA levels of CS-target genes will return to the base line after that [11]. In the contrast, the drug will reach and remain to a stable steady state after 6h in the chronic administration. Yet, the GR is greatly diminished in response to corticosteroids [14, 15, 19, 20], suggesting that the mRNA levels of CS-target genes in the chronic regimen should also return to the base line. This mechanism is corresponding to the first-phase regulation of target genes. However, almost all chronic patterns involve two phases of regulation and some (module 3 & 5) are only half-phase patterns i.e. persistent up or down without returning to the baseline. These complexities in expression patterns of CS-target genes can be explained by a number of possibilities previous studies have shown [11, 12], including multiple GR isoforms, multiple GREs with different affinities to the drug receptor complex, or some other receptors that can mediate the effect of corticosteroids and thus affected genes in this case can reach a new steady state in the presence of the drug.

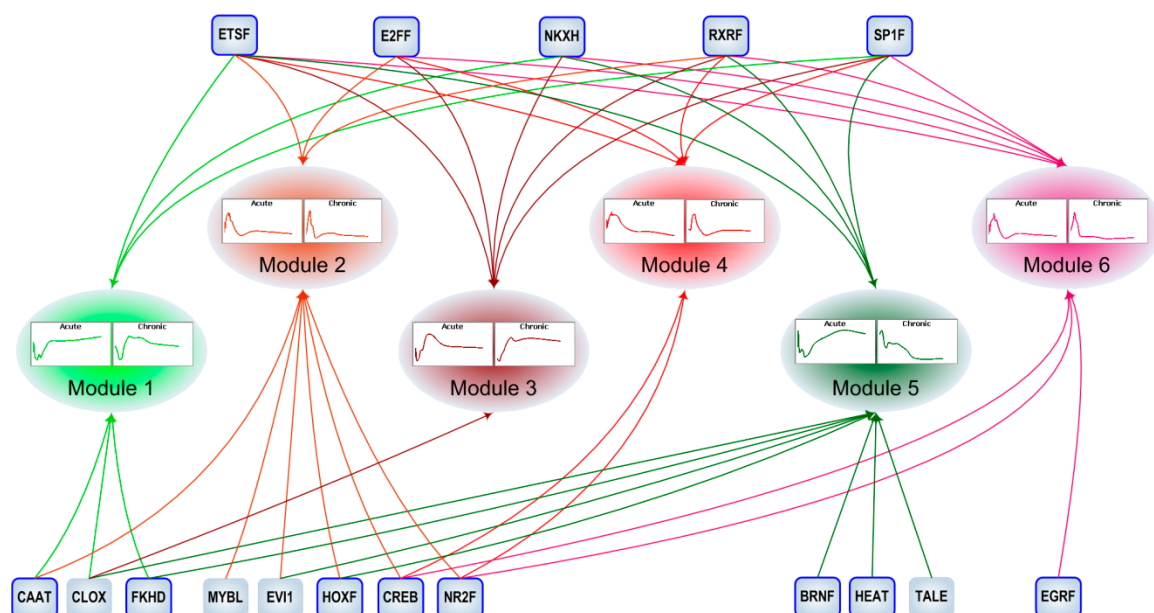
However, another possibility is a mechanism that results in the regulation of secondary biosignals which transcription factors are the most potential factors. After affected by corticosteroids, they in turn further modulate the expression of glucocorticoid-regulated genes as a continuing cascade of events that were initiated by the drug. As a result, this possibility suggests a possible interpretation of the complexities in expression changes of

multiple CS-target genes with the second phase of regulation (e.g. module 1, 2, 4, and 6). In order to reveal some underlying regulatory mechanism of these selected transcriptional modules, we start analyzing the promoter regions of genes to search for significant putative transcriptional regulators as well as possible relationships of regulation. The hypothesis we explore here is that if two or more genes have similar temporal profiles in response to multiple dosing regimens, they are more likely to share some common regulatory mechanisms.

For the 315 genes in six transcriptional modules, we extract 817 *Rattus norvegicus*'s promoter sequences, of which we only keep 194 genes with 502 promoters that include sufficient information of orthologous promoters for further analysis. Figure 7 shows the identified putative regulation between TF families and transcriptional modules. This finding highlights the possibility that secondary biosignals are involved in the regulatory complexities of expression changes for CS-affected genes. Almost all suggested TF families do consist of transcription factor members that are recognized as differentially expressed genes in one or both dosing regimens (see supplementary files – ‘functional\_characterization.xls’). Since transcription factors are characterized by pleiotropic effects, it is reasonable to observe a significant overlap across various transcriptional modules [329]. While comparing these regulatory combinations, we observe that some TF families seem to be common regulators for all modules (on the top of the figure) whereas some are very specific to particular modules (in the bottom of the figure). This could possibly explain the difference among the expression patterns of these modules. It is likely that the more similar the expression pattern of clusters the more likely they share a larger fraction of common regulators, e.g. TF families in this case. For



example, there are a large number of transcriptional regulators that are common between modules 2, 4 and 6 but it seems little overlap exists between the transcriptional regulators of modules 1 and 4, 1 and 6, 2 and 3, except common regulators on the top of the figure.



**Figure 4.8:** Putative regulation of CS transcriptional modules by enriched TFBSs. Those TF families with ‘blue’ border lines consist of transcription factors that are affected under corticosteroid administration in this study. The results show a putatively dynamic perspective of regulation between transcriptional regulators and involved sets of genes.

It has been also noticed that genes affected by CS include both immunosuppressive genes and metabolic genes. Upon the identification of putative transcriptional regulators, their relevance to immune response is demonstrated based on current literature evidence. Specifically, nine among the 29 recognized ETS transcription factors are known to regulate genes involved in immunity [330]; forkhead transcription factors (FKHD) play a major role in the control of apoptosis [331]; and especially CREB has been showed as an essential factor for interactions of glucocorticoid receptors to mediate gene expression [332, 333]. A number of others are overlapped with earlier *in silico* studies e.g. E2FF,

EGRF, HOXF, NKXH, SP1F [334]. However, given that the experiment of corticosteroid administration has been studied on normal rats, the relevance to adverse effects may be more important than the relevance to immune response. In fact, almost all enriched functions (gene ontologies, pathways) in these transcriptional modules are relevant to metabolic side-effects (see discussion below). Also, due to this reason NFkB and Ap-1 families widely considered as factors involved in inflammation are not present as direct transcriptional regulators for these sets of genes. Furthermore, we identify a number of transcriptional regulators known to be critical factors in metabolic syndrome including obesity, dyslipidemia, hypertension, insulin resistance, etc. e.g. RXRF [335], FKHD [336], SP1F [337]. For instance, the deletion of RXR in mouse liver results in abnormalities of all metabolic pathways regulated by retinoid X receptors heterodimers [338]; FoxOs, members of FKHD family, are able to increase hepatic glucose production, decrease insulin secretion, and affect glucose or lipid metabolism [336].

#### **4.7.2. Results from human endotoxemia model**

##### ***Identification of sets of ‘hypothetically’ co-regulated genes***

Upon identification of four significant patterns of gene expression, a number of inflammation-specific pathways are selected by evaluating the enrichment of corresponding subsets in inflammation-specific pathways, including Toll-like receptor signaling, Cytokine-Cytokine receptor interaction, Apoptosis and JAK-STAT signaling cascade, etc. (**Table 4.2**). It is now well established that Toll like receptor signaling pathway is the first arm of the host defence system that is activated when endotoxin is recognized by pathogen recognition receptors [339]. During the recognition process, LPS binds and interacts with its signaling receptor (TLR4) which triggers a signal transduction

cascade essential for the up-regulation of several pro-inflammatory mediators [340]. Such mediators including cytokines and chemokines interact with their appropriate receptors, giving rise to the Cytokine-Cytokine receptor signaling pathway that amplifies and propagates the inflammatory reaction throughout the cell until the system restores homeostasis [341]. Therefore, both Toll like receptor signaling and Cytokine-Cytokine receptor interaction pathways play a pivotal role in the pro-inflammatory response. Complementary to this, considerable attention has been given to the role of an excessive death of immune effector cells (apoptotic cells) during the progression of an aberrant inflammatory response [209]. The nature of apoptosis as a rectifying process has led researchers to the realization that identifying mediators that are critical in regulating the apoptotic-inflammatory imbalance might prove beneficial in treating human sepsis [342]. It is therefore reasonable to assume that apoptosis also plays a critical role in the endotoxin-induced inflammatory process. Along similar lines, JAK-STAT cascade is another highly enriched inflammation-specific pathway that exerts anti-inflammatory properties. Accordingly, recent data provide evidence that a STAT pathway from a receptor signaling system is a major determinant of key regulatory systems including feedback loops such as SOCS induction which subsequently suppresses the early induced Toll like receptor and cytokine signaling [210, 343]. Endotoxin-induced inflammation also causes a widespread suppression at the transcriptional response level of genes involved in mitochondrial energy production (Oxidative phosphorylation) and protein synthesis machinery (Ribosome). Such dysregulation in leukocyte bioenergetics together with persistent decrease in mitochondrial activity can lead to reduced cellular metabolism

[344], resulting in the participation of a number of critical metabolic pathways, e.g. Citrate cycle, Pyrimidine and Pyruvate metabolism.

**Table 4.2:** Data information and inflammation-relevant significant functions

Expression data (3,269 probesets <sup>+</sup> )			Relevant significant functions (p-value<0.05)	
Patterns	# of probesets (Total: 1703)	# of genes* (Total: 1213)	Pathways (KEGG)	Corresponding selected genes
Early-up	182	141	Apoptosis <sup>1</sup>	il1a, il1b, nfkb1a, tnfr
			Cytokine-cytokine receptor interaction <sup>1</sup>	ccl20, ccl4, cxcl1, cxcl2, il1a, il1b, il8, inhbb, tnfr
			Toll-like receptor signaling pathway <sup>1</sup>	ccl4, il1b, il8, map2k6, nfkb1a, tnfr
Middle-up	119	88	Apoptosis <sup>1</sup>	casp10, cflar, fas, irak3, myd88, nfkb1, nfkb2, rela
			Toll-like receptor signaling pathway <sup>1</sup>	myd88, nfkb1, nfkb2, rela
Late-up	284	185	Apoptosis <sup>1</sup>	casp8, il1r1, il1rap, irak4, pik3cg, tnfrsf10c, tnfrsf10
			Cytokine-cytokine receptor interaction <sup>1</sup>	ccr1, csf3r, il10rb, il13ra1, il1r1, il1rap, il8ra, il8rb, tnfrsf10c, tnfrsf10
			Toll-like receptor signaling pathway <sup>1</sup>	casp8, irak4, pik3cg, tlr1, tlr5, tlr8
			Jak-STAT signaling pathway <sup>1</sup>	csf3r, il10rb, il13ra1, pik3cg, stat2, stat5b
Down	1118	799	Citrate cycle (TCA cycle) <sup>2</sup>	acly, idh2, idh3a, mdh1, mdh2, suclg2
			Pyrimidine metabolism <sup>2</sup>	dck, dctd, dut, entpd6, pole3, polr2b, polr2e, polr2k, rpa1, uckl1
			Pyruvate metabolism <sup>2</sup>	akr1b1, glo1, ldhb, mdh1, mdh2, pdhb
			Ribosome <sup>1</sup>	fau, rpl10a, rpl12, rpl13a, rpl14, rpl18, rpl24, rpl27, rpl27a, rpl29, rpl3, rpl36a, rpl36al, rpl37a, rpl38, rpl8, rps2, rps24, rps7, rps9
			Oxidative phosphorylation <sup>2</sup>	atp5a1, atp5b, atp5f1, atp5g1, atp5g2, atp5g3, atp5h, atp5i, atp5j2, atp5l, atp5o, atp61f, cox4i1, cox5a, cox6c, cox7c, cyc1, nduf1, ndufa13, ndufa3, ndufa4, ndufa5, ndufa6, ndufab1, ndufb2, ndufb4, ndufb5, ndufb8, ndufc2, ndufs4, ndufs5, ndufs6, ndufs7, ndufs8, ppa2, ucr, uqcrb, uqcr2, uqcrh, uqcrq

\*: 3,269 significantly differentially expressed probesets were selected by ANOVA (p-value < 10<sup>-4</sup>) from the total 44,924 probesets; <sup>+</sup>: the number of corresponding genes with promoter annotation in Genomatix;

<sup>1</sup>: regulatory pathways; <sup>2</sup>: metabolic pathways.

### ***Biological characterization of identified transcription factors***

Predicated upon the hypothesis that subsets of co-expressed genes involved in the same biological pathway are more likely to be co-regulated, their transcriptional regulators are computationally predicted (**Table 4.3**). There is considerable evidence indicating the inflammatory relevance of the aforementioned inferred transcription factors including MEF2 [345], GATA [346], OCT1 [347], FKHD [331], ETSF [330], IRFF [348], NFKB [349] and CREB [332]. Specifically, the myocyte enhancer factor 2 (MEF2) transcription factor plays a central role in the transmission of extracellular signals to the genome and in the activation of genetic programs that control cell differentiation, proliferation, survival and apoptosis [350]. In addition to this, MEF2 proteins serve as the endpoints for multiple inflammatory signaling pathways including mitogen-activated protein kinase signaling pathway (MAPK) and thereby confer signal-responsiveness to downstream target genes [351]. Furthermore, the octamer transcription factor -1 (OCT-1) has also been shown to function as a stress sensor modulating the activity of genes important for the cellular response to stress [352]. Although OCT-1 is a ubiquitous transcription factor, it has recently been demonstrated that it promotes cell survival signifying its involvement in the apoptosis signaling [353]. Additional studies [354] document the involvement of octamer binding transcription factors (OCT-1) in regulating the expression of TLR4 in humans; thus making it a critical regulator of Toll like receptor signaling. Furthermore, Forkhead Transcription Factors (FKHD) also play a major role in the control of apoptosis perhaps by affecting the transcription of the gene encoding FASL [355]. Since these regulators can be the substrate of the protein kinase B (Akt) preventing their nuclear translocation, it is expected that FKHD regulators promote cellular survival and thereby

control the apoptotic machinery [356]. Moreover, IFN regulatory factors (IRFF) are a family of transcription factors that regulate expression of various pro-inflammatory and anti-inflammatory genes. Research findings reveal a critical role for these interferon regulatory proteins in the control of apoptosis [357, 358] while it has become evident [359, 360] that such regulators are also essential for TLR gene expression including the trans-acting factors, IRF-1 and IRF-2. This implies that in addition to up-regulation of pro-inflammatory gene expression, TLR stimulation also results in modulation of TLR gene expression itself via interferon transcription factors.

One of the most important cellular factors involved in the regulation of the host innate immune response is the nuclear factor (NF)- $\kappa$ B which can be activated by a variety of stimuli including bacterial products, inflammatory cytokines and growth factors [349, 361]. NF- $\kappa$ B is a pleiotropic transcription factor involved in the inducible expression of a diverse array of genes. As such, activation of the NF- $\kappa$ B signalling module involves not only the early up-regulation of pro-inflammatory cytokines but also the transcriptional control of apoptosis [362]. Oftentimes, transcriptional regulation requires the participation of several transcriptional factors through protein-protein interactions, known as transcriptional co-activators or co-repressors. For example, NF- $\kappa$ B encompasses an important family of inducible transcriptional activators critical in the regulation of the gene expression in response to injury and inflammatory stimuli. As such, the CREB-binding protein has been identified as co-activator of the NF- $\kappa$ B component p65 and might play an important role in the cytokine-induced expression of various immune and inflammatory genes [363]. Such observations emphasize the role of the CREB regulator in pro-inflammatory signaling pathways including TLR signaling pathway. Further

evidence [364] confers the involvement of over-expressed CREB in inducing apoptosis while the control of FASL induction which mediates programmed cell death in human T lymphocytes [365] appears to be accomplished through a series of regulatory interactions that implicate the role of NF-kB&CREB/ATF pathways [366].

**Table 4.3:** Critical transcription factors in human endotoxemia model

No.	Patterns	Functions	Transcription factors
1	Early-up	Apoptosis	BRNF, CLOX, E2FF, EKLF, ETSF, HEAT, HOXF, IRFF, MAZF, MYT1, NFKB, RXRF, SORY, SP1F
2	Middle-up	Apoptosis	AP4R, CREB, E2FF, ETSF, GATA, HEAT, MAZF, MZF1, NFKB, NKXH, PAX6, SP1F, ZBPF
3	Late-up	Apoptosis	ATBF, BRNF, CLOX, EBOX, ETSF, FKHD, GATA, HOMF, HOXF, IRFF, NKXH, OCT1, PARF, SORY, STAT, TBPf
4	Early-up	Toll-like receptor signaling pathway	EKLF, HEAT, MAZF, MYT1, SP1F
5	Middle-up	Toll-like receptor signaling pathway	CREB, E2FF, EGRF, EKLF, ETSF, EVI1, HEAT, MAZF, MYBL, MZF1, NFKB, NR2F, PAX6, SORY, SP1F, STAT, ZBPF
6	Late-up	Toll-like receptor signaling pathway	AP4R, ATBF, BRNF, CLOX, ETSF, EVI1, FKHD, GATA, HOMF, HOXF, IRFF, MEF2, NKXH, OCT1, PARF, SORY, STAT, TBPf
7	Early-up	Cytokine-cytokine receptor interaction	SORY, TBPf
8	Late-up	Cytokine-cytokine receptor interaction	AP4R, CLOX, EBOX, ETSF, EVI1, FKHD, GATA, HEAT, HOMF, HOXF, IRFF, MAZF, MEF2, NFKB, NR2F, OCT1, PARF, PAX6, RXRF, SORY, SP1F, TBPf
9	Late-up	Jak-STAT signaling pathway	AP4R, BRNF, CLOX, E2FF, EGRF, ETSF, HEAT, HOMF, HOXF, MAZF, MZF1, RXRF, SP1F, ZBPF
10	Down	Citrate cycle (TCA cycle)	ATBF, BRNF, EGRF, ETSF, FKHD, HEAT, HOMF, HOXF, MAZF, MEF2, MYBL, MYT1, MZF1, NR2F, RXRF, SP1F, STAT, TBPf, ZBPF
11	Down	Pyrimidine metabolism	CREB, E2FF, EBOX, ETSF, IRFF, MYBL, SP1F, ZBPF
12	Down	Pyruvate metabolism	HEAT*
13	Down	Ribosome	E2FF, ETSF, RXRF
14	Down	Oxidative phosphorylation	None

\*: present in TF-module '+HEAT\_\_+NRF1\_\_+NRSF'

Additionally, there is considerable evidence indicating the role of the early growth response-1 (member of EGR family) in regulating endotoxin induced SOCS-1 transcription [367]. SOCS-1 has been identified as a critical regulator of both adaptive cytokine signaling and innate immune responses and therefore understanding its transcriptional regulation under inflammatory conditions will no doubt be critical in

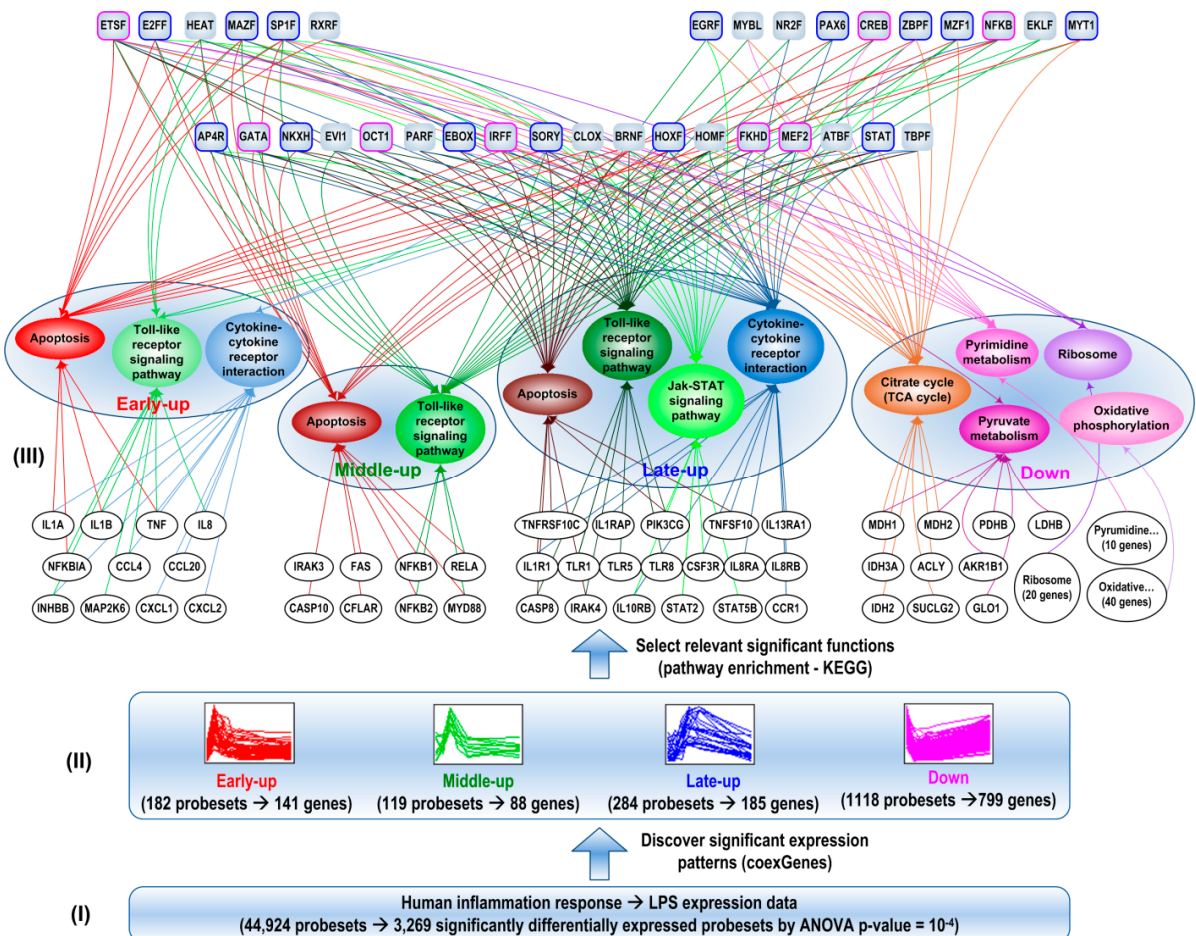
understanding its role in limiting inflammatory responses [368]. Interestingly, these results demonstrate an important role of regulatory members of EGR family in regulating the endotoxin induced activity of the SOCS-1 promoter; thereby validating its presence in our computational predictions. On the other hand, we also observe a significant overlap across various biological processes while comparing these sets of TFs but it is reasonable since transcription factors are characterized by pleiotropic effects [329]. TLR signaling appears to be the principal pathway that initiates the host response to endotoxin and via the cross-talk among other pathways (e.g. Apoptosis, JAK-STAT) amplifies and propagates the inflammatory reaction providing for complex non-linear responses [369]. Here, we also analyzed the reasons why no statistically significant CRM is found in the down-regulated gene batteries of the oxidative phosphorylation pathway (so-called OXPHOS group). OXPHOS itself is composed of genes that are coexpressed across numerous datasets under different conditions [370, 371] and it was proposed as a group of genes that might share a common regulatory mechanism [372]. However, we did not detect any complex-specific arrangement of TFBSs although it is highly enriched by a number of common TFBSs even when the promoter lengths are increased up to 1,000bp upstream. Although this conclusion is similar to the result of a previous study [372], we notice that subunits of each complex in OXPHOS group tend to have tighter coexpression with subunits of the same complex than subunits of other complex which is also proposed and discussed extensively in [372]. Based on the assumption that genes characterized by tightly coordinated expression levels are more likely to share common regulatory elements (proposed and demonstrated in [304]), we assume that genes belonging to the same complex might share some common set of regulatory signals. Therefore, we applied



the same procedure of finding statistically significant CRMs on the control regions of those subgroups of genes including complex I – 17 genes, complex III – 6 genes, complex IV – 4 genes, and complex V – 13 genes. Eventually, we identified statistically significant CRMs for each complex from which relevant transcriptional regulators can be inferred. As a result, from a promoter analysis standpoint we are highly confident that subunits of each complex in OXPHOS group are more likely to be under a common regulatory mechanism rather than all the genes in the entire group. However, from a computational standpoint this result raised another possibility related to whether a subset of genes within a gene battery can provide more statistically significant CRMs than the entire gene battery. Assuming that the possibility is correct, this raises two questions including: (i) what is an appropriate size of the subset as well as (ii) how genes in the subset are selected. In order to address this issue, we make a case-study by randomly selecting a subset of  $N$  genes within the OXPHOS group ( $N=17, 6, 4$ , and  $13$  respectively) and search for significant CRMs. The process is repeated 100 times and the average of minimum significance p-values is calculated. Results show that for  $N=4$ , the average of minimum p-values is comparable to the one with  $N$  genes randomly selected from the background set (**Figure 4.7**). Yet, for the other cases the average of minimum p-values is less significant than the ones from the background set, suggesting that random subsets of genes within a gene battery behave more or less similarly to the case from the background set. Certainly, some subsets can provide more significant CRMs than the entire gene battery but how to interpret those selected subsets and the corresponding results remains a challenge. Therefore, it should be emphasized that using prior biological knowledge might overcome some of these limitations.

