SYSTEMS ANALYSIS OF COMPLEX BIOLOGICAL DATA FOR BIOPROCESS ENHANCEMENT

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Salim Pyarali Charaniya

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Wei-Shou Hu

December, 2008

ACKNOWLEDGEMENTS

This work would not have been possible without the support and encouragement of many people whom I wish to thank. My advisor, Prof. Wei-Shou Hu, has been a constant source of guidance and encouragement throughout the course of this work. His enthusiasm and motivation have inspired me through the ups and downs of my graduate years. With deep gratitude, I thank him for this wonderful experience. I am also immensely grateful to Prof. George Karypis for his guidance and support (and the baklava!), especially during the early learning years. I would like to thank my thesis committee members Prof. Friedrich Srienc and Prof. Yiannis Kaznessis for agreeing to serve on my committee and taking the time to review this thesis.

I wish to thank the past and present members of the 'Hu group' who have made this a joyful and memorable experience. I thank Wei Lian, Katie Wlaschin, Mugdha Gadgil, Gargi Seth, Patrick Hossler, Sarika Mehra, David Umulis, Joon Chong Yee, Karthik Jayapal, CM Cameron, Anne Kantardjieff, Marlene Castro, Nitya Jacob, Siguang Sui, Bhanu Chandra Mulukutla, Kartik Subramanian, Yonsil Park, Jason Owens, Huong Le, Anushree Chatterjee, and Kathryn Johnson. I also wish to thank the undergraduate student Lewis Marshall who worked with me. I offer thanks to members of Karypis lab – Nikil Wale, Huzefa Rangwala, and Chris Kauffman, who have, over the years, patiently endured my naiveté in machine learning.

I would also like to acknowledge Keri Mills and Dr. Kevin Johnson from Genentech for the wonderful opportunity of collaboration.

The efforts of Zheng Jin Tu at the Minnesota Supercomputing Institute and Archana Deshpande at the Biomedical Genomics Center are greatly appreciated.

Many thanks to my friends in Minnesota and my cousin Nadya who convinced me that Minnesota is not that cold!

For the affectionate memories of yesterday and the beautiful dreams of tomorrow, I thank Shikha for her love and companionship. Thank you for being there for me. Lastly, I would like to acknowledge the unwavering love and support