### *Putative temporal program of transcriptional regulation*

The administration of a low dose of endotoxin to human subjects elicits dynamic and reproducible changes in the circulating leukocyte population by altering the expression level of numerous genes. Since the host response to endotoxin evolves dynamically, it is possible to observe a dynamic representation in the transcriptional regulatory program (Figure 4.9).



**Figure 4.9:** Putative temporal regulatory program in human endotoxemia plus schematic illustration of the integrated computational framework. The clustering and selection step extracts a ‘clusterable’ subset of differentially expressed probesets and cluster it into a number of expression patterns. Subsequently, pathway enrichment is performed in each

pattern and relevant significant pathways are selected based on literature information. The process of CRM searching is then applied to each gene battery which is a group of genes that belong to an expression pattern and a particular pathway. Eventually, 34 TFs are identified as human inflammation-relevant transcriptional regulators. The results show a highly dynamic perspective of regulation and interactions between genes, functions, and TF across the time.

Due to the fact that transcription factors are characterized by pleiotropic effects [329], it is also reasonable to anticipate a significant overlap among sets of transcriptional regulators across various biological processes. On the other hand, our results also illustrate the phenomenon in which genes involved in the same function (pathway) may exhibit different expression patterns and genes within an expression pattern can participate in different functions, implying that there are different regulatory mechanisms regulating genes in the same function or in the same expression pattern. Along with this dynamic response, the regulatory mechanisms can also be dynamic over the time, leading to the flexibility of the transcriptional network topology. Additionally, the results also reflect the phenomenon that a gene can participate in various functions and thus be regulated by different sets of transcriptional regulators based on the context (e.g. TNF, MYD88).

In order to assess whether coexpressed genes are more likely to be coregulated, we estimate p-values of CRMs in individual gene batteries vs. the corresponding entire pattern of expression (**Table 4.4**). The results show that the estimated p-values values are similar to those calculated for the background set, implying that the entire pattern of coexpressed genes behaves more likely the same as a random background rather than as a

set of genes that share a common regulatory mechanism (see Data S1, sheet ‘CRMs’ and ‘Middle-up TLR’). This supports our assumption related to the definition of a gene battery. Such preliminary results indicate that genes that are both coexpressed and functionally relevant are very likely to be governed by an underlying transcriptional regulatory program.

**Table 4.4:** Statistical significance of selected *cis*-regulatory modules\*

No.	TF-modules	avglen-minlen-maxlen	Common levels	vs. the background <sup>1</sup> (p-value <sup>2</sup> )	vs. the entire pattern <sup>2</sup> (p-value)
1	+AP4R__GATA__HEAT	288__169__485	0.75	1.88E-06	1.78E-05
2	+E2FF__+MOKF__E2FF	333.8__170__514	0.75	1.06E-05	9.32E-08
3	+MOKF__MZFI	168.7__95__236	0.75	3.36E-05	6.37E-07
4	+SP1F__ETSF__NFKB	189__110__268	0.75	3.58E-05	1.78E-05
5	+PAX6__SNAP	154.2__66__260	0.75	4.29E-05	3.82E-05
6	+MOKF__NKXH	101.3__37__194	0.875	4.51E-05	2.57E-05
7	+PAX6__ETSF__ZBPF	271.7__191__326	0.75	4.52E-05	3.82E-05
8	+NKXH__CREB__E2FF	518.3__403__788	0.75	6.85E-05	1.35E-04
9	+MAZF__E2FF	72.1__32__98	0.875	9.82E-05	6.96E-05
10	+NFKB__CREB__SP1F	246.2__117__529	0.75	9.91E-05	1.35E-04

\*: common significant *cis*-regulatory modules that are considered as transcriptional regulators for 8 genes in the middle-up expression pattern that belong to the apoptosis pathway; ‘+’ | ‘-’ TFBSs present on the forward | backward strand orientation; <sup>5</sup>: this CRM contains 3 TFBSs, binding sites of AP4R on the forward and of GATA, HEAT on the backward strand. Its average length is 288 bases while the minimum one has 169 bases and the maximum one has 485 bases. There are  $8 \times 0.75 = 6$  instances of this CRM over 8 control regions of 8 genes; <sup>1</sup>: the background consists of 5,000 randomly selected genes; <sup>2</sup>: the entire corresponding pattern of gene expression (88 genes in this case); <sup>3</sup>: hyper-geometric p-value of this group vs. the background set or vs. the entire pattern.

### ***Comparison with in vitro human endotoxemia model***

In order to assess the stability of our prediction, we applied the analysis to an in vitro human endotoxin model. Data are extracted from a culture of peripheral-blood-derived mononuclear cells stimulated by a high dose of LPS (100ng/ml) [106]. Clustering approach reveals that there exist five critical transcriptional responses. Three of them characterize inflammatory phases similar to those identified in the analysis of *in vivo* data

including an early-up response (284 probesets), a late-up response (700 probesets), and a down regulation (226 probesets). Due to a high dose of LPS administration, it would have expected an up- (367 probesets) and a down-regulation (319 probesets) without returning to the base line after 24hr of LPS administration. Subsequently, a similar analysis of pathway enrichment (using KEGG database) was applied for each set of genes characterizing a transcriptional response. In an overlap with the analysis of *in vivo* data, we select statistically inflammatory relevant significant pathways (p-value < 0.05) that were selected from the analysis on the *in vivo* human endotoxemia model. Accordingly, nine sets of genes that belong to a specific pathway and a pattern of gene expression were extracted, corresponding to nine genes batteries used to determine critical transcriptional regulators relevant to the inflammatory response in this study (**Table 4.5**).

**Table 4.5:** Critical transcription factors identified from the *in vitro* endotoxin study

No.	Patterns	Functions	Transcription factors
1	Early-up	Apoptosis	CLOX, E2FF, EGRF, EKLF, ETSF, FKHD, HOXC, HOXF, IRFF, MAZF, NKXH, NOLF, OCT1, RXRF, SORY, SPIF, STAT, XBBF
2	Late-up	Apoptosis	CREB, EKLF, MAZF, NFKB, SORY, ZBPF
3	Early-up	Toll-like receptor signaling pathway	APIR, CLOX, E2FF, EGRF, EKLF, ETSF, HOXC, IRFF, NFKB, NOLF, NR2F, OCT1, RXRF, SORY, SPIF, STAT, XBBF, ZBPF
4	Late-up	Toll-like receptor signaling pathway	ABDB, CLOX, ETSF, HOMF, HOXF, IRFF, NFKB, NKXH, RXRF, SORY, STAT, TBPf
5	Early-up	Cytokine-cytokine receptor interaction	CREB, ETSF, FKHD, HOXF, RXRF, STAT, TBPf
6	Late-up	Cytokine-cytokine receptor interaction	ABDB, HOXF, NR2F, OCT1, RXRF, SORY, STAT
7	Early-up	Jak-STAT signaling pathway	ABDB, APIR, AP4R, E2FF, EGRF, EKLF, ETSF, FKHD, HOMF, HOXF, IRFF, MAZF, NKXH, RXRF, SORY, SPIF, STAT, TBPf, XBBF, ZBPF
8	Late-up	Jak-STAT signaling pathway	ABDB, APIR, AP4R, CLOX, CREB, E2FF, ETSF, FKHD, HOMF, HOXC, HOXF, NKXH, NR2F, OCT1, RXRF, SORY, TBPf
9	Up-remained	Pyrimidine metabolism	AP4R, E2FF, EGRF, EKLF, ETSF, FKHD, HOXF, MAZF, NFKB, NKXH, NOLF, NR2F, RXRF, SPIF, XBBF, ZBPF

Subsequently, the proposed method has been applied to search for statistical significant CRMs which are decomposed into a list of TFBSs to infer associated TFs that may be

functional transcription factors in the regulation of inflammatory transcriptional responses. In a similar manner with the *in vivo* analysis, TFs that are present with the high frequency among gene batteries (at least three times) are reported (**Table 4.5**). We identify 27 critical TFs of which more than 80% are present in the list of relevant transcriptional regulators found in the analysis of the *in vivo* data including AP4R, CLOX, CREB, E2FF, EGRF, EKLF, ETSF, FKHD, HOMF, HOXF, IRFF, MAZF, NFKB, NKXH, NR2F, OCT1, RXRF, SORY, SP1F, STAT, TBPf, ZBPF. Given that different dosing amounts of LPS have been applied in two experiments, there may be different genes involved in the response of the same function between the *in vivo*- and *in vitro*- model, resulting in different TFs involved in the transcriptional regulation of the same gene battery between two cases. However, the significant overlap between two final lists of predicted TFs relevant to inflammatory transcriptional responses provides promising implications of the predictive performance of the method. Therefore, the proposed framework appears to be a robust and valuable methodology to identify critical transcriptional regulators relevant to biological responses under external stimuli.

#### **4.8. Limitations and advantages**

One of the key features in our analysis is the identification of significantly overrepresented CRMs in each gene battery. Based on the size of a gene battery, a corresponding significance threshold is applied to select statistically significant CRMs. Since these recognized CRMs are located on the control regions of many ‘hypothetically’ co-regulated genes, they are likely to be composed of functional binding sites that are activated upon the initiation of the transcriptional machinery. We therefore decompose these CRMs into a list of TFBSs to infer associated TFs which can be considered as

relevant transcriptional regulators of the corresponding gene battery. In particular, TFs that are present with the high frequency among gene batteries (at least three times across fourteen gene batteries) are assumed to play a key role in the biological process. We identify 34 TFs relevant to the human inflammatory responses, of which around 25% has been experimentally shown to be involved in the inflammatory and/or immune response based on literature evidence and more than half of the remaining have been computationally shown to play a critical role in the regulation of immune system [334].

Our analysis has attempted to reverse engineer the underlying regulatory network of the human blood leukocyte response to a prototypical inflammatory stimulus (LPS). Given the transcriptional profiling data of human blood leukocytes, an elementary set of temporal responses with putative transcriptional regulators have been identified. A key feature of the analysis is the exploration of the concept ‘gene battery’ which represents for a group of genes that are both co-expressed and functional relevant to identify inflammatory transcriptional regulators using a context-specific searching approach [373]. Novel heuristics regarding to computational issues e.g. eukaryotic genes consist of multiple alternative promoters leading to a huge complexity are also proposed. In order to provide a systematically unbiased *in silico* approach, CRM structural constraints are also adjusted so that no parameter is required except for the statistical significance thresholds. Furthermore, our analysis also allows for the reconstruction of a dynamic temporal regulatory network, making it a critical enabler for improving our understanding of how the transcriptional machinery ‘program’ effectively regulates key cellular processes.

Although no single analysis can identify all transcriptional regulators involved in a response, it has been demonstrated that the proposed framework can identify critical TFs

that are relevant to acute inflammatory responses. Despite the fact that many methods have been proposed in the literature to search for relevant transcriptional regulators, different approaches explore different biological assumptions resulting to different sets of putative TFs which may or may not significantly overlap each other. Since the true extent of all TFs involved in the regulation of a complex response under some external stimuli is unknown, these differences could not be interpreted as the high- or low- accuracy of the methods. Instead, all of found TFs may be involved in different processes of the response but because of the limitation of hypotheses used by the methods, they may not be recognized by a certain approach.

Novel methods are still proposed using different analytical approaches but generally they can be categorized into two main directions including mRNA expression-based [374-376] and TF binding pattern-based methods [320, 377-380]. The first direction somehow utilizes the fundamental hypothesis that the mRNA expression level of TFs is proportional to their protein concentration but this may not be appropriate especially in higher eukaryotes since TF activation is often regulated post-translationally and acts somewhat in an independent manner of expression level. Some methods also require multiple-condition data as the input which may not be applicable when practical data are only sampled under one condition/treatment [374-376]. In the meanwhile, a lot of methods following to the latter direction have been developed e.g. FrameWorker [320], CMA [379], CRÈME [377], ModuleMiner [378], CisModule [380], BioMoby [381] etc. of which ours is among them. These are not limited by the mRNA expression proportion hypothesis but they are limited by promoter identification, TF binding profiles, and the underlying assumption to select the input set of ‘co-regulated’ genes.



Therefore, we opt to extend an available computational tool, FrameWorker, to take into account the fact that genes of higher eukaryotes contain multiple alternative promoters exploring the rich information of the Genomatix database on promoters and TF binding profiles. The underlying assumption that coexpressed genes are more likely to share some common regulatory mechanism when they are functional-relevant has been explored to predict putative functional activation of TFs in a specific context. These factors make our method become incomparable or unnecessary to compare with available methods. However, given the future availability of more complete TF binding data and other resources, the method could be enhanced by integrating protein-protein interaction to refine selected CRMs or using other tools to support the selection of relevant functions e.g. Pathway-Express [382]. Since each single method or even each direction always contains its own limitations and advantages, one possibility in future improvements could be the development of a framework to obtain a consensus result under diverse underlying hypotheses from various outputs of different methods.

## **Chapter 5 – Cellular variability and circadian control in human endotoxemia**

### **5.1. Introduction**

Systemic inflammation is evoked by many stimuli including infection, trauma, invasive surgery and biological stressors in general; furthermore, it is typically observed in many critical illnesses [383]. While the host inflammatory response is essential to resolve the infection or repair the damage to restore homeostasis, it also plays a central pathogenic role in a wide spectrum of diseases [62]. Under normal circumstances, the inflammatory response is activated, initializes a repair process and then abates [23]. However when anti-inflammatory processes fail, an amplified pro-inflammatory signal can turn what is normally a beneficial reparative process into a detrimental physiological state of severe, uncontrolled systemic inflammation [24]. In order to gain a better understanding of the molecular mechanisms and physiological significance associated with inflammatory responses, alternative clinically relevant models have been proposed including the human endotoxemia model in which an intravenous administration of *E.coli* endotoxin (lipopolysaccharide – LPS) is given to healthy human subjects [67, 384]. Bacterial endotoxin, a component of the outer cell membrane of gram-negative bacteria, is an important mediator in the pathophysiology of gram-negative bacterial sepsis [385]. This complex macromolecule induces its injurious effects by a non-cytotoxic interaction with CD14-bearing inflammatory cells, such as macrophage-monocytes, circulating neutrophils and lung epithelial cells. These effector cells are activated through a family of

Toll-like receptors (TLR) and subsequently release a network of inflammatory products. While we do not argue that the human endotoxin challenge model precisely replicates an acute infectious or sepsis condition, we believe that human endotoxin challenge does serve as a useful model of TLR4 agonist-induced systemic inflammation while at the same time providing a reproducible experimental platform.

The inflammatory response is a complex non-linear process involving a cascade of events mediated by a large array of immune cells and inflammatory cytokines [386]. At the cellular level, innate immune cells are activated leading to the production and release of pro-inflammatory and anti-inflammatory cytokines into the systemic circulation for communication between cells [61, 387]. Anti-inflammatory cytokines counteract the effects of pro-inflammatory cytokines and the relative concentration or balance between them strongly affects to the degree and extent of the response [388, 389]. At a higher level, the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system (SNS) produces stress hormones [390] whose pattern of release follow broad circadian rhythms which play critical roles in immune responses [391-394]. The rhythmicity is regulated by the 24 hour light/dark cycle exerting diurnal effects on numerous inflammatory cytokines [395, 396]. The complexity of the overall response has encouraged the development of mathematical and computational models as a means of exploring the connections between multiple components.

Various modeling approaches have been proposed in the literature, but generally they can be classified into two main categories: equation- and agent-based modeling [73, 397, 398], including our prior approaches using deterministic ordinary differential equations (ODE) for developing models of human endotoxemia [83, 84, 399]. However,

deterministic ODE models assume homogeneity and perfect mixing within compartments, while ignoring spatial effects [73]. Given that stochasticity and heterogeneity can have profound effects on the function of biological systems [400-402], an alternative, more natural, approach – agent-based modeling (ABM) is explored. ABM is an object-oriented, rule-based, discrete, and stochastic modeling method [77, 403]. Interactions between agents (cells, molecules) in a model are nonlinear, stochastic, spatial, and are described by asynchronous movements through multiple compartments. Furthermore, every agent in the model is encapsulated with its own properties and behaviors making the system able to exhibit emergent behaviors arising from simple interactions between agents. The usefulness and applicability of ABMs vary but some have been applied to immunological problems and findings derived from these models generated a lot of insights into the interactions and dynamics at the cellular level in immune responses. For example, Jenkins and colleagues [404] investigated B-T cell interactions in the absence of directed cell chemotaxis during the first 50hr of a primary immune response to an antigen; Gary An and coworkers have pioneered many ABMs representations to evaluate the dynamics of the innate immune response and the efficacy of proposed interventions for SIRS/multiple organ failure (MOF) [70, 92] or examine the dynamics of the TLR4 signal transduction cascade with LPS preconditioning and dose-dependent pro-inflammatory response effects [405, 406]. They also developed a basic immune simulator (BIS) to qualitatively examine the interactions between innate and adaptive interactions of the immune responses to a viral infection [407]. Furthermore, there is a variety of successful agent-based simulators that have been constructed as

frameworks for immunology/disease understanding and exploration e.g. IMMSIM [408, 409], SIMMUNE [410], CyCells [411].

In order to investigate the cellular variability through the interactions and dynamics of inflammatory cytokines in acute inflammatory responses following endotoxin administration, we first construct a homeostatic model of human endotoxemia using the agent-based approach which naturally incorporates key biological features (e.g. stochasticity, heterogeneity, and discreteness) and physicochemical properties of biological molecules. While in our prior work [83, 84, 399], we focused on the possibility of modeling the transcriptional dynamics of cellular responses, we here attempt to capture stochastic variation in the transcriptional process, one of the key factors leading to phenotypic variation besides the genetic and environmental variability [412-415]. The main aim of this study is establishing a multi-scale modeling framework capable of simulating main characteristics of critical components in human endotoxemia to examine (i) the balance and distribution of inflammatory cytokines in a population of heterogeneous leukocytes and (ii) the interplay between circadian controls and endotoxin treatments through a novel quantity based on the cell-to-cell variability.

## **5.2. The in silico model of human endotoxemia**

### **5.2.1. The system dynamics model**

Based on our prior studies [83, 84, 399] high-dimensional transcriptional profiling data from human blood leukocytes following LPS administration are decomposed into four significant expression patterns, capturing the essence of three inflammatory phases including a pro-inflammatory response (‘early-up’ & ‘middle-up’ expression pattern, P), a counter-regulatory/anti-inflammatory response (‘late-up’ expression pattern, A), and a

dysregulation in leukocyte bioenergetics ('down' pattern, E) [104]. These define the basic elements (state variables) characterizing the leukocyte response to endotoxemia.

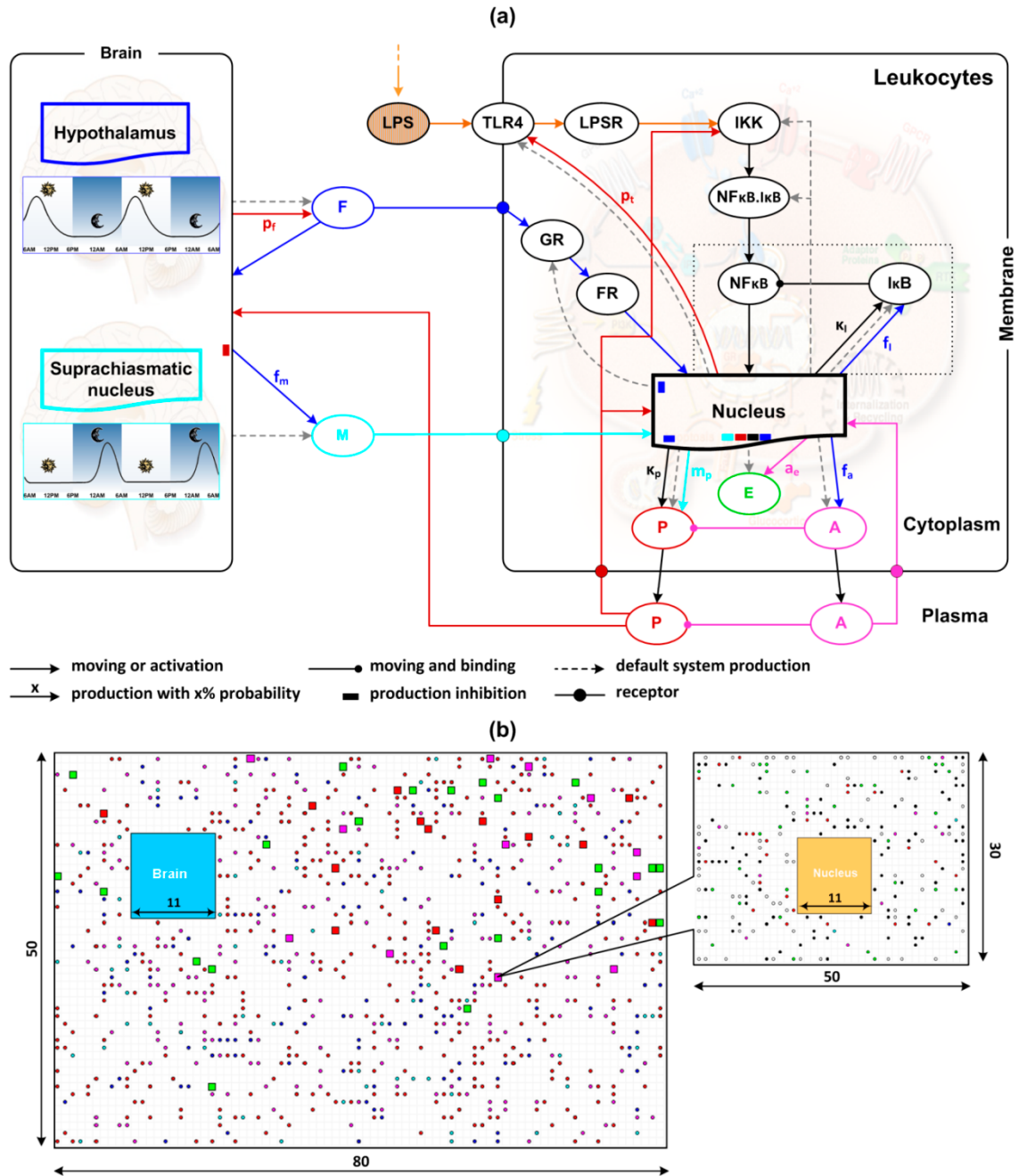
A number of assumptions have been made to construct the model, namely: (1) Peripheral blood leukocytes can be approximated as a community of leukocytes whose main behavior is characterized by asynchronous and stochastic activities without an intra-cellular spatial localization; (2) The dynamics of the pro-inflammatory response, the counter-regulatory response, and the dysregulation in leukocyte bioenergetics can be characterized by patterns of corresponding pro-inflammatory cytokines, anti-inflammatory cytokines, and bio-energetic proteins; (3) Different types of pro-inflammatory cytokines, anti-inflammatory cytokines, and bio-energetic proteins can be approximated with corresponding average delegators as P, A, and E respectively whose main behavior is associated with asynchronous and stochastic activities. Lastly, it has been observed that after LPS challenge many pro-inflammatory cytokines exhibit similar dynamics as is observed in their corresponding mRNA temporal profiles e.g.  $\text{TNF}\alpha$ , IL6, IL8, etc [416]. IL10 – a member of anti-inflammatory cytokines shows a slight difference between its mRNA and its protein temporal profile. While mRNA levels of IL10 dropped during the first hour and its protein levels rose very modestly, both profiles still exhibit up-regulation. Consequently in this context, we hypothesize that the common dynamics of pro- and anti-inflammatory cytokines can be characterized by their average mRNA expression profiles.

Such expression dynamics of inflammatory cytokines is assumed to be mainly regulated by the activation of relevant transcription factors (TFs). The nuclear factor-kappa B (NF $\kappa$ B) was selected as the representative signaling controller underpinning the

manifestation of transcriptional responses due to its essential role in the immune system [417, 418] and extensive prior computational analyses [419]. Furthermore, NF $\kappa$ B activities are primarily modulated by the kinase (IKK) activity and the inhibitor I $\kappa$ B through the Toll-like receptor (TLR) signaling pathway – a pivotal pathway subjected to crosstalk from other signals and pathways (e.g. JAK-STAT [369, 420]) [421, 422]. Such regulation can be characterized by the ubiquitous paradigm of a two-feedback mechanism: a positive- and a negative-feedback [422-425]. Therefore, we hypothesize that the dynamics of inflammatory cytokines are mainly regulated by intra-cellular signaling cascades and transcription factors whose activities can be characterized with a paradigm of two-feedback regulatory mechanism.

At the systemic level, pro-inflammatory cytokines released from the innate immune system induce signals activating the hypothalamic-pituitary adrenal (HPA) axis, thus controlling the secretion of glucocorticoids (cortisol in primates or corticosterone in rodents) [1, 426]. Of particular interest is the hormone melatonin given its potential role as a mediator in the crosstalk between the suprachiasmatic nucleus (SCN) and the immune system [427, 428]. The corresponding hormone levels exhibit a circadian pattern with strong effects on the production of inflammatory cytokines [395, 396]. While cortisol reaches its peak in the early morning [429], melatonin's peak production occurs late at night and remains at a low level for the rest of the day [428, 430]. Therefore, in the model developed here cortisol (F) is set under the control of hypothalamus (HPT) while melatonin (M) is regulated by the SCN. The system dynamics of the proposed human endotoxemia model including all components and associated interactions and a snapshot of the model representation are succinctly presented in **Figure 5.1**. Simulated molecular

types and their corresponding characteristics are shown in **Table 5.1**. Details of model components, rules, and parameters are discussed as follows.



**Figure 5.1:** *In silico* human endotoxemia model accounting for circadian variability. (a) The system dynamic model. At the cellular level, molecular interactions involve the propagation of LPS signaling on the transcriptional response level (P, A, E) through the



activation of NF- $\kappa$ B signaling module. At the systemic level, circulating stress hormones are released from the neuro-endocrine system coupled with their circadian rhythms. The dynamics of cortisol (F) and melatonin (M) signaling from the systemic level involve molecular behaviors at the cellular level. The activities of each agent are characterized by its corresponding color. (b) A snapshot of the implemented model. Molecules are displayed with solid circles (P: red-; A: magenta-; F: blue-; M: cyan-; NF $\kappa$ B: yellow-; E: green-; TLR & GR: white-; I $\kappa$ B, IKK, NF $\kappa$ B.I $\kappa$ B: black- circles). Cells are displayed with solid squares where green squares represent for cells with an approximate number of P and A, red squares for those with the number of P greater than 1.5 fold of the number of A and magenta squares for those with A more than 1.5 fold of P.

**Table 5.1:** Model components

No.	Components	Description	Approximate half-life (hr)	Initial population size*
1	LPS	Lipopolysaccharide (endotoxin)	1.0	n/a
2	TLR4	Toll-like receptor 4	2.0	40
3	LPSR	LPS-TLR4 complex – active form	2.0	n/a
4	IKK	I kappa-B kinase complex – activated by LPSR	2.5	50
5	NF $\kappa$ B.I $\kappa$ B	NF $\kappa$ B complex – inactive form	2.5	50
6	NF $\kappa$ B	NF $\kappa$ B – active form	2.0	n/a
7	I $\kappa$ B	I kappa-B – NF $\kappa$ B inhibitors	0.5	10
8	P	Pro-inflammatory proteins – active when imported	1.5	30
9	A	Anti-inflammatory proteins– active when imported	1.5	30
10	E	Bio-energetic proteins	2.0	40
11	F	Cortisol– active when imported <sup>§</sup>	1.0	n/a
12	GR	Glucocorticoid receptors	2.0	40
13	FR	Cortisol-receptor complex – active form	2.0	n/a
14	M	Melatonin– active when imported	1.0	n/a

\*: the initial corresponding number of molecules within a cell; <sup>§</sup>: the status of P, A, F, and M change to active when they are imported to the cytoplasm (cells) or brain compartment

### 5.2.2. Agent rules

Agents are simulated objects (cells, molecules) that follow specific instructions on how they behave and interact with other agents within or between compartments. The rule system is listed in **Table 5.2**. Briefly, when LPS is recognized by its receptors TLR-4 a signal transduction cascade triggers downstream intracellular signalling modules to ultimately activate the transcription of inflammatory genes. Such transcriptional processes are assumed to be mainly regulated by transcription factors for which NF $\kappa$ B serves as a proxy whose activities, including activities of IKK and I $\kappa$ B in the NF $\kappa$ B-signaling module, have critical role in the inflammatory response [431, 432]. Following the activation of NF $\kappa$ B through the phosphorylation of the inhibitor protein I $\kappa$ B by IKK, NF $\kappa$ B is translocated to the nucleus to activate the transcriptional processes resulting in the production of pro-inflammatory cytokines (e.g. TNF $\alpha$ ) and I $\kappa$ B [349, 433, 434]. After released to the systemic circulation, these pro-inflammatory cytokines may bind to their corresponding receptors on the membrane of leukocytes and either further activate the NF $\kappa$ B-signaling module [433, 435] or lead to production of additional TLR-4 molecules [436, 437]. On the other hand, they also act as hormone-like signals that converge to activate the HPA axis to produce glucocorticoids [1, 426] or suppress the nocturnal melatonin production [438-440]. While glucocorticoids have critical roles in the anti-inflammatory arm of the host defense system by inducing the expression of anti-inflammatory proteins such as I $\kappa$ B and anti-inflammatory cytokines (e.g. IL10) [1, 441], they also act as potential modulators that enhance the production of melatonin [439, 442, 443]. Melatonin in turn which can be also a regulator modulates the production of pro-inflammatory cytokines [395, 428]. To establish the link between the inflammatory

response and the cellular energetic state, we assumed that there are a number of bio-energetic molecules in each cell which control the production of new molecules in the cell and represent the overall cellular energetic status. If the number of bio-energetic molecules in a cell is positive, the cell will be able to produce new molecules and in the meantime the default production of E will be inhibited. Since anti-inflammatory cytokines are responsible for the counter-regulation of the pro-inflammatory responses, it is hypothesized that they have the role in increasing the amount of energy.

**Table 5.2:** Model rules

No.	Rule definition
1	LPSR and P imported to cells from plasma can activate IKK; activated IKK can activate NFκB.IκB to NFκB
2	An individual NFκB in the nucleus has a probability of $\kappa p/\kappa i$ to produce a new unit of P/IκB respectively
3	IκB inhibits NFκB activity by forming NFκB.IκB complex
4	P, A in the inactive form can be released to plasma if they lie on the membrane (boundary) of cells
5	P, A, F, M can be imported to cells from plasma if they hit a cell when moving in plasma
6	P, A, F, M after imported to cells from plasma will not be released to plasma again
7	An individual P in the nucleus has a probability of $p_t$ to produce a new unit of TLR4
8	An individual A in the nucleus has a probability of $a_e$ to produce a new unit of E
9	A inhibits P activity; both are degraded when they hit each other
10	An individual FR in the nucleus has a probability of $f_a/f_i$ to produce a new unit of A/IκB respectively
11	NFκB activity in the nucleus is inhibited if the number of NFκB is less than the number of FR in the nucleus
12	FR inhibits the default system production of GR when in the nucleus
13	An individual F in the brain has a probability of $f_m$ to produce a new unit of M
14	An individual M in the nucleus has a probability of $m_p$ to produce a new unit of P
15	An individual P in the brain has a probability of $p_f$ to produce a new unit of F
16	P in the brain prevents F from producing M if the number of P is two folds more than that of F in the brain
17	NFκB, active P, FR, active M, and IκB can be translocated to the nucleus; they inhibit the default system production of E if they stimulate the nucleus activity to produce a new unit
18	0.5% of individuals F in the homeostatic system are added with the probability of $\sin(0 \rightarrow \pi/2)$ for the time from 3:00AM to 9:00AM
19	2% of individuals M in the homeostatic system are added with the probability of $\sin(0 \rightarrow \pi/2)$ for the time from 10:00PM to 2:00AM
20	Molecules are degraded after $\sim t$ hr if there is no action except movements where $t/2$ is defined by the approximate half-life of molecular types

### 5.2.3. Agent movements and interactions

There are four types of compartments: the plasma, the brain, the cytoplasm, and the nucleus. The plasma contains the brain and all simulated cells (50 in this study); each cell contains a cytoplasm and a nucleus. All agents move in a random fashion following the 'random walk' model on a 2-dimensional grid (see Materials and methods). The Plasma and each cell have their own simulating grid while brain and nucleus directly occupy a region in the plasma and corresponding cell simulating grid respectively. There is no special spatial arrangement for agents. However, there are a number of restrictions on which compartment a molecule can be in. Specifically, LPS can only move in the plasma compartment; LPSR, IKK, NF $\kappa$ B.I $\kappa$ B, E, and GR are only present in cytoplasm; NF $\kappa$ B, I $\kappa$ B, and FR can be in both cytoplasm and nucleus; M and A cannot be in brain and F cannot be present in nucleus while P can move between any compartment. TLR4 molecules after produced are transferred to the cell membrane i.e. when they reach the boundary of the corresponding cell simulating grid they are fixed there until they are degraded.

Molecules are translocated between compartments based on an import- and export procedure. In plasma, if a molecule has the same position with a cell or reach the region of the brain, the system will check to determine whether it is imported or not. Except LPS, other molecule types are imported to the brain and cells with the approximate probability of LPS-binding TLR4 to simulate the probability of the binding to receptors. This is approximately to the initial number of TLR4 molecules in a cell divided by the number of positions on the boundary of the cell simulating grid which is about 30%. For LPS molecules, a random position on the boundary of the cell simulating grid is assigned;

if it is overlapped with the position of some TLR4 molecule, it will be imported. If imported to cells, the molecule status is changed to 'active'. In the cytoplasm compartment, active molecules are simply translocated to the nucleus compartment when they reach the nucleus regions in the corresponding simulating grid. On the other hand, when a molecule reaches the boundary of a compartment, it is exported to the outer compartment if it is not restricted.

Each agent moves in a random direction for a random number of times with a random delay time for each movement. However, two interactive molecules  $X_1, X_2$  with current positions  $\{P_x^{X_1}, P_y^{X_1}\}, \{P_x^{X_2}, P_y^{X_2}\}$  respectively will move towards the position where an interaction may occur if their distance is less than a threshold  $d(X_1, X_2) = \max\left\{\left|P_x^{X_1} - P_x^{X_2}\right|, \left|P_y^{X_1} - P_y^{X_2}\right|\right\} \leq \tau, \tau = 1$ . If two molecules have the same position on the simulating grid of the corresponding compartment, they will interact (activation, inhibition, or degradation) following the rules showed in **Figure 5.1a** and **Table 5.2** e.g. A and P with the same status in any compartment, LPSR and IKK, activated IKK and NF $\kappa$ B.I $\kappa$ B, F and GR in cytoplasm, and NF $\kappa$ B and I $\kappa$ B in cytoplasm or nucleus. The rule is also applied to the movement of molecules and cells in plasma to increase the probability of entering a cell for molecules in systemic circulation.

Finally, circadian controls are introduced in an attempt to simulate the daily patterns of stress hormones [428-430]. In our simulation, these rhythms are produced using sine waves. At every tick during the time from the onset of the production to the corresponding peak in a day (e.g. 3:00AM to 9:00AM for cortisol and 10:00PM to 2:00AM for melatonin), a constant number of F and M units ( $c_{fm}$ ) are added to the system where the probability for each adding such a unit is

$$t_F = T_{cur} \bmod N_{tph}; \quad t_M = (t_F + 2 \times N_{tph}) \bmod N_{tpd}; \quad N_{tpd} = 24 \times N_{tph}$$

$$prob_F = \sin \left[ (t_F - 3 \times N_{tph}) \times \frac{2\pi}{N_{tpd}} \right]; \quad prob_M = \sin \left[ t_M \times \frac{3\pi}{N_{tpd}} \right]$$

Simulated time is scaled from ‘ticks’, which is the simulation counter, to hours.  $T_{cur}$  is the current tick of the simulation counter which expresses the current simulated time.  $N_{tph}$  is the number of ticks corresponding to one simulated hour.  $c_{fm}$  is selected to have the peaks of F and M approximately triple their corresponding homeostatic levels ( $c_{fm} = 3$  in this study). These activities are assumed to be controlled and taken place in the brain compartment since they are all associated with behaviors in brain. Definition of the time-scale and the homeostatic system will be discussed in the following section.

### ***‘Random walk’ model***

Agents (cells, molecules) move on a 2-dimensional grid in a random fashion depending on two main factors: the time agents wait before each movement and the number of times agents move in a direction. For a specific agent U, at time t, let  $\gamma(t)$  be the time (number of ticks) U has to wait before moving and  $\lambda(t)$  be the number of times U will move in direction D, we have

$$\gamma(t+1) = \begin{cases} \gamma(t) - 1 & \text{if } \gamma(t) > 0 \\ \text{rand}\{0,1\} + \text{status}(U) & \text{if } \gamma(t) = 0 \end{cases}; \quad \lambda(t+1) = \begin{cases} \lambda(t) & \text{if } \gamma(t) > 0 \text{ and } \lambda(t) > 0 \\ \lambda(t) - 1 & \text{if } \gamma(t) = 0 \text{ and } \lambda(t) > 0 \\ \text{rand}\{2,3,4\} \times (N_{comp} + 1) & \text{if } \lambda(t) = 0 \end{cases}$$

$$\text{where } \text{status}(U) = \begin{cases} \text{rand}\{1,2\} & \text{(initial value)} \\ 0 & \text{if } U \text{ is in the active form} \end{cases}$$

$$N_{comp} = \{0, 0, 1, 2\} \text{ if } U \text{ in } \{\text{nucleus, brain, cell, plasma}\} \text{ respectively}$$

Each compartment or each cell has its own 2-dimensional simulating grid. When  $\gamma(t)$  is zero, U will move to the next grid-space in the Moore neighbourhood of the corresponding simulating grid which consists of 8 spaces immediately adjacent to and surrounding the current position based on the current direction D. D is one of 8 directions  $\{N, NE, E, ES, S, SW, W, WN\}$  (N: north, E: east, S: south, and W: west). Let  $P_x(t), P_y(t)$  be the current position of U in a 2-dimensional simulating grid, its next position is defined as follows

$$\begin{aligned}
 P_x(t+1) &= P_x(t) + H_x(\gamma(t), D(t)); \quad P_y(t+1) = P_y(t) + H_y(\gamma(t), D(t)) \\
 \text{where } D(t+1) &= \begin{cases} D(t) & \text{if } \lambda(t) > 0 \\ \text{rand}\{N, NE, E, ES, S, SW, W, WN\} & \text{if } \lambda(t) = 0 \end{cases} \\
 H_x(\gamma(t), D(t)) &= \begin{cases} 0 & \text{if } \gamma(t) > 0 \text{ or } D(t) \in \{N, S\} \\ +1 & \text{if } \gamma(t) = 0 \text{ and } D(t) \in \{NE, E, ES\} \\ -1 & \text{if } \gamma(t) = 0 \text{ and } D(t) \in \{SW, W, WN\} \end{cases} \\
 H_y(\gamma(t), D(t)) &= \begin{cases} 0 & \text{if } \gamma(t) > 0 \text{ or } D(t) \in \{E, W\} \\ +1 & \text{if } \gamma(t) = 0 \text{ and } D(t) \in \{WN, N, NE\} \\ -1 & \text{if } \gamma(t) = 0 \text{ and } D(t) \in \{ES, S, SW\} \end{cases}
 \end{aligned}$$

#### 5.2.4. Model parameters

Model parameters are classified into two categories: default- and production- parameters. Default parameters are those related to system settings and physicochemical properties of cells and molecules, such as compartment extensions, simulation scales, molecular lifetimes, or initial populations. For simplicity, in this study all compartments are simulated with unitless rectangular grids. The plasma is represented by an 80 x 50 rectangular unitless grid and the cell with 40 x 30. The cell nucleus is about 10% of the total cell volume and thus it occupies a region of about 11 x 11 on the cell simulating

grid. Similarly, the brain compartment is also simulated by a region of 11 x 11 on the plasma simulating grid.

Since it is still unclear about the relationship between the system response time and the system production rate, we would like to define two scales in this simulation including (1) the life-scale ( $L$ ) that characterizes for the lifetime of molecules and the system production rate and (2) the time-scale ( $N_{ph}$ ) that characterizes for circadian controls and system responses. The time-scale is initially equal to the life-scale but adjusted later to match *in silico* system responses with *in vivo* transcriptional responses. In order to identify the life-scale, the system is set to have no activity except the default system production and the protein degradation and thus the number of units of each molecule type in a cell should be unchanged. Given the default production rate is  $R\%$ , after an hour a cell will produce  $R \times L$  new units for a molecule type and thus there must be  $R \times L$  units of this molecule type degraded to keep the cell at homeostasis ( $R = 50\%$  in this study). Consequently, if a molecule has a certain lifetime, its average lifetime will be approximately to its number of units divided by  $R \times L$ . In other words, the initial number of units of a molecule type should be set equally to its average lifetime multiplied by  $R \times L$ .

In this simulation, the average lifetime of a unit is double its approximate half-life which is listed in **Table 5.1**. Specifically, I $\kappa$ B half-life is about 0.5 hour and the NF $\kappa$ B.I $\kappa$ B complex have five-fold more than that of I $\kappa$ B [432, 444]; inflammatory cytokines and stress hormones have the average half-life about 1 hour [445, 446]; the largest protein IKK is assumed to have its half-life equal to that of the NF $\kappa$ B.I $\kappa$ B complex; and the rest are assumed to have the average half-life about 2 hours. Let  $f$  be the initial number of



units for I $\kappa$ B in a cell, the initial population of NF $\kappa$ B, I $\kappa$ B, IKK, P, A, E, TLR4, and GR in a cell will be (5f, 5f, 3f, 3f, 4f, 4f, 4f) respectively. Since the cell protein occupies 15–35% of cell volume [447], we here assume that there are approximately 300 units in a normal cell which is 25% of the cell volume. Therefore, the total initial number of units in a cell under the assumption of the homeostatic system will be 29f, resulting in  $f = \frac{300}{29} \approx 10$  units. The estimated initial population size of each molecule type in a cell is given in **Table 5.1**. The life-scale  $L$  therefore is 2f which is 20 ticks per hour which is equal to the lifetime of I $\kappa$ B. Also, the initial number of units for P/A in plasma is initialized by 10% of all P units in all cells in the system. Further, the default production of F and M is set to the activities of the brain (see Figure S1 for the programming architecture and initial parameter values).

Production parameters are the probabilities of producing new units of a molecule when present in the nucleus or brain compartment as indicated in the system dynamics model. It is hypothesized that there is a balance between protein synthesis and degradation in the homeostatic system [448]. Thus, without any external stimulation and circadian influences production parameters need to be adjusted so that the number of units of each molecule type in the system does not change significantly over the time (**Table 5.3**). Techniques from process trending analysis are utilized to obtain the set of adjusted parameters whose values remain unchanged for subsequently added mechanisms e.g. circadian rhythms, endotoxin treatments [449, 450] (see Materials and methods). The current configuration of the homeostatic system including all agents and their properties is saved for further experiments.

**Table 5.3:** Model production parameters

No.	Parameters	Initial probability (%)	Adjusted probability (%)
1	$\kappa p$ (NF $\kappa$ B $\rightarrow$ P)*	70.00	69.44
2	$\kappa i$ (NF $\kappa$ B $\rightarrow$ I $\kappa$ B)	70.00	70.08
3	$f i$ (F $\rightarrow$ I $\kappa$ B)	70.00	70.08
4	$f a$ (F $\rightarrow$ A)	70.00	74.77
5	$f m$ (F $\rightarrow$ M)	70.00	27.48
6	$m p$ (M $\rightarrow$ P)	70.00	69.44
7	$p f$ (P $\rightarrow$ F)	70.00	24.98
8	$p t$ (P $\rightarrow$ TLR4)	70.00	70.00
9	$a e$ (A $\rightarrow$ E)	70.00	75.72

\*x (Y  $\rightarrow$  Z): x is the probability that a single unit Y can produce an individual unit Z when Y is in the nucleus (or brain) compartment.

### ***Parameter tuning***

Based on the trend of the dynamics of each particular molecule type  $X$ , we adjust the probability of the associated production parameter  $p_X$  (**Table 5.3**) so that the total number of  $X$  in the system does not change significantly over the time. For each simulated day ( $24N_{tph}$  ticks), we sample the level of  $X$  at each hour and determine whether there is a significant change based on the sample vector using the ordinary least square regression and significant mean difference [449].

Let  $x_j$  be the number of molecules  $X$  in the system at hour  $j$ ,  $j = 1, \dots, J, J = 24$ . The regression model used in this approach is  $x_j = \alpha + \beta j + \varepsilon_j$  where  $\alpha$  is the intercept,  $\beta$  is the slope, and  $\varepsilon_j$  are random errors which are assumed to be independent and identically distributed. The estimates of the slope and intercept are given by

$$\hat{\beta} = \frac{\sum_j (j - \bar{j})(x_j - \bar{x})}{\sum_j (j - \bar{j})^2}; \quad \hat{\alpha} = \bar{x} - \hat{\beta} \bar{j}; \quad \bar{x} = \frac{1}{J} \sum_j x_j$$

The standard error of the slope will be  $SE(\hat{\beta}) = \sqrt{\frac{\sum_j (x_j - \hat{\alpha} - \hat{\beta} j)^2}{(J-2) \sum_j (j - \bar{j})^2}}$

A 95% confidence interval for the slope  $\beta$  is  $\hat{\beta} \pm t_{0.975, J-2} SE(\hat{\beta})$ . If zero is not contained in the interval, we conclude that the trend of change is significant.

Let  $m_1, m_2$  be the means of the first and last half of the sample vector

$$m_1 = \sum_{j=1}^{J/2} x_j; \quad m_2 = \sum_{j=J/2+1}^J x_j$$

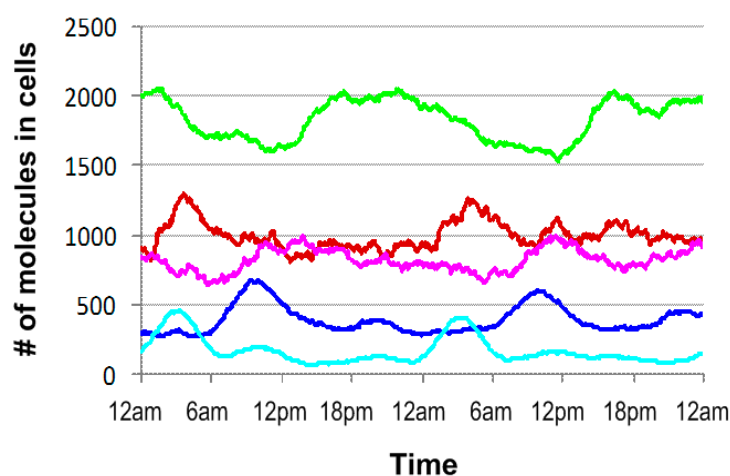
If the percentage change between the first and last half of the sample vector  $\frac{m_2 - m_1}{m_1}$  is more than 10%, we determine that the change is significant and adjust the corresponding production parameter. If the trend of the dynamics is increasing, the parameter value  $p_x$  will be decreased. Otherwise, if the trend of the dynamics is decreasing, we increase  $p_x$ . In order to estimate the changing amount of  $p_x$ , we assume that the percentage change of the parameter would be approximately to the percentage change of the molecule level between the first and last half of the sample vector but set under the opposite effect. Therefore, the estimate for the adjusted parameter value will be

$$\frac{p'_x - p_x}{p_x} = -\frac{m_2 - m_1}{m_1} \Rightarrow p'_x = p_x \left( 1 - \frac{m_2 - m_1}{m_1} \right)$$

In the case that there are two associated production parameters, the amount each parameter is changed will be half of that in the normal case. The process is repeated until there is no change of all production parameters in three consecutive simulated days.

### 5.3. Qualitative assessment of model behaviors with experimental observations

Circadian rhythms play an important role in many physiological and metabolic processes in almost all organisms. In mammals, it is recognized that a bidirectional communication between circadian controls and the immune system exists, and that glucocorticoids and melatonin are important hormones that show strong circadian expression patterns and play critical roles in mediating cytokine production [427-429]. Since melatonin and cortisol are associated with the production of pro-inflammatory and anti-inflammatory cytokines respectively, their expression rhythms will contribute to the dynamic patterns of cytokine expression [394-396, 429], resulting in the rhythms of P and A as observed in **Figure 5.2**.



**Figure 5.2:** Dynamics patterns of selected components under circadian control. Circadian control is regulated by the rhythms of cortisol (F) and melatonin (M) which in turn drive the patterns of other components in the system. Pro-inflammatory cytokines (P) driven by melatonin secretion are up-regulated to reach the peak ~4:00AM whereas anti-inflammatory cytokines (A) are down regulated due to the increase of pro-inflammatory cytokines and then up-regulated under the effects of cortisol rhythms. These behaviors result in the circadian variation of bio-energetic proteins (E) and others.

Pro-inflammatory cytokines (e.g.  $\text{TNF}\alpha$ , IL6,  $\text{IFN}\gamma$ ) are regulated in part by melatonin, reaching a maximum early in the morning and subsequently subsiding as cortisol levels induce the production of anti-inflammatory cytokines (e.g. IL10). As pro- and anti-inflammatory cytokines have opposing effects on cellular immunity, changes in their concentration and thus their balance would be anticipated to influence host fitness. Additionally, since transcription in the nucleus requires energy, each time a nucleus produces a new unit beyond default system production, the corresponding cell will exhaust some unit of energy, representing by the deletion of one bio-energetic protein (E) in this simulation. Consequently, energy balance and/or energetic protein abundance relevant to metabolic processes also exhibit daily circadian variations [451]. These observations provide a validation for our model's behaviors.

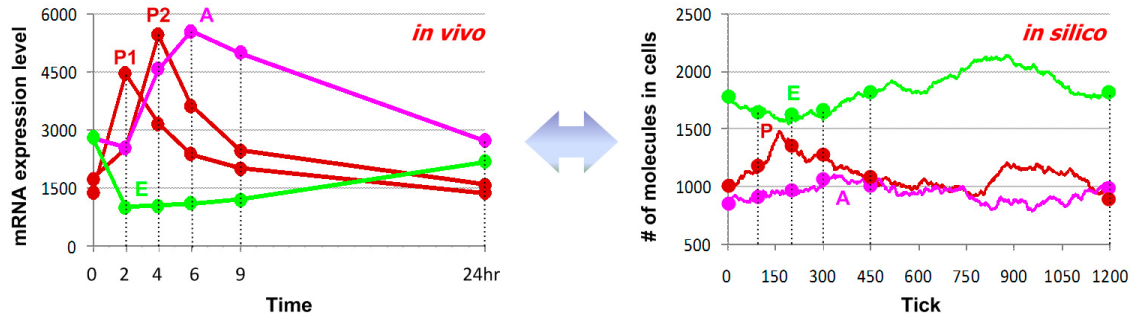
*In silico* administration of endotoxin is simulated by 'injecting' a number of new LPS molecules into the system at tick  $T$  which is corresponding to time  $(T \bmod N_{tpd})/N_{tph}$  of the day. In order to simulate *in vivo* endotoxin administration at 9:00am, we introduced 1000 LPS molecules randomly to the plasma compartment at the corresponding tick and track the cellular responses. Due to lack of information to evaluate the corresponding

dose and influence of other system factors (e.g. cell density) to the actual effects of those LPS molecules, we measure the effective concentration of LPS in our system by a

definition as follows:  $C_{eff} = \frac{1}{V_C} \max_t \left\{ \frac{\sum_i LPS_i(t)}{N_C} \right\}$ ,  $i = 1, \dots, N_C$ ,  $t = T, \dots, T + 2N_{tph}$ .  $N_C$  is

the total number of simulated cells (50 in this study) and  $LPS_i(t)$  is the number of LPSR molecules in cell  $i$  at time  $t$ .  $V_C$  is the volume of cells which is equal to  $30 \times 40 = 1200$  in this context. The effective concentration of this experiment is about 0.33%. The current default time-scale  $N_{tph}$  is 20 ticks per hour as discussed in the ‘Model parameters’ section. Since this time-scale calibration does not provide a corresponding mapping of the times between *in vivo* and *in silico* inflammatory responses, we vary  $N_{tph}$  to search for a timing match between *in vivo* and *in silico* patterns by gradually increasing the number of ticks per hour to 30, 40, 50, etc. The search ends up with a new time-scale  $N_{tph} = 50$ .

Main inflammatory responses of *in vivo* and *in silico* human endotoxemia are presented in **Figure 5.3**. Following endotoxin treatment, the pro-inflammatory response exhibits a fast and robust increase, peaking between 2 and 4hr after treatment and eventually resuming normal rhythms. The anti-inflammatory response which is normally down-regulated around mid-day keeps increasing following LPS administration. The systemic energy balance also continues to reduce for around 2hr more before returning to its normal rhythm. The system resumes normal daily rhythms about 24h post LPS administration.



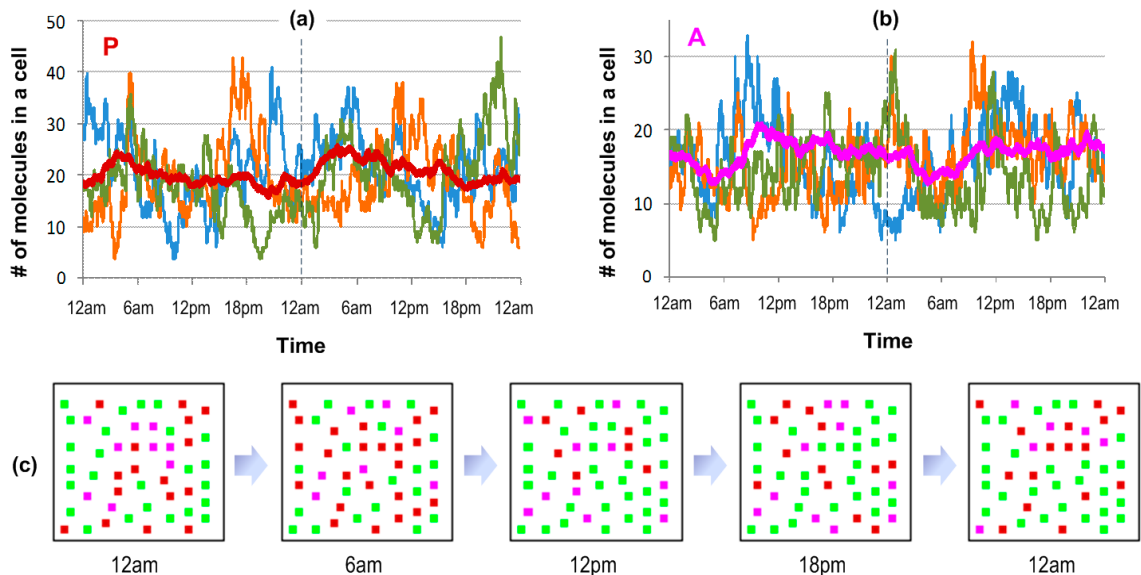
**Figure 5.3:** Correspondence between *in vivo*- and *in silico*- system responses to endotoxin. The left-panel presents average expression patterns of critical inflammatory responses under endotoxin treatment at 9:00AM. Early-up (red) and middle-up (black) patterns are characterized for pro-inflammatory responses, late-up pattern (magenta) for anti-inflammatory responses, and down pattern (green) for energetic responses. The right-panel displays corresponding simulated responses. The patterns between *in vivo*- and *in silico*- responses are matched to define the time-scale for the system.

## 5.4. Cellular variability and stochastic behaviors

### 5.4.1. Variability-based fitness

Since stochasticity is an inherent property of our individual-based simulation, stochastic transcriptional activities especially those relevant to the NF $\kappa$ B-signaling module have large impacts on cellular variability [452-454]. Simulated cells behave differently from one to another and no individual cell behaves like the average one. For example, dynamics patterns of pro- and anti-inflammatory protein levels oscillate stochastically between different cells and even different days although their average patterns exhibit some common daily patterns (**Figure 5.4a, b**). In general, these patterns are similar to corresponding system responses. Specifically, the average level of pro-inflammatory cytokines is induced early due to the increasing level of melatonin at the onset of the day

and then gradually abates while the level of cortisol increases. The average level of anti-inflammatory cytokines is transiently down regulated and then starts increasing to restore the balance between pro- and anti-inflammatory cytokines under the opposing effects and acutely altered patterns of melatonin and cortisol. From the system perspective, we assume that a cell will be (1) in the pro-inflammatory state (expressed by red squares) if the level of pro- is much greater than the level of anti-inflammatory cytokines ( $P > 1.5A$ ), (2) in the anti-inflammatory state (expressed by magenta squares) if  $A > 1.5P$ , and (3) otherwise in the homeostatic state (expressed by green squares). Interestingly, the status change of the cellular system also follows a common daily pattern although the status of a single cell is always dynamic over time, even for the same time the next day (**Figure 5.4c**). At the beginning of a day, pro-inflammatory cells predominate and then make room for anti-inflammatory cells in the late morning. Since the status of the cellular system is in some part associated with the protein abundance level of corresponding cytokine types, the balance between pro- and anti-inflammatory cytokines is anticipated to be dynamic over time but follow some common daily pattern.

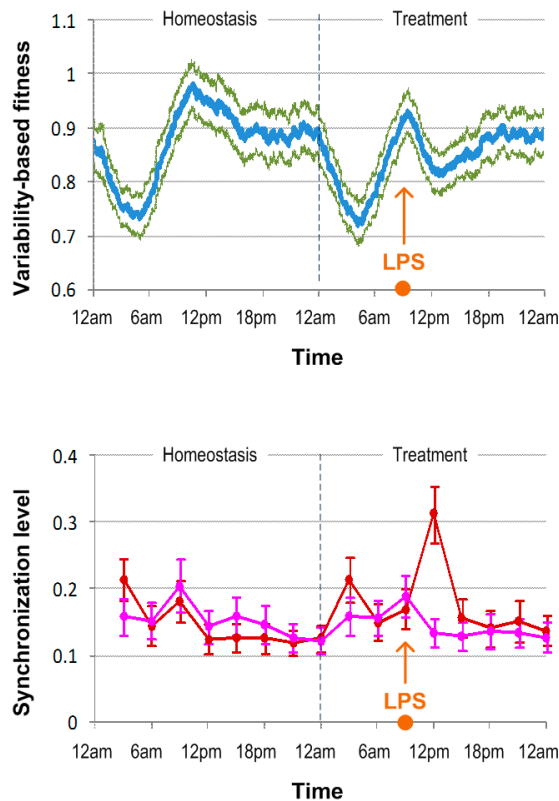




**Figure 5.4:** Stochastic dynamics in cell population. The stochastic behaviors of pro-inflammatory cytokines (a) and anti-inflammatory cytokines (b) in three different cells are shown in the top-panel. Although cellular patterns are different from cell to cell and from day to day, the average pattern still exhibits some daily common pattern. The dynamics of the homeostatic system in a simulated day are present in (c). Cells are displayed with solid squares where green squares represent for cells with an approximate number of P and A, red squares for those with the number of P much greater than the number of A and magenta squares for those with  $A \gg P$ .

Recent studies have implied that there is an association between patient fitness and the balance between the levels of pro- and anti-inflammatory cytokines [455, 456]. However, the protein abundance level in a population of genetically identical cells is proportional to the expression variance of the corresponding protein [457-459]. Consequently, the cell-to-cell variability potentially conveys information beyond the simple mean level of protein abundance in characterizing the dynamic kinetics of the entire system at the single cell level. Cellular variability can account for the stochastic transcriptional activities and thus not only the consequence but also the mechanisms that lead to the fluctuation of a protein between cells. As a result, we hereby define a novel quantity to characterize the entire status of the system in homeostasis or under treated conditions, so-called the variability-based fitness ( $F_{\text{var}}$ ), based on the ratio between the expression variance of anti-inflammatory cytokines and pro-inflammatory cytokines from the population of simulated leukocytes. In order to characterize the cytokine expression variance among cells, we utilize Shannon entropy to estimate the cellular variability based on the distribution of pro- or anti-inflammatory cytokines through the cell

population (see Materials and methods). This measurement somewhat reflects changes in host fitness, since the anti-inflammatory arm characterizes for the ‘fitness’ restoration and the pro-inflammatory arm serves for the ‘fitness’ dysregulation. In homeostasis, the ratio is anticipated to remain at some optimal level while its normal rhythm has some daily common fluctuations in the first half of a day due to the circadian secretion of melatonin and cortisol (**Figure 5.5-top**). Following endotoxin treatment (at 9:00AM in this case), the variability-based fitness immediately reduces to the minimum point around 3-4hr post injection and then gradually returns to the optimal level when the systemic manifestation of endotoxin abates, implying that the affect of endotoxin treatment can be quantified through this method.



**Figure 5.5:** Cellular variability and synchronization behaviors. The top-panel displays the pattern of variability-based fitness of a simulated day in the homeostatic system and of

the day where endotoxin is treated at 9:00AM. Two parallel curves present corresponding standard errors of  $N$  times of simulation ( $N=100$  in this study). The bottom panel shows the synchronization level of specific behaviors among all cells of the system in the interval  $[t - 3\text{hr}, t]$ ,  $t = 3, 6 \dots 24\text{hr}$ . The error bars are corresponding standard errors of  $N$  times of simulation.

#### ***Definition of the variability-based fitness ( $F_{\text{var}}$ )***

Given  $N_c$  cells, let  $x_i(t)$  be the number of molecules  $X$  in cell  $i$ ,  $i = 1, \dots, N_c$  at a specific time  $t$ . Since the distribution of  $x_i(t)$  values may be sparse, we first contract the range of  $x_i(t)$  by a whole number division of  $r$  for all  $x_i(t)$  ( $r = 5$  in this study).

$$y_i = \left\lfloor \frac{x_i(t)}{r} \right\rfloor, i = 1, \dots, N_c$$

Let  $p(y_i)$  be the probability of the presence of  $y_i$  value in the contracted array  $y = \{y_i, i = 1, \dots, N_c\}$ . The variability-based fitness  $F_{\text{var}}$  at a specific time  $t$  is defined as follows

$$F_{\text{var}}(t) = \frac{V_{C2C}^A}{V_{C2C}^P} \quad \text{where } V_{C2C}^X = - \sum_{i=1}^{N_c} p(y_i) \ln p(y_i), X = [A, P]$$

#### **5.4.2. Synchronization**

Even in the presence of the type of the large variability observed in some molecules from cell kinetics observed in the population of cells, external stimulus signals (e.g.  $\text{TNF}\alpha$ ) can cause cell synchronization for a short period of time [452, 453]. The synchronization behavior of cellular responses is therefore examined to get an insight into how pro- and anti-inflammatory cytokines act under endotoxin treatments. Quantitatively, the

synchronization level of a response (e.g. a molecule type) is defined as the average correlation coefficient between all individual response patterns of cells and the average response pattern of the cell population in a period of time (e.g. 3hr in this study) (see Materials and methods). LPS-induced cell synchronization has been examined for pro- and anti-inflammatory responses (**Figure 5.5-bottom**). Although the cellular pro-inflammatory responses are different from cell to cell, under an external stimulus their responses expose an increment of similarity in the first period after the treatment. However, anti-inflammatory responses among cells do not propose a significant trend of synchronization. This phenomenon results from the fact that all cells follow the only path that activates the NF $\kappa$ B-signaling module to produce pro-inflammatory cytokines under the primary stimulus signal while the path to produce anti-inflammatory cytokines is secondary and set under the effects of pro-inflammatory inhibitors. After the first period, stochastic oscillations resume in the population of cells although the systemic manifestation of inflammation does not quite abate.

### ***Definition of the synchronization***

Let  $x_i(t)$  be the number of molecules  $X$  in cell  $i$  and  $\bar{x}(t)$  be the average number of molecules  $X$  from  $N_c$  cells at time  $t$ . The synchronization level of molecule  $X$  in the population of cells for a period of time from 0:00 to 3:00AM is

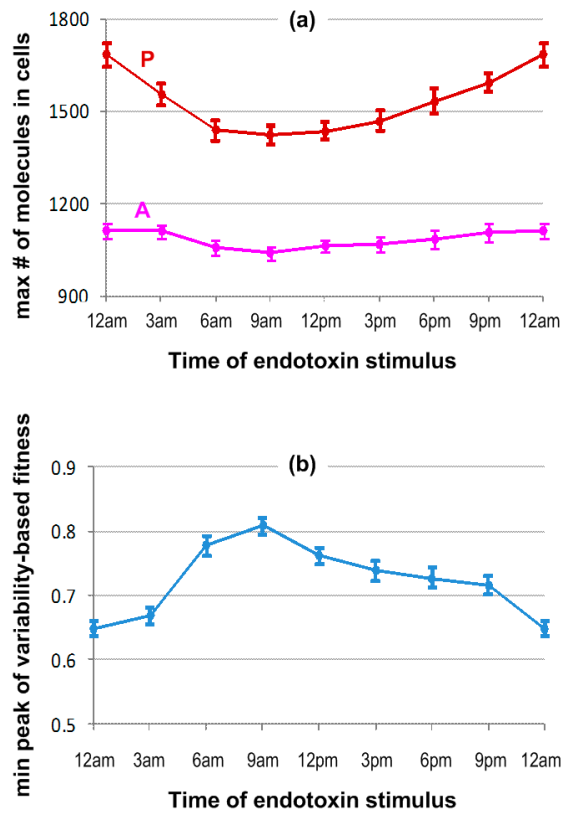
$$\begin{aligned}
Syn(0 \rightarrow 3AM) &= \frac{1}{N_C} \sum_i \varphi(x_i(t), xx(t)) \\
\text{where } \varphi(x_i(t), xx(t)) &= \frac{\sum_t (x_i(t) - \bar{x})(xx(t) - \overline{xx})}{\sqrt{\sum_t (x_i(t) - \bar{x})^2 \sum_t (xx(t) - \overline{xx})^2}} \\
xx(t) &= \frac{1}{N_C} \sum_i x_i(t); \quad \bar{x} = \frac{1}{T} \sum_t x_i(t); \quad \overline{xx} = \frac{1}{T} \sum_t xx(t) \\
i &= 1, \dots, N_C, t = 1, \dots, T = 3N_{ph}
\end{aligned}$$

## 5.5. Other relevant issues

### 5.5.1. Time-dependent effects under endotoxin treatment

As observed in previous studies, there are clearly significant effects of circadian rhythms on the dosing time in therapeutic treatments. For instance, ‘low dose prednisolone has more effect on rheumatoid arthritis at 2:00AM than at 7:00AM’ [394, 396] and ‘bedtime dosing with nifedipine gastrointestinal therapeutic system for antihypertensive medications is more effective than morning dosing’ [460]. We therefore explore the time-dependent effects of endotoxin administration by executing *in silico* experiments with endotoxin injection at different times of the day (3hr intervals from 0 to 24hr). We quantitatively examined the peaks of inflammatory responses following endotoxin administration at different times throughout the day. Results are characterized by the maximum numbers of pro- and anti-inflammatory cytokines as well as the dysregulation peak of the variability-based fitness versus the treated times of endotoxin (**Figure 5.6**). Simulation shows that endotoxin administrated in the morning (around 9:00AM) has the least pronounced effect, while the largest response occurs around midnight. Although the maximum numbers of anti-inflammatory cytokines in different cases seem to be approximately equal, there is a significant trend in the effects of administration times of

endotoxin on the production of pro-inflammatory cytokines. Characterizing these phenomena is the change of the variability-based fitness versus the administration time, implying somehow the loss of the host fitness. Periods of highly vulnerable effects are those around the midnight peak of melatonin secretion where the production of pro-inflammatory cytokines is set under two paths, NF $\kappa$ B-signaling and the melatonin-induced pathway. On the contrary, high concentration of plasma cortisol in the morning provides an inhibition to the activation of the NF $\kappa$ B-signaling module, resulting in the reduced effects of endotoxin administration.



**Figure 5.6:** Time-dependence system responses to endotoxin administration. The strength of the inflammatory response or the vulnerability of the host fitness is characterized by (a) the maximal peak of pro-inflammatory cytokines ( $P_{\max}$ ) and (b) the

minimum peak of variability-based fitness versus the time of endotoxin treatment. The error bars are corresponding standard errors of N times of simulation (N=100).

### 5.5.2. Sensitivity analysis

Sensitivity analysis was performed to explore how the perturbations in production parameter values affect the overall system behavior which is characterized by the variability-based fitness. Following previous studies [399, 419], we sequentially perturbed each production parameter and estimated the sensitivity coefficient which is defined as the percentage change of the fitness ( $DF_{\text{var}}$ ) over the percentage change of the parameter ( $Dp = \delta p / p$  where  $\delta p$  is the changing amount of parameter p) (see Materials and methods). In this case, 75% is selected as the cutoff to have a clear impact on the percentage change of the fitness, which is estimated from 10 simulated days with circadian controls and without external stimuli. Results are showed in **Table 5.4**. Two parameters that have great impact are  $k_i$  and  $f_a$  respectively where  $k_i$  is responsible for I $\kappa$ B production from the NF $\kappa$ B activities and  $f_a$  is directly responsible for the production of anti-inflammatory cytokines. Since the sensitivity coefficient is mainly relied on the change of the variability-based fitness where the dynamics of pro- and anti-inflammatory cytokines take place, parameters relevant to the production of these cytokines should have large impact. However, since  $k_i$  affects I $\kappa$ B production from NF $\kappa$ B activities in nucleus which in turn directly control back NF $\kappa$ B activities in regulating the production of pro-inflammatory cytokines, a small change on the value of  $k_i$  can have a large impact on the regulation of pro-inflammatory cytokine production. Therefore,  $k_p$  and  $m_p$ , two parameters directly relevant to the production of pro-inflammatory cytokines, have lesser impacts on the variability-based fitness than  $k_i$  does.

**Table 5.4:** Effects of production parameters on system behaviors

No.	Parameters	Change*	DF <sub>var</sub> /Dp	Change*	DF <sub>var</sub> /Dp
1	κp	↓	0.2085	↑	0.2461
2	κi	↓	<b>0.4481</b>	↑	<b>0.2679</b>
3	fi	↓	0.1861	↑	0.2520
4	fa	↓	<b>0.3260</b>	↑	<b>0.3485</b>
5	fm	↓	0.1560	↑	0.1440
6	mp	↓	0.1883	↑	0.2452
7	pf	↓	0.1607	↑	0.1585
8	pt	↓	0.1391	↑	0.1429
9	ae	↓	0.1425	↑	0.3303

\*: decrease/increase 75% of the current value; if greater than 1.0, set to 1.0.

### ***Definition of the percentage change of fitness***

In order to evaluate how a change impacts to the system behaviors, we define a so-called percentage change of the fitness as a ratio of the total changing amount between the variability-based fitness of the original system and that of the new system over the total amount in the original system during a period of time ( $n_{day} = 10$  in this study)

$$DF_{var} = \frac{\sum_t |F'_{var}(t) - F_{var}(t)|}{\sum_t F_{var}(t)}; t = 1, \dots, n_{day} \times N_{tpd}$$

where  $F_{var}(t)$  and  $F'_{var}(t)$  are the variability-based fitness at time  $t$  of the original system and the new system (e.g. the system with new parameter values).

## **5.6. Conclusions**

We have proposed a multi-level homeostatic system of human endotoxemia using the individual-based simulation to examine the dynamic kinetics of the inflammatory response at the single cell level under circadian control and endotoxin treatment. The



model naturally captures the stochastic and discrete nature of biological processes; specifically, it models the transcriptional dynamics at the cellular level and the linking of processes at multiple scales. Physicochemical properties of biological molecules and cellular properties have been incorporated to construct the model. Novel solutions for parameter tuning and time-scale estimation are also proposed to refine the model. The model is validated by its ability to reproduce *in vivo* homeostatic circadian rhythms and capture critical inflammatory responses under endotoxin treatment.

One of the most critical questions raised here is what information cellular variability can contribute to clinical implications. By defining novel hypothetical quantities such as the variability-based fitness and the synchronization level, we provided a step forward to the exploration of cell-to-cell variability and stochastic dynamics of inflammatory proteins. Daily common patterns of such measurements in homeostatic and LPS-treated systems are examined. Furthermore, the effects of time-dependent endotoxin administration characterized by variability-based fitness and the synchronization level of inflammatory cytokines are also studied. Although a full understanding of how cell-to-cell variability impacts clinical symptoms and pharmacological treatments is beyond the scope of this manuscript, proposed concepts in this study may actually be applicable in the near future as single-cell studies become increasingly common. Also, the proposed framework provides an effective model to generate testable hypotheses for a number of ‘*what if*’ scenarios to understand the connectivity of critical components in the immune system and the interplay between circadian controls and endotoxin treatments.

## Chapter 6 – Summary and Future Perspectives

### 6.1. Summary

The advance of high-throughput technologies has enabled a new generation of massive amounts of biological data, facilitating a dramatic increase as well as challenges in the degree of quantification applied to modern biological research. In this dissertation, we explore alternative- and/or propose novel- hypotheses delving into the complexity of high-dimensional datasets with the aim of extracting critical components and rules that govern their behaviors as well as the mysteries and complexities of the transcriptional regulatory gene network. The unifying hypothesis is that what is observed is usually an outcome of orchestrated interactions between critical modules in the form of a network. Therefore, our overall goal for this study is set for the development of bioinformatics tools and systems biology approaches towards the analysis and modeling of transcriptional dynamics and the understanding of gene regulatory network. Two *in vivo* models, namely corticosteroid pharmacogenomics in rat and human endotoxemia in human, have been investigated.

First, we examined the complexity of high-dimensional transcriptional responses under corticosteroid administration. As we all know that glucocorticoids are a class of steroid hormones present in almost every animal cell and play a central role in a wide range of physiological responses. Because of their potent anti-inflammatory and immunosuppressive effects, synthetic glucocorticoids referred as corticosteroids have been used widely in pharmacology as a therapeutic option for a wide range of autoimmune and inflammatory diseases. However, beneficial effects are derived from

magnifying the physiological actions of endogenous glucocorticoids, causing a variety of side effects following long-term treatment with this kind of drugs. The physiological and pharmacological effects of corticosteroids are complex and manifest themselves with expression changes of many genes across multiple tissues. As such, we ask that whether we can explore the complexity of gene expression changes to provide a better understanding of corticosteroid pharmacogenomic effects or understand how the drug alter systemic physiology and contribute to adverse-effects within individual tissues and across multiple tissues.

In the meantime, we also put efforts to explore the human endotoxemia model which is a well accepted surrogate model for studying acute inflammation and elicits significant dynamic inflammatory transcriptional responses. To gain a better understanding of the molecular mechanisms and physiological significance associated with inflammatory responses, clinically relevant models have been proposed including the human endotoxemia model in which an intravenous administration of *E.coli* endotoxin is given to healthy human subjects. Bacterial endotoxin, a component of the outer cell membrane of gram-negative bacteria, is an important mediator in the pathophysiology of gram-negative bacterial sepsis. This complex macromolecule induces its injurious effects by a non-cytotoxic interaction with CD14-bearing inflammatory cells, such as macrophage-monocytes, circulating neutrophils and lung epithelial cells. These effector cells are activated through a family of Toll-like receptors (TLR) and subsequently release a network of inflammatory products. While we do not argue that the human endotoxin challenge model precisely replicates an acute infectious or sepsis condition, we believe that human endotoxin challenge does serve as a useful model of TLR4 agonist-induced

systemic inflammation while at the same time providing a reproducible experimental platform.

In order to discover potential rules hidden in high-dimensional transcriptional profiles, we start by handling the uncertainty in microarray experimental measurements. It has been noticed that experimental data usually contain potential sources of uncertainty and thus many experiments are now designed with repeated measurements to better assess such inherent variability. Several computational methods have been proposed to account for the variability in replicates. As yet, there is no model to output expression profiles accounting for replicate information so that a variety of computational models that take the expression profiles as the input data can explore this information without any modification. Thus we proposed a methodology which integrates replicate variability into expression profiles, to generate so-called ‘true’ expression profiles. The clustering effectiveness when using ‘true’ profiles coupled with clustering techniques has been demonstrated through synthetic data where several models with the error information integrated are compared.

We next explore the hypothesis that the more clusterable the data is the more biologically relevant it is and utilize the concepts of consensus clustering to identify, within a set of differentially expressed genes, a subset of genes that are either highly co-expressed or highly non-coexpressed with the hope of extracting a more biologically relevant subset of genes. The purpose of this approach is to enable a systematic identification of smaller, clusterable, subsets of gene expression data exploring the concept of consensus clustering. The fundamental assumption of our approach is that an appropriate weighting of multiple alternative methods would eliminate the biases associated with specific

clustering methods. Also, it must be emphasized that the proposed framework is not designed, or proposed, in order to replace more refined clustering analysis, but is advocated as a critical preliminary steps in order to identify putatively informative subsets of genes given a high-dimensional expression dataset.

Additionally, we also proposed a framework to identify significant coexpressed clusters of genes across multiple datasets. Following the orientation of meta-analysis, an extended computational approach that explores the concept of agreement matrix from consensus clustering has been proposed with the aims of identifying gene clusters that share common expression patterns across multiple dosing regimens as well as handling challenges in the analysis of microarray data from heterogeneous sources, e.g. different platforms and time-grids in this study. Analysis on rich *in vivo* datasets of corticosteroid time-series yielded significant insights into the pharmacogenomic effects of corticosteroids, especially the relevance to metabolic side-effects. This has been illustrated through enriched metabolic functions in those transcriptional modules and the presence of GRE binding motifs in those enriched pathways, providing significant modules for further analysis on pharmacogenomic corticosteroid effects.

After identification of significant gene sets representing for critical transcriptional responses within individual or across multiple conditions/tissues, we ask that whether we can go one-level further up to understand more about those relevant to the regulation of these transcriptional responses. Consequently, we have developed computational strategies with the aim of providing significant insights into the potential regulatory interactions among transcriptional factors and their target genes which is a crucial step towards quantitative modelling of transcriptional regulatory networks. One of the key

features in our analysis is the identification of significantly overrepresented CRMs in each gene battery. Since these recognized CRMs are located on the control regions of many ‘hypothetically’ co-regulated genes, they are likely to be composed of functional binding sites that are activated upon the initiation of the transcriptional machinery. Furthermore, our analyses also allow for the reconstruction of a dynamic temporal regulatory network, making it a critical enabler for improving our understanding of how the transcriptional machinery ‘program’ effectively regulates key cellular processes.

Finally, we proposed a multi-level homeostatic system of human endotoxemia using the individual-based simulation to examine the dynamic kinetics of inflammatory responses at the single cell level under circadian controls and endotoxin treatments. The model naturally captures the stochastic and discrete nature of biological processes; especially the stochasticity of the transcriptional dynamics, one of the main reasons leading to phenotypic variations, at the cellular level and the linking of processes at multiple scales. Physicochemical properties of biological molecules and cellular properties have been incorporated to construct the model. With novel hypothetical quantities such as the variability-based fitness and the synchronization level, we provided a step forward to the exploration of cell-to-cell variability and predictive implications inferred from cellular variability.

## **6.2. Future perspectives**

Gene transcription is one of the main biological processes that govern an organism’s response to external stimuli. Unraveling the mysteries and complexities of transcriptional regulation is of paramount importance in modern biology. What causes a stem cell to commit to a particular lineage, what makes a cell response to an external perturbation or

an organism to a drug is largely determined by the transcriptional machinery that is the control mechanisms that dictate the up- or down-regulation of genes. In that context, the part of the non-coding regions of genes located upstream the transcription start site holds the keys to this mystery. The main theater of controlling transcriptional regulation is hypothesized by the interplay between TFs and their associated TFBSs located on the proximal promoters of target genes.

Our analysis has attempted to reverse engineer the underlying regulatory network under external stimuli e.g. the human blood leukocyte response to endotoxin. Given the transcriptional profiling data, an elementary set of temporal responses with putative transcriptional regulators have been identified. A key feature of the analysis is the exploration of the concept 'gene battery' which represents for a group of genes that are both co-expressed and functionally relevant to identify inflammation-relevant TFs using a context-specific searching approach [373]. Although no single analysis can identify all regulators involved in a response, it has been demonstrated that the proposed framework can identify critical TFs that are relevant to acute inflammatory responses. Despite the fact that many methods have been proposed in the literature to search for relevant transcriptional regulators, different approaches explore different biological assumptions resulting to different sets of putative TFs which may or may not significantly overlap each other. Since the true extent of all TFs involved in the regulation of a complex response under some external stimuli is unknown, these differences could not be interpreted as the high- or low- accuracy of the methods. Instead, all of found TFs may be involved in different processes of the response but because of the limitation of hypotheses used by the methods, they may not be recognized by a certain approach.

Novel methods are still proposed using different analytical approaches but generally they can be categorized into two main directions including mRNA expression-based and TF binding pattern-based methods. The first direction somehow utilizes the fundamental hypothesis that the mRNA expression level of TFs is proportional to their protein concentration but this may not be appropriate especially in higher eukaryotes since TF activation is often regulated post-translationally and acts somewhat in an independent manner of expression level. Some methods also require multiple-condition data as the input which may not be applicable when practical data are only sampled under one condition/treatment. In the meanwhile, a lot of methods following to the latter direction have been developed of which ours is among them. These are not limited by the mRNA expression proportion hypothesis but they are limited by promoter identification, TF binding profiles, and the underlying assumption to select the input set of ‘co-regulated’ genes. Consequently, given the future availability of more complete TF binding data and other resources, the method could be enhanced by integrating protein-protein interaction to refine selected CRMs. Since each single method or even each direction always contains its own limitations and advantages, one possibility in future improvements could be the development of a framework to obtain a consensus result under diverse underlying hypotheses from various outputs of different methods.

Previous analyses bring back an overview as well as a better understanding of how the system responses under external stimuli. To examine cellular behaviors and regulatory mechanisms, more specifically the interplay between circadian control and endotoxin challenge, we construct an *in silico* human endotoxemia that can mimic critical aspects of the physiological human endotoxemia model. The model naturally captures the stochastic



and discrete nature of biological processes; especially the stochasticity of the transcriptional dynamics, one of the main reasons leading to phenotypic variations, at the cellular level and the linking of processes at multiple scales. One of the most critical questions we ask is what information cellular variability can contribute to clinical implications. By defining novel hypothetical quantities e.g. the so-called variability-based fitness and the synchronization level, we provided a step forward to the exploration of cell-to-cell variability and stochastic dynamics of inflammatory proteins as well as predictive implications relevant to clinical outcomes.

The proposed framework provides an effective approach to generate testable hypotheses for a number of '*what if*' scenarios to understand the connectivity of critical components in the immune system and the interplay between circadian controls and endotoxin treatments. In future studies, we may examine in detail the impacts of system default parameters e.g. cell-population, molecule-interaction blocking on the system behaviors concerning homeostatic daily patterns and inflammatory responses under endotoxin administration. Eventually, our work aims at establishing the core of an extensive model with multiple body-in-systems including activities of different inflammatory cytokines, specific mechanisms of critical immune cell types (e.g. macrophages, T-cells, dendritic cells), and ultimately specific mechanisms of immune-relevant systems (e.g. the immune system, the central nervous system, the stress system, the cardiovascular system) to explore systemic responses and clinical implications of inflammatory diseases.

## Bibliography

1. Rhen T, Cidlowski JA: **Antiinflammatory action of glucocorticoids--new mechanisms for old drugs.** *N Engl J Med* 2005, **353**(16):1711-1723.
2. Barnes PJ: **Corticosteroid effects on cell signalling.** *Eur Respir J* 2006, **27**(2):413-426.
3. Baxter JD: **Advances in glucocorticoid therapy.** *Adv Intern Med* 2000, **45**:317-349.
4. Bialas MC, Routledge PA: **Adverse effects of corticosteroids.** *Adverse Drug React Toxicol Rev* 1998, **17**(4):227-235.
5. Frauman AG: **An overview of the adverse reactions to adrenal corticosteroids.** *Adverse Drug React Toxicol Rev* 1996, **15**(4):203-206.
6. Schacke H, Docke WD, Asadullah K: **Mechanisms involved in the side effects of glucocorticoids.** *Pharmacol Ther* 2002, **96**(1):23-43.
7. Locsey L, Asztalos L, Kincses Z, Gyorfi F, Berczi C: **Dyslipidaemia and hyperlipidaemia following renal transplantation.** *Int Urol Nephrol* 1996, **28**(3):419-430.
8. Almon RR, Dubois DC, Jin JY, Jusko WJ: **Pharmacogenomic responses of rat liver to methylprednisolone: an approach to mining a rich microarray time series.** *Aaps J* 2005, **7**(1):E156-194.
9. Almon RR, DuBois DC, Piel WH, Jusko WJ: **The genomic response of skeletal muscle to methylprednisolone using microarrays: tailoring data mining to the structure of the pharmacogenomic time series.** *Pharmacogenomics* 2004, **5**(5):525-552.
10. Almon RR, Lai W, DuBois DC, Jusko WJ: **Corticosteroid-regulated genes in rat kidney: mining time series array data.** *Am J Physiol Endocrinol Metab* 2005, **289**(5):E870-882.
11. Almon RR, DuBois DC, Jusko WJ: **A microarray analysis of the temporal response of liver to methylprednisolone: a comparative analysis of two dosing regimens.** *Endocrinology* 2007, **148**(5):2209-2225.
12. Almon RR, DuBois DC, Yao Z, Hoffman EP, Ghimbovski S, Jusko WJ: **Microarray analysis of the temporal response of skeletal muscle to methylprednisolone: comparative analysis of two dosing regimens.** *Physiol Genomics* 2007, **30**(3):282-299.
13. Yao Z, Hoffman EP, Ghimbovski S, Dubois DC, Almon RR, Jusko WJ: **Mathematical modeling of corticosteroid pharmacogenomics in rat muscle following acute and chronic methylprednisolone dosing.** *Mol Pharm* 2008, **5**(2):328-339.
14. Ramakrishnan R, DuBois DC, Almon RR, Pyszczynski NA, Jusko WJ: **Pharmacodynamics and pharmacogenomics of methylprednisolone during 7-day infusions in rats.** *J Pharmacol Exp Ther* 2002, **300**(1):245-256.
15. Sun YN, DuBois DC, Almon RR, Jusko WJ: **Fourth-generation model for corticosteroid pharmacodynamics: a model for methylprednisolone effects on receptor/gene-mediated glucocorticoid receptor down-regulation and tyrosine aminotransferase induction in rat liver.** *J Pharmacokinet Biopharm* 1998, **26**(3):289-317.

16. Dong Y, Poellinger L, Gustafsson JA, Okret S: **Regulation of glucocorticoid receptor expression: evidence for transcriptional and posttranslational mechanisms.** *Mol Endocrinol* 1988, **2**(12):1256-1264.
17. Oakley RH, Cidlowski JA: **Homologous down regulation of the glucocorticoid receptor: the molecular machinery.** *Crit Rev Eukaryot Gene Expr* 1993, **3**(2):63-88.
18. Vedeckis WV, Ali M, Allen HR: **Regulation of glucocorticoid receptor protein and mRNA levels.** *Cancer Res* 1989, **49**(8 Suppl):2295s-2302s.
19. Almon RR, DuBois DC, Brandenburg EH, Shi W, Zhang S, Straubinger RM, Jusko WJ: **Pharmacodynamics and pharmacogenomics of diverse receptor-mediated effects of methylprednisolone in rats using microarray analysis.** *J Pharmacokinet Pharmacodyn* 2002, **29**(2):103-129.
20. Sun YN, DuBois DC, Almon RR, Pyszczynski NA, Jusko WJ: **Dose-dependence and repeated-dose studies for receptor/gene-mediated pharmacodynamics of methylprednisolone on glucocorticoid receptor down-regulation and tyrosine aminotransferase induction in rat liver.** *J Pharmacokinet Biopharm* 1998, **26**(6):619-648.
21. Nystrom PO: **The systemic inflammatory response syndrome: definitions and aetiology.** *J Antimicrob Chemother* 1998, **41** Suppl A:1-7.
22. Annane D, Bellissant E, Cavaillon JM: **Septic shock.** *Lancet* 2005, **365**(9453):63-78.
23. Hotchkiss RS, Karl IE: **The pathophysiology and treatment of sepsis.** *N Engl J Med* 2003, **348**(2):138-150.
24. Tetta C, Fonsato V, Ronco C, Camussi G: **Recent insights into the pathogenesis of severe sepsis.** *Crit Care Resusc* 2005, **7**(1):32-39.
25. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR: **Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care.** *Crit Care Med* 2001, **29**(7):1303-1310.
26. Balk RA: **Optimum treatment of severe sepsis and septic shock: evidence in support of the recommendations.** *Dis Mon* 2004, **50**(4):168-213.
27. Poeze M, Ramsay G, Gerlach H, Rubulotta F, Levy M: **An international sepsis survey: a study of doctors' knowledge and perception about sepsis.** *Crit Care* 2004, **8**(6):R409-413.
28. Heron M, Hoyert DL, Murphy SL, Xu J, Kochanek KD, Tejada-Vera B: **Deaths: final data for 2006.** *Natl Vital Stat Rep* 2009, **57**(14):1-134.
29. Annane D, Aegerter P, Jars-Guincestre MC, Guidet B: **Current epidemiology of septic shock: the CUB-Rea Network.** *Am J Respir Crit Care Med* 2003, **168**(2):165-172.
30. Martin GS, Mannino DM, Eaton S, Moss M: **The epidemiology of sepsis in the United States from 1979 through 2000.** *N Engl J Med* 2003, **348**(16):1546-1554.
31. Levy MM, Dellinger RP, Townsend SR, Linde-Zwirble WT, Marshall JC, Bion J, Schorr C, Artigas A, Ramsay G, Beale R *et al*: **The Surviving Sepsis Campaign: results of an international guideline-based performance improvement program targeting severe sepsis.** *Crit Care Med* 2010, **38**(2):367-374.

32. Heyland DK, Hopman W, Coe H, Tranmer J, McColl MA: **Long-term health-related quality of life in survivors of sepsis. Short Form 36: a valid and reliable measure of health-related quality of life.** *Crit Care Med* 2000, **28**(11):3599-3605.
33. Longo CJ, Heyland DK, Fisher HN, Fowler RA, Martin CM, Day AG: **A long-term follow-up study investigating health-related quality of life and resource use in survivors of severe sepsis: comparison of recombinant human activated protein C with standard care.** *Crit Care* 2007, **11**(6):R128.
34. Deans KJ, Haley M, Natanson C, Eichacker PQ, Minneci PC: **Novel therapies for sepsis: a review.** *J Trauma* 2005, **58**(4):867-874.
35. Freeman BD, Natanson C: **Anti-inflammatory therapies in sepsis and septic shock.** *Expert Opin Investig Drugs* 2000, **9**(7):1651-1663.
36. Riedemann NC, Guo RF, Ward PA: **Novel strategies for the treatment of sepsis.** *Nat Med* 2003, **9**(5):517-524.
37. Zeni F, Freeman B, Natanson C: **Anti-inflammatory therapies to treat sepsis and septic shock: a reassessment.** *Crit Care Med* 1997, **25**(7):1095-1100.
38. Minneci PC, Deans KJ, Banks SM, Eichacker PQ, Natanson C: **Meta-analysis: the effect of steroids on survival and shock during sepsis depends on the dose.** *Ann Intern Med* 2004, **141**(1):47-56.
39. Cohen J: **Adjunctive therapy in sepsis: a critical analysis of the clinical trial programme.** *Br Med Bull* 1999, **55**(1):212-225.
40. Opal SM, Gluck T: **Endotoxin as a drug target.** *Crit Care Med* 2003, **31**(1 Suppl):S57-64.
41. Johnson GB, Brunn GJ, Platt JL: **Cutting edge: an endogenous pathway to systemic inflammatory response syndrome (SIRS)-like reactions through Toll-like receptor 4.** *J Immunol* 2004, **172**(1):20-24.
42. Johnson GB, Brunn GJ, Samstein B, Platt JL: **New insight into the pathogenesis of sepsis and the sepsis syndrome.** *Surgery* 2005, **137**(4):393-395.
43. Gao H, Evans TW, Finney SJ: **Bench-to-bedside review: sepsis, severe sepsis and septic shock - does the nature of the infecting organism matter?** *Crit Care* 2008, **12**(3):213.
44. Vincent JL, Sun Q, Dubois MJ: **Clinical trials of immunomodulatory therapies in severe sepsis and septic shock.** *Clin Infect Dis* 2002, **34**(8):1084-1093.
45. Eichacker PQ, Parent C, Kalil A, Esposito C, Cui X, Banks SM, Gerstenberger EP, Fitz Y, Danner RL, Natanson C: **Risk and the efficacy of antiinflammatory agents: retrospective and confirmatory studies of sepsis.** *Am J Respir Crit Care Med* 2002, **166**(9):1197-1205.
46. Reinhart K, Karzai W: **Anti-tumor necrosis factor therapy in sepsis: update on clinical trials and lessons learned.** *Crit Care Med* 2001, **29**(7 Suppl):S121-125.
47. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, Steingrub JS, Garber GE, Helterbrand JD, Ely EW *et al*: **Efficacy and safety of recombinant human activated protein C for severe sepsis.** *N Engl J Med* 2001, **344**(10):699-709.
48. Kerschen EJ, Fernandez JA, Cooley BC, Yang XV, Sood R, Mosnier LO, Castellino FJ, Mackman N, Griffin JH, Weiler H: **Endotoxemia and sepsis**

- mortality reduction by non-anticoagulant activated protein C. *J Exp Med* 2007, **204**(10):2439-2448.
49. Greven D, Leng L, Bucala R: **Autoimmune diseases: MIF as a therapeutic target.** *Expert Opin Ther Targets* 2010, **14**(3):253-264.
  50. Wang H, Zhu S, Zhou R, Li W, Sama AE: **Therapeutic potential of HMGB1-targeting agents in sepsis.** *Expert Rev Mol Med* 2008, **10**:e32.
  51. Guo RF, Ward PA: **Role of C5a in inflammatory responses.** *Annu Rev Immunol* 2005, **23**:821-852.
  52. Guo RF, Ward PA: **C5a, a therapeutic target in sepsis.** *Recent Pat Antiinfect Drug Discov* 2006, **1**(1):57-65.
  53. Parrish WR, Gallowitsch-Puerta M, Czura CJ, Tracey KJ: **Experimental therapeutic strategies for severe sepsis: mediators and mechanisms.** *Ann N Y Acad Sci* 2008, **1144**:210-236.
  54. Lolis E, Bucala R: **Therapeutic approaches to innate immunity: severe sepsis and septic shock.** *Nat Rev Drug Discov* 2003, **2**(8):635-645.
  55. Silva E, Passos Rda H, Ferri MB, de Figueiredo LF: **Sepsis: from bench to bedside.** *Clinics (Sao Paulo)* 2008, **63**(1):109-120.
  56. Remick DG: **Pathophysiology of sepsis.** *Am J Pathol* 2007, **170**(5):1435-1444.
  57. Neugebauer EA, Willy C, Sauerland S: **Complexity and non-linearity in shock research: reductionism or synthesis?** *Shock* 2001, **16**(4):252-258.
  58. Tjardes T, Neugebauer E: **Sepsis research in the next millennium: concentrate on the software rather than the hardware.** *Shock* 2002, **17**(1):1-8.
  59. **Food and Drug Administration. Innovation or stagnation: challenge and opportunity on the critical path to new medical products.** 2004, **1**:1-38.
  60. Beutler B, Rietschel ET: **Innate immune sensing and its roots: the story of endotoxin.** *Nat Rev Immunol* 2003, **3**(2):169-176.
  61. Opal SM, DePalo VA: **Anti-inflammatory cytokines.** *Chest* 2000, **117**(4):1162-1172.
  62. Nathan C: **Points of control in inflammation.** *Nature* 2002, **420**(6917):846-852.
  63. Santos AA, Wilmore DW: **The systemic inflammatory response: perspective of human endotoxemia.** *Shock* 1996, **6 Suppl 1**:S50-56.
  64. Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, Cho RJ, Chen RO, Brownstein BH, Cobb JP, Tschoeke SK *et al*: **A network-based analysis of systemic inflammation in humans.** *Nature* 2005, **437**(7061):1032-1037.
  65. Talwar S, Munson PJ, Barb J, Fiuza C, Cintron AP, Logun C, Tropea M, Khan S, Reda D, Shelhamer JH *et al*: **Gene expression profiles of peripheral blood leukocytes after endotoxin challenge in humans.** *Physiol Genomics* 2006, **25**(2):203-215.
  66. Copeland S, Warren HS, Lowry SF, Calvano SE, Remick D: **Acute inflammatory response to endotoxin in mice and humans.** *Clin Diagn Lab Immunol* 2005, **12**(1):60-67.
  67. Lowry SF: **Human endotoxemia: a model for mechanistic insight and therapeutic targeting.** *Shock* 2005, **24 Suppl 1**:94-100.
  68. Van Zee KJ, Coyle SM, Calvano SE, Oldenburg HS, Stiles DM, Pribble J, Catalano M, Moldawer LL, Lowry SF: **Influence of IL-1 receptor blockade on the human response to endotoxemia.** *J Immunol* 1995, **154**(3):1499-1507.

69. van Deventer SJ, Buller HR, ten Cate JW, Aarden LA, Hack CE, Sturk A: **Experimental endotoxemia in humans: analysis of cytokine release and coagulation, fibrinolytic, and complement pathways.** *Blood* 1990, **76**(12):2520-2526.
70. An G: **Agent-based computer simulation and sirs: building a bridge between basic science and clinical trials.** *Shock* 2001, **16**(4):266-273.
71. An G, Mi Q, Dutta-Moscato J, Vodovotz Y: **Agent-based models in translational systems biology.** *WIREs* 2009, **1**:159-171.
72. Vodovotz Y, Constantine G, Faeder J, Mi Q, Rubin J, Bartels J, Sarkar J, Squires RH, Jr., Okonkwo DO, Gerlach J *et al*: **Translational systems approaches to the biology of inflammation and healing.** *Immunopharmacol Immunotoxicol* 2010, **32**(2):181-195.
73. Vodovotz Y, Constantine G, Rubin J, Csete M, Voit EO, An G: **Mechanistic simulations of inflammation: current state and future prospects.** *Math Biosci* 2009, **217**(1):1-10.
74. An G, Hunt CA, Clermont G, Neugebauer E, Vodovotz Y: **Challenges and rewards on the road to translational systems biology in acute illness: four case reports from interdisciplinary teams.** *J Crit Care* 2007, **22**(2):169-175.
75. An G, Faeder J, Vodovotz Y: **Translational systems biology: introduction of an engineering approach to the pathophysiology of the burn patient.** *J Burn Care Res* 2008, **29**(2):277-285.
76. Vodovotz Y, Csete M, Bartels J, Chang S, An G: **Translational systems biology of inflammation.** *PLoS Comput Biol* 2008, **4**(4):e1000014.
77. An G: **Introduction of an agent-based multi-scale modular architecture for dynamic knowledge representation of acute inflammation.** *Theor Biol Med Model* 2008, **5**:11.
78. Foteinou PT, Calvano SE, Lowry SF, Androulakis IP: **Translational Potential of Systems-based Models of Inflammation** *Clinical and Translational Science* 2008 (accepted for publication).
79. Chow CC, Clermont G, Kumar R, Lagoa C, Tawadrous Z, Gallo D, Betten B, Bartels J, Constantine G, Fink MP *et al*: **The acute inflammatory response in diverse shock states.** *Shock* 2005, **24**(1):74-84.
80. Vodovotz Y, Clermont G, Chow C, An G: **Mathematical models of the acute inflammatory response.** *Curr Opin Crit Care* 2004, **10**(5):383-390.
81. Scholl HJ: **Agent based and system dynamics modeling: a call for cross study and joint research.** *The 34th Hawaii International Conference on System Sciences* 2001.
82. Day J, Rubin J, Vodovotz Y, Chow CC, Reynolds A, Clermont G: **A reduced mathematical model of the acute inflammatory response II. Capturing scenarios of repeated endotoxin administration.** *J Theor Biol* 2006, **242**(1):237-256.
83. Foteinou PT, Calvano SE, Lowry SF, Androulakis IP: **Modeling endotoxin-induced systemic inflammation using an indirect response approach.** *Math Biosci* 2009, **217**(1):27-42.

84. Foteinou PT, Calvano SE, Lowry SF, Androulakis IP: **In silico simulation of corticosteroids effect on an NFkB- dependent physicochemical model of systemic inflammation.** *PLoS One* 2009, **4**(3):e4706.
85. Foteinou PT, Calvano SE, Lowry SF, Androulakis IP: **A Multi-scale Model for the Assessment of Autonomic Dysfunction in Human Endotoxemia.** *Physiol Genomics* 2010:doi: 0.1152/physiolgenomics.00184.02009
86. Reynolds A, Rubin J, Clermont G, Day J, Vodovotz Y, Bard Ermentrout G: **A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation.** *J Theor Biol* 2006, **242**(1):220-236.
87. Vodovotz Y: **Deciphering the complexity of acute inflammation using mathematical models.** *Immunol Res* 2006, **36**(1-3):237-245.
88. Wakeland WW, Gallaher EJ, Macovsky LM, Aktipis CA: **A comparison of system dynamics and agent-based simulation applied to the study of cellular receptor dynamics.** *The 37th Hawaii International Conference on System Sciences* 2004.
89. Goldstein B, Faeder JR, Hlavacek WS: **Mathematical and computational models of immune-receptor signalling.** *Nat Rev Immunol* 2004, **4**(6):445-456.
90. Brown EN, Meehan PM, Dempster AP: **A stochastic differential equation model of diurnal cortisol patterns.** *Am J Physiol Endocrinol Metab* 2001, **280**(3):E450-461.
91. Klerman EB, Adler GK, Jin M, Maliszewski AM, Brown EN: **A statistical model of diurnal variation in human growth hormone.** *Am J Physiol Endocrinol Metab* 2003, **285**(5):E1118-1126.
92. An G: **In silico experiments of existing and hypothetical cytokine-directed clinical trials using agent-based modeling.** *Crit Care Med* 2004, **32**(10):2050-2060.
93. Li NY, Verdolini K, Clermont G, Mi Q, Rubinstein EN, Hebda PA, Vodovotz Y: **A patient-specific in silico model of inflammation and healing tested in acute vocal fold injury.** *PLoS One* 2008, **3**(7):e2789.
94. Mi Q, Riviere B, Clermont G, Steed DL, Vodovotz Y: **Agent-based model of inflammation and wound healing: insights into diabetic foot ulcer pathology and the role of transforming growth factor-beta1.** *Wound Repair Regen* 2007, **15**(5):671-682.
95. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**(2):101-113.
96. Lemon B, Tjian R: **Orchestrated response: a symphony of transcription factors for gene control.** *Genes Dev* 2000, **14**(20):2551-2569.
97. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**(6945):147-151.
98. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome.** *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
99. Jin JY, Almon RR, DuBois DC, Jusko WJ: **Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays.** *J Pharmacol Exp Ther* 2003, **307**(1):93-109.

100. Pavlidis P: **Using ANOVA for gene selection from microarray studies of the nervous system.** *Methods* 2003, **31**(4):282-289.
101. Nguyen TT, Nowakowski RS, Androulakis IP: **Unsupervised Selection of Highly Coexpressed and Noncoexpressed Genes Using a Consensus Clustering Approach.** *Omics* 2009.
102. Almon RR, Yang E, Lai W, Androulakis IP, DuBois DC, Jusko WJ: **Circadian variations in rat liver gene expression: relationships to drug actions.** *J Pharmacol Exp Ther* 2008, **326**(3):700-716.
103. Cobb JP, Mindrinos MN, Miller-Graziano C, Calvano SE, Baker HV, Xiao W, Laudanski K, Brownstein BH, Elson CM, Hayden DL *et al*: **Application of genome-wide expression analysis to human health and disease.** *Proc Natl Acad Sci U S A* 2005, **102**(13):4801-4806.
104. Nguyen TT, Nowakowski RS, Androulakis IP: **Unsupervised selection of highly coexpressed and noncoexpressed genes using a consensus clustering approach.** *Omics* 2009, **13**(3):219-237.
105. Yeung KY, Medvedovic M, Bumgarner RE: **Clustering gene-expression data with repeated measurements.** *Genome Biol* 2003, **4**(5):R34.
106. Roach JC, Smith KD, Strobe KL, Nissen SM, Haudenschild CD, Zhou D, Vasicek TJ, Held GA, Stolovitzky GA, Hood LE *et al*: **Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells.** *Proc Natl Acad Sci U S A* 2007, **104**(41):16245-16250.
107. Pei Y ZO: **A Synthetic Data Generator for Clustering and Outlier Analysis.** *Technical report, University of Alberta*, 2006:TR06-15.
108. Medvedovic M, Yeung KY, Bumgarner RE: **Bayesian mixture model based clustering of replicated microarray data.** *Bioinformatics* 2004, **20**(8):1222-1232.
109. Ideker T TV, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner RE, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systemically perturbed metabolic network.** *Science* 2001, **292**:929-934.
110. Yao J, Chang C, Salmi ML, Hung YS, Loraine A, Roux SJ: **Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient.** *BMC Bioinformatics* 2008, **9**:288.
111. Genomatix: <http://www.genomatix.de>.
112. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21**(13):2933-2942.
113. Altman N: **Replication, variation and normalisation in microarray experiments.** *Appl Bioinformatics* 2005, **4**(1):33-44.
114. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32 Suppl**:490-495.
115. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci U S A* 2000, **97**(18):9834-9839.



116. Lonnstedt I, Speed T: **Replicated microarray data**. *Statistica Sinica* 2002, **12**:31-46.
117. Pan W, Lin J, Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach**. *Genome Biol* 2002, **3**(5):research0022.
118. Pavlidis P, Li Q, Noble WS: **The effect of replication on gene expression microarray experiments**. *Bioinformatics* 2003, **19**(13):1620-1627.
119. Rocke DM, Durbin B: **A model for measurement error for gene expression arrays**. *J Comput Biol* 2001, **8**(6):557-569.
120. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data**. *Bioinformatics* 2002, **18 Suppl 1**:S105-110.
121. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18 Suppl 1**:S96-104.
122. Lin SM, Du P, Huber W, Kibbe WA: **Model-based variance-stabilizing transformation for Illumina microarray data**. *Nucleic Acids Res* 2008, **36**(2):e11.
123. Motakis ES, Nason GP, Fryzlewicz P, Rutter GA: **Variance stabilization and normalization for one-color microarray data using a data-driven multiscale approach**. *Bioinformatics* 2006, **22**(20):2547-2553.
124. Celeux G, Martin O, Lavergne C: **Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments**. *Statistical Modelling* 2005, **5**(3):243-267.
125. Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng SW: **A mixture model with random-effects components for clustering correlated gene-expression profiles**. *Bioinformatics* 2006, **22**(14):1745-1752.
126. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments**. *Bioinformatics* 2002, **18**(4):546-554.
127. Begun A: **Power estimation of the t test for detecting differential gene expression**. *Funct Integr Genomics* 2008, **8**(2):109-113.
128. de Menezes RX, Boer JM, van Houwelingen HC: **Microarray data analysis: a hierarchical T-test to handle heteroscedasticity**. *Appl Bioinformatics* 2004, **3**(4):229-235.
129. McCarthy DJ, Smyth GK: **Testing significance relative to a fold-change threshold is a TREAT**. *Bioinformatics* 2009, **25**(6):765-771.
130. Churchill GA: **Using ANOVA to analyze microarray data**. *Biotechniques* 2004, **37**(2):173-175, 177.
131. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116-5121.
132. Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW: **Significance analysis of time course microarray experiments**. *Proc Natl Acad Sci U S A* 2005, **102**(36):12837-12842.

133. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al*: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **102**(1):109-126.
134. Medvedovic M, Sivaganesan S: **Bayesian infinite mixture model based clustering of gene expression profiles**. *Bioinformatics* 2002, **18**(9):1194-1206.
135. Tjaden B: **An approach for clustering gene expression data with error information**. *BMC Bioinformatics* 2006, **7**:17.
136. Phang TL, Neville MC, Rudolph M, Hunter L: **Trajectory clustering: a non-parametric method for grouping gene expression time courses, with applications to mammary development**. *Pac Symp Biocomput* 2003:351-362.
137. Matsumoto S, Aisaki K, Kanno J: **Mass distributed clustering: a new algorithm for repeated measurements in gene expression data**. *Genome Inform* 2005, **16**(2):183-194.
138. Asyali MH, Colak D, Demirkaya O, Inan MS: **Gene expression profile classification: a review**. *Current Bioinformatics* 2006, **1**:55-73.
139. Fraley, Raftery A: **mclust: Model-Based Clustering / Normal Mixture Modeling**. *R packages* 2007.
140. Yan J: **som: Self-Organizing Map**. *R packages* 2004.
141. Yang E, Maguire T, Yarmush ML, Berthiaume F, Androulakis IP: **Bioinformatics analysis of the early inflammatory response in a rat thermal injury model**. *BMC Bioinformatics* 2007, **8**:10.
142. **R Development Core Team: The R stats package**. *R packages* 2008.
143. Jang RJ: **DCPR (Data Clustering and Pattern Recognition) Toolbox**. <http://www.wcsnhtuedutw/~jang>.
144. Maechler M, Rousseeuw P, Struyf A, Hubert M: **cluster: Cluster Analysis Basics and Extensions**. *R packages* 2005.
145. Hubert L, Arabie P: **Comparing partitions**. *J Classification* 1985, **2**(1):193-218.
146. Eptter S KM, Zaki M: **Clusterability detection and initial seed selection in large datasets**. *Technical report, Rensselaer Polytechnic Institute* 1999.
147. Bolshakova N, Azuaje F: **Estimating the number of clusters in DNA microarray data**. *Methods Inf Med* 2006, **45**(2):153-157.
148. Dudoit S, Fridlyand J: **A prediction-based resampling method for estimating the number of clusters in a dataset**. *Genome Biol* 2002, **3**(7):RESEARCH0036.
149. Monti S TP, Mesirov J, Golub T: **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Mach Learn* 2003, **52**:91-118.
150. Ressom H, Wang D, Natarajan P: **Adaptive double self-organizing maps for clustering gene expression profiles**. *Neural Netw* 2003, **16**(5-6):633-640.
151. Sharan R, Shamir R: **CLICK: a clustering algorithm with applications to gene expression analysis**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:307-316.
152. Tibshirani R, Walther G, Hastie T: **Estimating the Number of Clusters in a Data Set via the Gap Statistic**. *J Royal Statistical Society* 2001, **63**(2):411-423.
153. Yan M, Ye K: **Determining the number of clusters using the weighted gap statistic**. *Biometrics* 2007, **63**(4):1031-1037.
154. Yu Z, Wong HS, Wang H: **Graph-based consensus clustering for class discovery from gene expression data**. *Bioinformatics* 2007, **23**(21):2888-2896.

155. Ester M KH, Sander J, Xu X: **A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.** *Proc KDD' 96, AAAI Press, Menlo Park* 1996:226-231.
156. Belacel N, Wang Q, Cuperlovic-Culf M: **Clustering methods for microarray gene expression data.** *OMICS* 2006, **10**(4):507-531.
157. Huang D, Pan W: **Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data.** *Bioinformatics* 2006, **22**(10):1259-1268.
158. Hunter L, Taylor RC, Leach SM, Simon R: **GEST: a gene expression search tool based on a novel Bayesian similarity metric.** *Bioinformatics* 2001, **17 Suppl 1**:S115-122.
159. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: **Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.** *J Mol Biol* 2001, **314**(5):1053-1066.
160. Chipman H, Tibshirani R: **Hybrid hierarchical clustering with applications to microarray data.** *Biostatistics* 2006, **7**(2):286-301.
161. van der Laan MJ PK: **Hybrid clustering of gene expression data with visualization and the bootstrap.** *J of Stat Planning and Inference* 2003, **117**:275-303.
162. Kustra R ZA: **Incorporating Gene Ontology in Clustering Gene Expression Data.** *IEEE on Computer-Based Medical Systems, CBMS' 06* 2006:555-563.
163. Fang Z, Yang J, Li Y, Luo Q, Liu L: **Knowledge guided analysis of microarray data.** *J Biomed Inform* 2006, **39**(4):401-411.
164. Pan W: **Incorporating gene functions as priors in model-based clustering of microarray gene expression data.** *Bioinformatics* 2006, **22**(7):795-801.
165. Munneke B, Schlauch KA, Simonsen KL, Beavis WD, Doerge RW: **Adding confidence to gene expression clustering.** *Genetics* 2005, **170**(4):2003-2011.
166. Zhang K, Zhao H: **Assessing reliability of gene clusters from gene expression data.** *Funct Integr Genomics* 2000, **1**(3):156-173.
167. Strehl A, Ghosh J: **Cluster Ensembles A Knowledge Reuse Framework for Combining Multiple Partitions.** *Journal on Machine Learning Research* 2002, **3**:583-617.
168. Grotkjaer T, Winther O, Regenber B, Nielsen J, Hansen LK: **Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm.** *Bioinformatics* 2006, **22**(1):58-67.
169. Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, Kellam P: **Consensus clustering and functional interpretation of gene-expression data.** *Genome Biol* 2004, **5**(11):R94.
170. Hirsch M, Swift S, Liu X: **Optimal search space for clustering gene expression data via consensus.** *J Comput Biol* 2007, **14**(10):1327-1341.
171. Laderas T, McWeeney S: **Consensus framework for exploring microarray data using multiple clustering methods.** *Omic*s 2007, **11**(1):116-128.
172. Topchy A, Jain AK, Punch W: **Clustering ensembles: models of consensus and weak partitions.** *IEEE Trans Pattern Anal Mach Intell* 2005, **27**(12):1866-1881.

173. Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A: **e1071: Misc Functions of the Department of Statistics. R packages** 2006.
174. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80.
175. Ihaka R, Gentleman R: **R: A Language for Data Analysis and Graphics.** *J Comp Graphical Statistics* 1996, **5**(3):299-314 (<http://www.R-project.org>).
176. Gibbons FD, Roth FP: **Judging the quality of gene expression-based clustering methods using gene annotation.** *Genome Res* 2002, **12**(10):1574-1581.
177. Gentleman RC, Carey V, Huber W: **genefilter: methods for filtering genes from microarray experiments.** *R packages*.
178. Dangalchev C: **Residual closeness in networks.** *Physica A* 2006, **365**(2):556-564.
179. Rand WM: **Objective criteria for the evaluation of clustering methods.** *J American Statistical Association* 1971, **66**:846-850.
180. Jiang D, Tang C, Zhang A: **Cluster analysis for gene expression data: a survey.** *IEEE Transactions on Knowledge and Data Engineering* 2004, **16**(11):1370-1386.
181. Friedman JH RL: **Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests.** *Ann Statist* 1979, **7**:697-717.
182. Smith SP JA: **Testing for uniformity in multidimensional data.** *IEEE Trans Pattern Anal Machine Intell* 1984, **PAMI-6**(1):73-80.
183. Hardiman G: **Microarray platforms--comparisons and contrasts.** *Pharmacogenomics* 2004, **5**(5):487-502.
184. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O: **Are data from different gene expression microarray platforms comparable?** *Genomics* 2004, **83**(6):1164-1168.
185. Pedotti P, t Hoen PA, Vreugdenhil E, Schenk GJ, Vossen RH, Ariyurek Y, de Hollander M, Kuiper R, van Ommen GJ, den Dunnen JT *et al*: **Can subtle changes in gene expression be consistently detected with different microarray platforms?** *BMC Genomics* 2008, **9**:124.
186. Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV: **Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies.** *Bioinformatics* 2004, **20**(17):3166-3178.
187. Morris JS, Yin G, Baggerly KA, Wu C, Zhang L: **Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer.** *Methods of Microarray Data Analysis IV* 2005, New York: Springer-Verlag:51-66.
188. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G *et al*: **Multiple-laboratory comparison of microarray platforms.** *Nat Methods* 2005, **2**(5):345-350.
189. Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, Tsai CJ, Zhang S: **Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.** *BMC Bioinformatics* 2004, **5**:81.

190. Kim KY, Ki DH, Jeong HJ, Jeung HC, Chung HC, Rha SY: **Novel and simple transformation algorithm for combining microarray data sets.** *BMC Bioinformatics* 2007, **8**:218.
191. Park T, Yi SG, Shin YK, Lee S: **Combining multiple microarrays in the presence of controlling variables.** *Bioinformatics* 2006, **22**(14):1682-1689.
192. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB: **Merging two gene-expression studies via cross-platform normalization.** *Bioinformatics* 2008, **24**(9):1154-1160.
193. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z: **Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements.** *BMC Bioinformatics* 2005, **6**:107.
194. Lu J, Lee JC, Salit ML, Cam MC: **Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays.** *BMC Bioinformatics* 2007, **8**:108.
195. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z: **Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.** *Nucleic Acids Res* 2004, **32**(9):e74.
196. Morris JS, Wu C, Coombes KR, Baggerly KA, Wang J, Zhang L: **Alternative probeset definitions for combining microarray data across studies using different versions of affymetrix oligonucleotide arrays.** *Meta-Analysis in Genetics* 2006, New York: Chapman-Hall:1-21.
197. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18**(3):405-412.
198. Ramasamy A, Mondry A, Holmes CC, Altman DG: **Key issues in conducting a meta-analysis of gene expression microarray datasets.** *PLoS Med* 2008, **5**(9):e184.
199. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM: **Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.** *Funct Integr Genomics* 2003, **3**(4):180-188.
200. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling interstudy variation.** *Bioinformatics* 2003, **19** Suppl 1:i84-90.
201. Hu P, Greenwood CM, Beyene J: **Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.** *BMC Bioinformatics* 2005, **6**:128.
202. Stevens JR, Doerge RW: **Combining Affymetrix microarray results.** *BMC Bioinformatics* 2005, **6**:57.
203. Conlon EM, Song JJ, Liu A: **Bayesian meta-analysis models for microarray data: a comparative study.** *BMC Bioinformatics* 2007, **8**:80.
204. Liang Y, Kelemen A: **Bayesian models and meta analysis for multiple tissue gene expression data following corticosteroid administration.** *BMC Bioinformatics* 2008, **9**:354.

205. Nguyen TT, Almon RR, DuBois DC, Jusko WJ, Androulakis IP: **Importance of replication in analyzing time-series gene expression data: Corticosteroid dynamics and circadian patterns in rat liver.** *BMC Bioinformatics* 2010(accepted).
206. Tong W, Cao X, Harris S, Sun H, Fang H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R *et al*: **ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research.** *Environ Health Perspect* 2003, **111**(15):1819-1826.
207. Chow JC, Young DW, Golenbock DT, Christ WJ, Gusovsky F: **Toll-like receptor-4 mediates lipopolysaccharide-induced signal transduction.** *J Biol Chem* 1999, **274**(16):10689-10692.
208. Carmody RJ, Chen YH: **Nuclear factor-kappaB: activation and regulation during toll-like receptor signaling.** *Cell Mol Immunol* 2007, **4**(1):31-41.
209. Hotchkiss RS, Nicholson DW: **Apoptosis and caspases regulate death and inflammation in sepsis.** *Nat Rev Immunol* 2006, **6**(11):813-822.
210. Murray PJ: **The JAK-STAT signaling pathway: input and output integration.** *J Immunol* 2007, **178**(5):2623-2629.
211. Singer M, De Santis V, Vitale D, Jeffcoate W: **Multiorgan failure is an adaptive, endocrine-mediated, metabolic response to overwhelming systemic inflammation.** *Lancet* 2004, **364**(9433):545-548.
212. Kafatos FC: **A revolutionary landscape: the restructuring of biology and its convergence with medicine.** *J Mol Biol* 2002, **319**(4):861-867.
213. van Driel R, Fransz PF, Verschure PJ: **The eukaryotic genome: a system regulated at different hierarchical levels.** *J Cell Sci* 2003, **116**(Pt 20):4067-4075.
214. Werner T, Fessele S, Maier H, Nelson PJ: **Computer modeling of promoter organization as a tool to study transcriptional coregulation.** *Faseb J* 2003, **17**(10):1228-1237.
215. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16**(1):1-10.
216. Heintzman ND, Ren B: **The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome.** *Cell Mol Life Sci* 2007, **64**(4):386-400.
217. Barrera LO, Ren B: **The transcriptional regulatory code of eukaryotic cells--insights from genome-wide analysis of chromatin organization and transcription factor binding.** *Curr Opin Cell Biol* 2006, **18**(3):291-298.
218. Dillon N: **Gene regulation and large-scale chromatin organization in the nucleus.** *Chromosome Res* 2006, **14**(1):117-126.
219. Mateos-Langerak J, Goetze S, Leonhardt H, Cremer T, van Driel R, Lanctot C: **Nuclear architecture: Is it important for genome function and can we prove it?** *J Cell Biochem* 2007, **102**(5):1067-1075.
220. Schneider R, Grosschedl R: **Dynamics and interplay of nuclear architecture, genome organization, and gene expression.** *Genes Dev* 2007, **21**(23):3027-3043.

221. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20**(9):1377-1419.
222. Landry JR, Mager DL, Wilhelm BT: **Complex controls: the role of alternative promoters in mammalian genomes.** *Trends Genet* 2003, **19**(11):640-648.
223. Singer GA, Wu J, Yan P, Plass C, Huang TH, Davuluri RV: **Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array.** *BMC Genomics* 2008, **9**:349.
224. Sandve GK, Drablos F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct* 2006, **1**:11.
225. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5**(1):201.
226. Lewin B: **Gene IX - Promoters and Enhancers.** 2007, **ch.24**:609-635.
227. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM: **Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.** *Genome Res* 2006, **16**(1):1-10.
228. Maston GA, Evans SK, Green MR: **Transcriptional Regulatory Elements in the Human Genome.** *Annu Rev Genomics Hum Genet* 2006, **7**:29-59.
229. Butler JEF, Kadonaga JT: **The RNA polymerase II core promoter: a key component in the regulation of gene expression.** *Genes Dev* 2002, **16**(20):2583-2592.
230. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72**:449-479.
231. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes.** *Annu Rev Genet* 2000, **34**:77-137.
232. Périer RC, Junier T, Bucher P: **The Eukaryotic Promoter Database EPD.** *Nucleic Acids Res* 1997, **26**(1):353-357.
233. Périer RC, Praz V, Junier T, Bonnard C, Bucher P: **The eukaryotic promoter database (EPD).** *Nucleic Acids Res* 2000, **28**(1):302-303.
234. Suzuki Y, Yamashita R, Nakai K, Sugano S: **DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs.** *Nucleic Acids Res* 2002, **30**(1):328-331.
235. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K: **DBTSS: DataBase of Human Transcription Start Sites, progress report 2006.** *Nucleic Acids Res* 2007:1-5.
236. Gershenzon NI, Ioshikhes IP: **Synergy of human Pol II core promoter elements revealed by statistical sequence analysis.** *Bioinformatics* 2005, **21**(8):1295-1300.
237. Xi H, Yu Y, Fu Y, Foley J, Halees A, Weng Z: **Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1.** *Genome Res* 2007, **17**(6):798-806.
238. McKnight SL, Kingsbury R: **Transcriptional control signals of a eukaryotic protein-coding gene.** *Science* 1982, **217**:316-324.
239. Blackwood EM, Kadonaga JT: **Going the Distance: A Current View of Enhancer Action.** *Science* 1998, **281**(5373):60 - 63.

240. Li Q, Peterson KR, Fang X, Stamatoyannopoulos G: **Locus control regions.** *Blood* 2002, **100**(9):3077-3086.
241. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE: **High-resolution mapping and characterization of open chromatin across the genome.** *Cell* 2008, **132**(2):311-322.
242. Genomatix: <http://www.genomatix.de/>.
243. Scherf M, Klingenhoff A, Werner T: **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.** *J Mol Biol* 2000, **297**(3):599-606.
244. Bajic VB, Seah SH: **Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units.** *Genome Res* 2003, **13**(8):1923-1929.
245. Won HH, Kim MJ, Kim S, Kim JW: **EnsemPro: an ensemble approach to predicting transcription start sites in human genomic DNA sequences.** *Genomics* 2008, **91**(3):259-266.
246. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL: **Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment.** *Genome Biol* 2006, **7** Suppl 1:S3 1-13.
247. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **The biology of eukaryotic promoter prediction--a review.** *Comput Chem* 1999, **23**(3-4):191-207.
248. Qiu P: **Recent advances in computational promoter analysis in understanding the transcriptional regulatory network.** *Biochem Biophys Res Commun* 2003, **309**(3):495-501.
249. Werner T: **The state of the art of mammalian promoter recognition.** *Brief Bioinform* 2003, **4**(1):22-30.
250. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH: **The functional consequences of alternative promoter use in mammalian genomes.** *Trends Genet* 2008, **24**(4):167-177.
251. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8**(6):413-423.
252. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA: **Mammalian RNA polymerase II core promoters: insights from genome-wide studies.** *Nat Rev Genet* 2007, **8**(6):424-436.
253. Cornish-Bowden A: **Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.** *Nucleic Acids Res* 1985, **13**(9):3021-3030.
254. Stormo GD: **Consensus patterns in DNA.** *Methods Enzymol* 1990, **183**:211-221.
255. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23**(23):4878-4884.
256. Chambers A, Stanway C, Tsang JS, Henry Y, Kingsman AJ, Kingsman SM: **ARS binding factor 1 binds adjacent to RAP1 at the UASs of the yeast glycolytic genes PGK and PYK1.** *Nucleic Acids Res* 1990, **18**(18):5393-5399.
257. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16**(1):16-23.



258. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31**(13):3576-3579.
259. Salzberg SL: **A method for identifying splice sites and translational start sites in eukaryotic mRNA.** *Comput Appl Biosci* 1997, **13**(4):365-376.
260. Bulyk ML, Johnson PL, Church GM: **Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.** *Nucleic Acids Res* 2002, **30**(5):1255-1261.
261. Man TK, Stormo GD: **Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay.** *Nucleic Acids Res* 2001, **29**(12):2471-2478.
262. Ellrott K, Yang C, Sladek FM, Jiang T: **Identifying transcription factor binding sites through Markov chain optimization.** *Bioinformatics* 2002, **18** Suppl 2:S100-109.
263. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**(1):78-94.
264. Durbin R, Eddy SR, Krogh A, Mitchison G: **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.** *Cambridge University Press* 1998.
265. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.** *Bioinformatics* 2001, **17**(12):1113-1122.
266. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks.** *Bioinformatics* 2005, **21**(11):2657-2666.
267. Pavesi G, Mauri G, Pesole G: **In silico representation and discovery of transcription factor binding sites.** *Brief Bioinform* 2004, **5**(3):217-236.
268. Chekmenev DS, Haid C, Kel AE: **P-Match: transcription factor binding site search by combining patterns and weight matrices.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W432-437.
269. Gershenzon NI, Stormo GD, Ioshikhes IP: **Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites.** *Nucleic Acids Res* 2005, **33**(7):2290-2301.
270. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *J Mol Biol* 2004, **338**(2):207-215.
271. Hannenhalli S, Wang LS: **Enhanced position weight matrices using mixture models.** *Bioinformatics* 2005, **21** Suppl 1:i204-212.
272. Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.** *Genome Res* 2006, **16**(12):1455-1464.
273. Hertz GZ, Hartzell GW, 3rd, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6**(2):81-92.

274. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**(5131):208-214.
275. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
276. Das MK, Dai HK: **A survey of DNA motif finding algorithms.** *BMC Bioinformatics* 2007, **8 Suppl 7**:S21.
277. Tung NT, Yang E, Androulakis IP: **Machine learning approaches in promoter sequence analysis.** *Machine Learning Research Progress*, Nova Science Publishers, Inc 2008.
278. Marsan L, Sagot MF: **Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.** *J Comput Biol* 2000, **7**(3-4):345-362.
279. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**(7-8):563-577.
280. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34**(Database issue):D95-97.
281. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24**(1):238-241.
282. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome.** *Science* 2001, **291**(5507):1304-1351.
283. Brazma A, Jonassen I, Eidhammer I, Gilbert D: **Approaches to the automatic discovery of patterns in biosequences.** *J Comput Biol* 1998, **5**(2):279-305.
284. Friberg M, von Rohr P, Gonnet G: **Scoring functions for transcription factor binding site prediction.** *BMC Bioinformatics* 2005, **6**:84.
285. Li N, Tompa M: **Analysis of computational approaches for motif discovery.** *Algorithms Mol Biol* 2006, **1**:8.
286. Doniger SW, Huh J, Fay JC: **Identification of functional transcription factor binding sites using closely related *Saccharomyces* species.** *Genome Res* 2005, **15**(5):701-709.
287. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301**(5629):71-76.
288. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE *et al*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428**(6982):493-521.
289. Morgenstern B: **DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment.** *Bioinformatics* 1999, **15**(3):211-218.

290. Schmollinger M, Nieselt K, Kaufmann M, Morgenstern B: **DIALIGN P: fast pair-wise and multiple sequence alignment using parallel processors.** *BMC Bioinformatics* 2004, **5**:128.
291. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB: **Benchmarking tools for the alignment of functional noncoding DNA.** *BMC Bioinformatics* 2004, **5**:6.
292. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al*: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**(7004):99-104.
293. Lee HG, Lee HS, Jeon SH, Chung TH, Lim YS, Huh WK: **High-resolution analysis of condition-specific regulatory modules in *Saccharomyces cerevisiae*.** *Genome Biol* 2008, **9**:R2.
294. McCord RP, Berger MF, Philippakis AA, Bulyk ML: **Inferring condition-specific transcription factor function from DNA binding and gene expression data.** *Mol Syst Biol* 2007, **3**:100.
295. Smith AD, Sumazin P, Zhang MQ: **Tissue-specific regulatory elements in mammalian promoters.** *Mol Syst Biol* 2007, **3**:73.
296. Yu X, Lin J, Zack DJ, Qian J: **Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors.** *BMC Bioinformatics* 2007, **8**:437.
297. Fessele S, Maier H, Zischek C, Nelson PJ, Werner T: **Regulatory context is a crucial part of gene function.** *Trends Genet* 2002, **18**(2):60-63.
298. Allocco DJ, Kohane IS, Butte AJ: **Quantifying the relationship between co-expression, co-regulation and gene function.** *BMC Bioinformatics* 2004, **5**:18.
299. Long F, Liu H, Hahn C, Sumazin P, Zhang MQ, Zilberstein A: **Genome-wide prediction and analysis of function-specific transcription factor binding sites.** *In Silico Biol* 2004, **4**(4):395-410.
300. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16**(10):939-945.
301. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**(3):281-285.
302. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405**(6788):827-836.
303. Flintoft L: **Gene regulation: The many paths to coexpression.** *Nature Reviews Genetics* 2007, **8**:827.
304. Choi D, Fang Y, Mathers WD: **Condition-specific coregulation with cis-regulatory motifs and modules in the mouse genome.** *Genomics* 2006, **87**(4):500-508.
305. Huang R, Wallqvist A, Covell DG: **Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen.** *Genomics* 2006, **87**(3):315-328.
306. Britten RJ, Davidson EH: **Gene regulation for higher cells: a theory.** *Science* 1969, **165**(891):349-357.

307. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome.** *Proc Natl Acad Sci U S A* 2002, **99**(2):757-762.
308. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12**(7):1019-1028.
309. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo.** *BMC Bioinformatics* 2002, **3**:30.
310. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34**(2):166-176.
311. Ivan A, Halfon MS, Sinha S: **Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs.** *Genome Biol* 2008, **9**(1):R22.
312. Nelander S, Larsson E, Kristiansson E, Mansson R, Nerman O, Sigvardsson M, Mostad P, Lindahl P: **Predictive screening for regulators of conserved functional gene modules (gene batteries) in mammals.** *BMC Genomics* 2005, **6**(1):68.
313. Altman RB, Raychaudhuri S: **Whole-genome expression analysis: challenges beyond clustering.** *Curr Opin Struct Biol* 2001, **11**(3):340-347.
314. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.** *Genome Res* 2003, **13**(5):773-780.
315. Brown CD, Johnson DS, Sidow A: **Functional architecture and evolution of transcriptional elements that drive gene coexpression.** *Science* 2007, **317**(5844):1557-1560.
316. Hannonhalli S, Levy S: **Transcriptional regulation of protein complexes and biological pathways.** *Mamm Genome* 2003, **14**(9):611-619.
317. Balmer JE, Blomhoff R: **Anecdotes, data and regulatory modules.** *Biol Lett* 2006, **2**(3):431-434.
318. Davidson EH: **Genomic Regulatory Systems: Development and Evolution.** *Academic Press* 2001.
319. Klepper K, Sandve GK, Abul O, Johansen J, Drablos F: **Assessment of composite motif discovery methods.** *BMC Bioinformatics* 2008, **9**:123.
320. Frech K, Danescu-Mayer J, Werner T: **A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter.** *J Mol Biol* 1997, **270**(5):674-687.
321. Schones DE, Smith AD, Zhang MQ: **Statistical significance of cis-regulatory modules.** *BMC Bioinformatics* 2007, **8**:19.
322. Fridman JS, Lowe SW: **Control of apoptosis by p53.** *Oncogene* 2003, **22**(56):9030-9040.
323. Vousden KH, Lu X: **Live or let die: the cell's response to p53.** *Nat Rev Cancer* 2002, **2**(8):594-604.

324. Chipuk JE, Kuwana T, Bouchier-Hayes L, Droin NM, Newmeyer DD, Schuler M, Green DR: **Direct activation of Bax by p53 mediates mitochondrial membrane permeabilization and apoptosis.** *Science* 2004, **303**(5660):1010-1014.
325. Ding HF, Lin YL, McGill G, Juo P, Zhu H, Blenis J, Yuan J, Fisher DE: **Essential role for caspase-8 in transcription-independent apoptosis triggered by p53.** *J Biol Chem* 2000, **275**(49):38905-38911.
326. Moll UM, Wolff S, Speidel D, Deppert W: **Transcription-independent pro-apoptotic functions of p53.** *Curr Opin Cell Biol* 2005, **17**(6):631-636.
327. Caelles C, Helmborg A, Karin M: **p53-dependent apoptosis in the absence of transcriptional activation of p53-target genes.** *Nature* 1994, **370**(6486):220-223.
328. Wagner AJ, Kokontis JM, Hay N: **Myc-mediated apoptosis requires wild-type p53 in a manner independent of cell cycle arrest and the ability of p53 to induce p21waf1/cip1.** *Genes Dev* 1994, **8**(23):2817-2830.
329. Rodriguez-Caso C, Medina MA, Sole RV: **Topology, tinkering and evolution of the human transcription factor network.** *Febs J* 2005, **272**(24):6423-6434.
330. Gallant S, Gilkeson G: **ETS transcription factors and regulation of immunity.** *Arch Immunol Ther Exp (Warsz)* 2006, **54**(3):149-163.
331. Coffey PJ, Burgering BM: **Forkhead-box transcription factors and their role in the immune system.** *Nat Rev Immunol* 2004, **4**(11):889-899.
332. McKay LI, Cidlowski JA: **CBP (CREB binding protein) integrates NF-kappaB (nuclear factor-kappaB) and glucocorticoid receptor physical interactions and antagonism.** *Mol Endocrinol* 2000, **14**(8):1222-1234.
333. Sulser F: **The role of CREB and other transcription factors in the pharmacotherapy and etiology of depression.** *Ann Med* 2002, **34**(5):348-356.
334. Hutton JJ, Jegga AG, Kong S, Gupta A, Ebert C, Williams S, Katz JD, Aronow BJ: **Microarray and comparative genomics-based identification of genes and gene regulatory regions of the mouse immune system.** *BMC Genomics* 2004, **5**(1):82.
335. Shulman AI, Mangelsdorf DJ: **Retinoid x receptor heterodimers in the metabolic syndrome.** *N Engl J Med* 2005, **353**(6):604-615.
336. Nakae J, Oki M, Cao Y: **The FoxO transcription factors and metabolic regulation.** *FEBS Lett* 2008, **582**(1):54-67.
337. Solomon SS, Majumdar G, Martinez-Hernandez A, Raghoebar R: **A critical role of Sp1 transcription factor in regulating gene expression in response to insulin and other hormones.** *Life Sci* 2008, **83**(9-10):305-312.
338. Wan YJ, An D, Cai Y, Repa JJ, Hung-Po Chen T, Flores M, Postic C, Magnuson MA, Chen J, Chien KR *et al*: **Hepatocyte-specific mutation establishes retinoid X receptor alpha as a heterodimeric integrator of multiple physiological processes in the liver.** *Mol Cell Biol* 2000, **20**(12):4436-4444.
339. Aderem A, Smith KD: **A systems approach to dissecting immunity and inflammation.** *Semin Immunol* 2004, **16**(1):55-67.
340. Takeda K, Akira S: **Toll-like receptors in innate immunity.** *Int Immunol* 2005, **17**(1):1-14.

341. Frankenstein Z, Alon U, Cohen IR: **The immune-body cytokine network defines a social architecture of cell interactions.** *Biol Direct* 2006, **1**:32.
342. Wesche-Soldato DE, Swan RZ, Chung CS, Ayala A: **The apoptotic pathway as a therapeutic target in sepsis.** *Curr Drug Targets* 2007, **8**(4):493-500.
343. Barton GM, Medzhitov R: **Toll-like receptor signaling pathways.** *Science* 2003, **300**(5625):1524-1525.
344. Singer M, Brealey D: **Mitochondrial dysfunction in sepsis.** *Biochem Soc Symp* 1999, **66**:149-166.
345. Han J, Jiang Y, Li Z, Kravchenko VV, Ulevitch RJ: **Activation of the transcription factor MEF2C by the MAP kinase p38 in inflammation.** *Nature* 1997, **386**(6622):296-299.
346. Shapira M, Hamlin BJ, Rong J, Chen K, Ronen M, Tan MW: **A conserved role for a GATA transcription factor in regulating epithelial innate immune responses.** *Proc Natl Acad Sci U S A* 2006, **103**(38):14086-14091.
347. Serfling E, Avots A, Neumann M: **The architecture of the interleukin-2 promoter: a reflection of T lymphocyte activation.** *Biochim Biophys Acta* 1995, **1263**(3):181-200.
348. Taniguchi T: **Transcription factors IRF-1 and IRF-2: linking the immune responses and tumor suppression.** *J Cell Physiol* 1997, **173**(2):128-130.
349. Tak PP, Firestein GS: **NF-kappaB: a key role in inflammatory diseases.** *J Clin Invest* 2001, **107**(1):7-11.
350. Potthoff MJ, Olson EN: **MEF2: a central regulator of diverse developmental programs.** *Development* 2007, **134**(23):4131-4140.
351. Olson EN: **Undermining the endothelium by ablation of MAPK-MEF2 signaling.** *J Clin Invest* 2004, **113**(8):1110-1112.
352. Tantin D, Schild-Poulter C, Wang V, Hache RJ, Sharp PA: **The octamer binding transcription factor Oct-1 is a stress sensor.** *Cancer Res* 2005, **65**(23):10750-10758.
353. Schild-Poulter C, Shih A, Tantin D, Yarymowich NC, Soubeyrand S, Sharp PA, Hache RJ: **DNA-PK phosphorylation sites on Oct-1 promote cell survival following DNA damage.** *Oncogene* 2007, **26**(27):3980-3988.
354. Rehli M, Poltorak A, Schwarzfischer L, Krause SW, Andreesen R, Beutler B: **PU.1 and interferon consensus sequence-binding protein regulate the myeloid expression of the human Toll-like receptor 4 gene.** *J Biol Chem* 2000, **275**(13):9773-9781.
355. Kitada S, Pedersen IM, Schimmer AD, Reed JC: **Dysregulation of apoptosis genes in hematopoietic malignancies.** *Oncogene* 2002, **21**(21):3459-3474.
356. Brunet A, Bonni A, Zigmond MJ, Lin MZ, Juo P, Hu LS, Anderson MJ, Arden KC, Blenis J, Greenberg ME: **Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor.** *Cell* 1999, **96**(6):857-868.
357. Cuesta N, Nhu QM, Zudaire E, Polumuri S, Cuttitta F, Vogel SN: **IFN regulatory factor-2 regulates macrophage apoptosis through a STAT1/3- and caspase-1-dependent mechanism.** *J Immunol* 2007, **178**(6):3602-3611.
358. Cuesta N, Salkowski CA, Thomas KE, Vogel SN: **Regulation of lipopolysaccharide sensitivity by IFN regulatory factor-2.** *J Immunol* 2003, **170**(11):5739-5747.

359. Hertzog PJ, O'Neill LA, Hamilton JA: **The interferon in TLR signaling: more than just antiviral.** *Trends Immunol* 2003, **24**(10):534-539.
360. Nhu QM, Cuesta N, Vogel SN: **Transcriptional regulation of lipopolysaccharide (LPS)-induced Toll-like receptor (TLR) expression in murine macrophages: role of interferon regulatory factors 1 (IRF-1) and 2 (IRF-2).** *J Endotoxin Res* 2006, **12**(5):285-295.
361. Tripathi P, Aggarwal A: **NF- $\kappa$ B transcription factor: a key player in the generation of immune response.** *Current Science* 2006, **90**(4):519-531.
362. Ward C, Chilvers ER, Lawson MF, Pryde JG, Fujihara S, Farrow SN, Haslett C, Rossi AG: **NF- $\kappa$ B activation is a critical regulator of human granulocyte apoptosis in vitro.** *J Biol Chem* 1999, **274**(7):4309-4318.
363. Gerritsen ME, Williams AJ, Neish AS, Moore S, Shi Y, Collins T: **CREB-binding protein/p300 are transcriptional coactivators of p65.** *Proc Natl Acad Sci U S A* 1997, **94**(7):2927-2932.
364. Saeki K, Yuo A, Suzuki E, Yazaki Y, Takaku F: **Aberrant expression of cAMP-response-element-binding protein ('CREB') induces apoptosis.** *Biochem J* 1999, **343 Pt 1**:249-255.
365. Alderson MR, Tough TW, Davis-Smith T, Braddy S, Falk B, Schooley KA, Goodwin RG, Smith CA, Ramsdell F, Lynch DH: **Fas ligand mediates activation-induced cell death in human T lymphocytes.** *J Exp Med* 1995, **181**(1):71-77.
366. Chen X, Zachar V, Zdravkovic M, Guo M, Ebbesen P, Liu X: **Role of the Fas/Fas ligand pathway in apoptotic cell death induced by the human T cell lymphotropic virus type I Tax transactivator.** *J Gen Virol* 1997, **78 ( Pt 12)**:3277-3285.
367. Mosteck J, Showalter BM, Rothman PB: **Early growth response-1 regulates lipopolysaccharide-induced suppressor of cytokine signaling-1 transcription.** *J Biol Chem* 2005, **280**(4):2596-2605.
368. Ilangumaran S, Rottapel R: **Regulation of cytokine receptor signaling by SOCS1.** *Immunol Rev* 2003, **192**:196-211.
369. Natarajan M, Lin KM, Hsueh RC, Sternweis PC, Ranganathan R: **A global analysis of cross-talk in a mammalian cellular signalling network.** *Nat Cell Biol* 2006, **8**(6):571-580.
370. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14**(6):1085-1094.
371. Bergmann S, Ihmels J, Barkai N: **Similarities and differences in genome-wide expression data of six organisms.** *PLoS Biol* 2004, **2**(1):E9.
372. van Waveren C, Moraes CT: **Transcriptional co-expression and co-regulation of genes coding for components of the oxidative phosphorylation system.** *BMC Genomics* 2008, **9**:18.
373. Nguyen TT, Androulakis IP: **Recent Advances in the Computational Discovery of Transcription Factor Binding Sites.** *Algorithms* 2009, **2**(1):582-605.
374. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ: **Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.** *Nat Biotechnol* 2005, **23**(3):377-383.

375. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T: **Module networks revisited: computational assessment and prioritization of model predictions.** *Bioinformatics* 2009, **25**(4):490-496.
376. Reverter A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP: **Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data.** *Bioinformatics* 2010, **26**(7):896-904.
377. Sharan R, Ben-Hur A, Loots GG, Ovcharenko I: **CREME: Cis-Regulatory Module Explorer for the human genome.** *Nucleic Acids Res* 2004, **32**(Web Server issue):W253-256.
378. Van Loo P, Aerts S, Thienpont B, De Moor B, Moreau Y, Marynen P: **ModuleMiner - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues?** *Genome Biol* 2008, **9**(4):R66.
379. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A: **Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W541-545.
380. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci U S A* 2004, **101**(33):12114-12119.
381. Kerhornou A, Guigo R: **BioMoby web services to support clustering of co-regulated genes based on similarity of promoter configurations.** *Bioinformatics* 2007, **23**(14):1831-1833.
382. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis.** *Genome Res* 2007, **17**(10):1537-1545.
383. Medzhitov R: **Origin and physiological roles of inflammation.** *Nature* 2008, **454**(7203):428-435.
384. Andreasen AS, Krabbe KS, Krogh-Madsen R, Taudorf S, Pedersen BK, Moller K: **Human endotoxemia as a model of systemic inflammation.** *Curr Med Chem* 2008, **15**(17):1697-1705.
385. Opal SM, Scannon PJ, Vincent JL, White M, Carroll SF, Palardy JE, Parejo NA, Pribble JP, Lemke JH: **Relationship between plasma levels of lipopolysaccharide (LPS) and LPS-binding protein in patients with severe sepsis and septic shock.** *J Infect Dis* 1999, **180**(5):1584-1589.
386. Rankin JA: **Biological mediators of acute inflammation.** *AACN Clin Issues* 2004, **15**(1):3-17.
387. McInnes IB, Schett G: **Cytokines in the pathogenesis of rheumatoid arthritis.** *Nat Rev Immunol* 2007, **7**(6):429-442.
388. Bone RC: **Immunologic dissonance: a continuing evolution in our understanding of the systemic inflammatory response syndrome (SIRS) and the multiple organ dysfunction syndrome (MODS).** *Ann Intern Med* 1996, **125**(8):680-687.
389. Sivalingam SP, Thumboo J, Vasoo S, Thio ST, Tse C, Fong KY: **In vivo pro- and anti-inflammatory cytokines in normal and patients with rheumatoid arthritis.** *Ann Acad Med Singapore* 2007, **36**(2):96-99.



390. Sternberg EM: **Neural regulation of innate immunity: a coordinated nonspecific host response to pathogens.** *Nat Rev Immunol* 2006, **6**(4):318-328.
391. Coogan AN, Wyse CA: **Neuroimmunology of the circadian clock.** *Brain Res* 2008, **1232**:104-112.
392. Levi F, Schibler U: **Circadian rhythms: mechanisms and therapeutic implications.** *Annu Rev Pharmacol Toxicol* 2007, **47**:593-628.
393. Sukumaran S, Almon RR, DuBois DC, Jusko WJ: **Circadian rhythms in gene expression: Relationship to physiology, disease, drug disposition and drug action.** *Adv Drug Deliv Rev* 2010, **62**(9-10):904-917.
394. Cutolo M, Serio B, Cravio C, Pizzorni C, Sulli A: **Circadian rhythms in RA.** *Ann Rheum Dis* 2003, **62**(7):593-596.
395. Lissoni P, Rovelli F, Brivio F, Brivio O, Fumagalli L: **Circadian secretions of IL-2, IL-12, IL-6 and IL-10 in relation to the light/dark rhythm of the pineal hormone melatonin in healthy humans.** *Nat Immun* 1998, **16**(1):1-5.
396. Petrovsky N, McNair P, Harrison LC: **Diurnal rhythms of pro-inflammatory cytokines: regulation by plasma cortisol and therapeutic implications.** *Cytokine* 1998, **10**(4):307-312.
397. An G, Mi Q, Dutta-Moscato J, Vodovotz Y: **Agent-based models in translational systems biology.** *Wiley Interdiscip Rev Syst Biol Med* 2009, **1**(2):159-171.
398. Bauer AL, Beauchemin CA, Perelson AS: **Agent-based modeling of host-pathogen systems: The successes and challenges.** *Inf Sci (Ny)* 2009, **179**(10):1379-1389.
399. Scheff JD, Calvano SE, Lowry SF, Androulakis IP: **Modeling the influence of circadian rhythms on the acute inflammatory response.** *J Theor Biol* 2010, **264**(3):1068-1076.
400. Bahcall OG: **Single cell resolution in regulation of gene expression.** *Mol Syst Biol* 2005, **1**:2005 0015.
401. Blake WJ, M KA, Cantor CR, Collins JJ: **Noise in eukaryotic gene expression.** *Nature* 2003, **422**(6932):633-637.
402. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB: **Gene regulation at the single-cell level.** *Science* 2005, **307**(5717):1962-1965.
403. Chavali AK, Gianchandani EP, Tung KS, Lawrence MB, Peirce SM, Papin JA: **Characterizing emergent properties of immunological systems with multi-cellular rule-based computational modeling.** *Trends Immunol* 2008, **29**(12):589-599.
404. Catron DM, Itano AA, Pape KA, Mueller DL, Jenkins MK: **Visualizing the first 50 hr of the primary immune response to a soluble antigen.** *Immunity* 2004, **21**(3):341-347.
405. An G: **A model of TLR4 signaling and tolerance using a qualitative, particle-event-based method: introduction of spatially configured stochastic reaction chambers (SCSRC).** *Math Biosci* 2009, **217**(1):43-52.
406. An GC, Faeder JR: **Detailed qualitative dynamic knowledge representation using a BioNetGen model of TLR-4 signaling and preconditioning.** *Math Biosci* 2009, **217**(1):53-63.

407. Folcik VA, An GC, Orosz CG: **The Basic Immune Simulator: an agent-based model to study the interactions between innate and adaptive immunity.** *Theor Biol Med Model* 2007, **4**:39.
408. Baldazzi V, Castiglione F, Bernaschi M: **An enhanced agent based model of the immune system response.** *Cell Immunol* 2006, **244**(2):77-79.
409. Celada F, Seiden PE: **A computer model of cellular interactions in the immune system.** *Immunol Today* 1992, **13**(2):56-62.
410. Meier-Schellersheim M, Xu X, Angermann B, Kunkel EJ, Jin T, Germain RN: **Key role of local regulation in chemosensing revealed by a new molecular interaction-based modeling method.** *PLoS Comput Biol* 2006, **2**(7):e82.
411. Warrender C, Forrest S, Koster F: **Modeling intercellular interactions in early Mycobacterium infection.** *Bull Math Biol* 2006, **68**(8):2233-2261.
412. Kaern M, Elston TC, Blake WJ, Collins JJ: **Stochasticity in gene expression: from theories to phenotypes.** *Nat Rev Genet* 2005, **6**(6):451-464.
413. Kilfoil ML, Lasko P, Abouheif E: **Stochastic variation: from single cells to superorganisms.** *Hfsp J* 2009, **3**(6):379-385.
414. Niepel M, Spencer SL, Sorger PK: **Non-genetic cell-to-cell variability and the consequences for pharmacology.** *Curr Opin Chem Biol* 2009, **13**(5-6):556-561.
415. Raser JM, O'Shea EK: **Noise in gene expression: origins, consequences, and control.** *Science* 2005, **309**(5743):2010-2013.
416. Prabhakar U, Conway TM, Murdock P, Mooney JL, Clark S, Hedge P, Bond BC, Jazwinska EC, Barnes MR, Tobin F *et al*: **Correlation of protein and gene expression profiles of inflammatory proteins after endotoxin challenge in human subjects.** *DNA Cell Biol* 2005, **24**(7):410-431.
417. Li Q, Verma IM: **NF-kappaB regulation in the immune system.** *Nat Rev Immunol* 2002, **2**(10):725-734.
418. Vallabhapurapu S, Karin M: **Regulation and function of NF-kappaB transcription factors in the immune system.** *Annu Rev Immunol* 2009, **27**:693-733.
419. Ihekweba AE, Broomhead DS, Grimley RL, Benson N, Kell DB: **Sensitivity analysis of parameters controlling oscillatory signalling in the NF-kappaB pathway: the roles of IKK and IkappaBalpha.** *Syst Biol (Stevenage)* 2004, **1**(1):93-103.
420. Hu X, Chen J, Wang L, Ivashkiv LB: **Crosstalk among Jak-STAT, Toll-like receptor, and ITAM-dependent pathways in macrophage activation.** *J Leukoc Biol* 2007, **82**(2):237-243.
421. Akira S, Takeda K: **Toll-like receptor signalling.** *Nat Rev Immunol* 2004, **4**(7):499-511.
422. O'Neill LA: **When signaling pathways collide: positive and negative regulation of toll-like receptor signal transduction.** *Immunity* 2008, **29**(1):12-20.
423. Croker BA, Kiu H, Nicholson SE: **SOCS regulation of the JAK/STAT signalling pathway.** *Semin Cell Dev Biol* 2008, **19**(4):414-422.
424. Shuai K, Liu B: **Regulation of JAK-STAT signalling in the immune system.** *Nat Rev Immunol* 2003, **3**(11):900-911.

425. Zi Z, Cho KH, Sung MH, Xia X, Zheng J, Sun Z: **In silico identification of the key components and steps in IFN-gamma induced JAK-STAT signaling pathway.** *FEBS Lett* 2005, **579**(5):1101-1108.
426. Webster JL, Tonelli L, Sternberg EM: **Neuroendocrine regulation of immunity.** *Annu Rev Immunol* 2002, **20**:125-163.
427. Carrillo-Vico A, Guerrero JM, Lardone PJ, Reiter RJ: **A review of the multiple actions of melatonin on the immune system.** *Endocrine* 2005, **27**(2):189-200.
428. Guerrero JM, Reiter RJ: **Melatonin-immune system relationships.** *Curr Top Med Chem* 2002, **2**(2):167-179.
429. Hermann C, von Aulock S, Dehus O, Keller M, Okigami H, Gantner F, Wendel A, Hartung T: **Endogenous cortisol determines the circadian rhythm of lipopolysaccharide-- but not lipoteichoic acid--inducible cytokine release.** *Eur J Immunol* 2006, **36**(2):371-379.
430. Skwarlo-Sonta K, Majewski P, Markowska M, Oblap R, Olszanska B: **Bidirectional communication between the pineal gland and the immune system.** *Can J Physiol Pharmacol* 2003, **81**(4):342-349.
431. Hoffmann A, Levchenko A, Scott ML, Baltimore D: **The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation.** *Science* 2002, **298**(5596):1241-1245.
432. O'Dea EL, Barken D, Peralta RQ, Tran KT, Werner SL, Kearns JD, Levchenko A, Hoffmann A: **A homeostatic model of IkappaB metabolism to control constitutive NF-kappaB activity.** *Mol Syst Biol* 2007, **3**:111.
433. Tripathi P, Aggarwal A: **NF-kB transcription factor: a key player in the generation of immune response.** *Current Science* 2006, **90**:519-531.
434. Kearns JD, Basak S, Werner SL, Huang CS, Hoffmann A: **IkappaBepsilon provides negative feedback to control NF-kappaB oscillations, signaling dynamics, and inflammatory gene expression.** *J Cell Biol* 2006, **173**(5):659-664.
435. Karin M, Delhase M: **The I kappa B kinase (IKK) and NF-kappa B: key elements of proinflammatory signalling.** *Semin Immunol* 2000, **12**(1):85-98.
436. Calandra T, Roger T: **Macrophage migration inhibitory factor: a regulator of innate immunity.** *Nat Rev Immunol* 2003, **3**(10):791-800.
437. Roger T, David J, Glauser MP, Calandra T: **MIF regulates innate immune responses through modulation of Toll-like receptor 4.** *Nature* 2001, **414**(6866):920-924.
438. Cavadini G, Petrzilka S, Kohler P, Jud C, Tobler I, Birchler T, Fontana A: **TNF-alpha suppresses the expression of clock genes by interfering with E-box-mediated transcription.** *Proc Natl Acad Sci U S A* 2007, **104**(31):12843-12848.
439. Fernandes PA, Cecon E, Markus RP, Ferreira ZS: **Effect of TNF-alpha on the melatonin synthetic pathway in the rat pineal gland: basis for a 'feedback' of the immune response on circadian timing.** *J Pineal Res* 2006, **41**(4):344-350.
440. Pontes GN, Cardoso EC, Carneiro-Sampaio MM, Markus RP: **Pineal melatonin and the innate immune response: the TNF-alpha increase after cesarean section suppresses nocturnal melatonin production.** *J Pineal Res* 2007, **43**(4):365-371.

441. Smoak KA, Cidlowski JA: **Mechanisms of glucocorticoid receptor signaling during inflammation.** *Mech Ageing Dev* 2004, **125**(10-11):697-706.
442. Fernandes PA, Bothorel B, Clesse D, Monteiro AW, Calgari C, Raison S, Simonneaux V, Markus RP: **Local corticosterone infusion enhances nocturnal pineal melatonin production in vivo.** *J Neuroendocrinol* 2009, **21**(2):90-97.
443. Ferreira ZS, Fernandes PA, Duma D, Assreuy J, Avellar MC, Markus RP: **Corticosterone modulates noradrenaline-induced melatonin synthesis through inhibition of nuclear factor kappa B.** *J Pineal Res* 2005, **38**(3):182-188.
444. Pando MP, Verma IM: **Signal-dependent and -independent degradation of free and NF-kappa B-bound IkappaBalpha.** *J Biol Chem* 2000, **275**(28):21278-21286.
445. Veldhuis JD, Iranmanesh A, Lizarralde G, Johnson ML: **Amplitude modulation of a burstlike mode of cortisol secretion subserves the circadian glucocorticoid rhythm.** *Am J Physiol* 1989, **257**(1 Pt 1):E6-14.
446. Waage A, Brandtzaeg P, Halstensen A, Kierulf P, Espevik T: **The complex pattern of cytokines in serum from patients with meningococcal septic shock. Association between interleukin 6, interleukin 1, and fatal outcome.** *J Exp Med* 1989, **169**(1):333-338.
447. Brown GC: **Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells.** *J Theor Biol* 1991, **153**(2):195-203.
448. Rothman S: **How is the balance between protein synthesis and degradation achieved?** *Theor Biol Med Model* 2010, **7**:25.
449. Hess A, Iyera H, Malmb W: **Linear trend analysis: a comparison of methods.** *Atmospheric Environment* 2001, **35**(30):5211-5222.
450. Maha RSH, Tamhanea AC, Tunga SH, Patela AN: **Process trending with piecewise linear smoothing.** *Computers & Chemical Engineering* 1995, **19**(2):129-137.
451. Bellet MM, Sassone-Corsi P: **Mammalian circadian clock and metabolism - the epigenetic link.** *J Cell Sci* 2010, **123**(Pt 22):3837-3848.
452. Nelson DE, Ihekweba AE, Elliott M, Johnson JR, Gibney CA, Foreman BE, Nelson G, See V, Horton CA, Spiller DG *et al*: **Oscillations in NF-kappaB signaling control the dynamics of gene expression.** *Science* 2004, **306**(5696):704-708.
453. Lipniacki T, Paszek P, Brasier AR, Luxon BA, Kimmel M: **Stochastic regulation in early immune response.** *Biophys J* 2006, **90**(3):725-742.
454. Schooley K, Zhu P, Dower SK, Qwarnstrom EE: **Regulation of nuclear translocation of nuclear factor-kappaB relA: evidence for complex dynamics at the single-cell level.** *Biochem J* 2003, **369**(Pt 2):331-339.
455. Gori AM, Cesari F, Marcucci R, Giusti B, Paniglia R, Antonucci E, Gensini GF, Abbate R: **The balance between pro- and anti-inflammatory cytokines is associated with platelet aggregability in acute coronary syndrome patients.** *Atherosclerosis* 2009, **202**(1):255-262.
456. Jerin A, Pozar-Lukanovic N, Sojar V, Stanisavljevic D, Paver-Erzen V, Osredkar J: **Balance of pro- and anti-inflammatory cytokines in liver surgery.** *Clin Chem Lab Med* 2003, **41**(7):899-903.

- 457. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N: **Noise in protein expression scales with natural protein abundance.** *Nat Genet* 2006, **38**(6):636-643.
- 458. Rausenberger J, Kollmann M: **Quantifying origins of cell-to-cell variations in gene expression.** *Biophys J* 2008, **95**(10):4523-4528.
- 459. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Liron Y, Rosenfeld N, Danon T, Perzov N, Alon U: **Variability and memory of protein levels in human cells.** *Nature* 2006, **444**(7119):643-646.
- 460. Hermida RC, Ayala DE, Calvo C: **Optimal timing for antihypertensive dosing: focus on valsartan.** *Ther Clin Risk Manag* 2007, **3**(1):119-131.

# Curriculum Vitae

Tung T. Nguyen

---

## Education

02/2012 –	Postdoctoral Research Fellow, <i>University of California – San Diego</i>
09/2006 – 1/2012	Ph.D., BioMaPS Institute for Quantitative Biology, <i>Rutgers - The State University of New Jersey</i>
09/2001 – 05/2004	M.S., Computer Science Department, <i>University of Natural Sciences, VNU-HCMC</i>
09/1996 – 09/2000	B.S., Computer Science Department, <i>University of Natural Sciences, VNU-HCMC</i>

## Publications

1. **Nguyen, T.T.**, Foteinou, P.T., Calvano, S.E., Lowry, S.F., and Androulakis, I.P. Computational identification of transcriptional programs in human endotoxemia. *PLoS One*, 6(5):e18889 (2011)
2. **Nguyen, T.T.**, Calvano, S.E., Lowry, S.F., Androulakis, I.P. Agent based of human endotoxemia accounting for circadian variability. *J. Critical Care*, 26(2):e6-e7 (2011)
3. Swiss, V.A., **Nguyen, T.T.**, Dugas, J.C., Ibrahim, A., Barres, B.A., Androulakis, I.P., and Casaccia, P. Identification of a gene regulatory network necessary for the initiation of oligodendrocyte differentiation. *PLoS One*, 6(4):e18088 (2011)

4. Orman, M.A., **Nguyen, T.T.**, Ierapetritou, M.G., Berthiaume, F., and Androulakis, I.P. Comparison of Cytokine Dynamics of the Early Inflammatory Response in Models of Burn Injury and Infection. *Cytokines*, 55(3):362-71 (2011)
5. Yang, Q., Mattick, J.S.A., Orman, M.A., **Nguyen, T.T.**, Ierapetritou, M.G., Berthiaume, F., Androulakis, I.P. Dynamics of hepatic gene expression profile in a rat cecal ligation and puncture model. *Journal of Surgical Research (in press)*
6. **Nguyen, T.T.**, Almon, R.R., DuBois, D.C., Jusko, W.J., and Androulakis, I.P. Comparative analysis of acute and chronic corticosteroid pharmacogenomic effects in rat liver: Transcriptional dynamics and regulatory structures. *BMC Bioinformatics*, 11:515 (2010).
7. **Nguyen, T.T.**, R.R. Almon, D.C. DuBois, W.J. Jusko and I.P. Androulakis, Importance of replication in analyzing time-series gene expression data: Corticosteroid dynamics and circadian patterns in rat liver. *BMC Bioinformatics*, 11:279 (2010)
8. **Nguyen T.T.**, Foteinou P.T., Calvano, S.E., Lowry, S.F., and Androulakis, I.P. Dynamic complexity of the temporal transcriptional regulation program in human endotoxemia. *IEEE Intl Conf on Bioinfo. and Bioeng. BIBE*, pp. 112-117 (2010)
9. **Nguyen, T.T.**, R. Nowakoski, and I.P. Androulakis, Unsupervised selection of highly coexpressed and non-coexpressed genes. *OMICS*, 13(3): 219-237 (2009)
10. **Nguyen T.T.** and Androulakis I. P. Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*. 2(1):582-605 (2009)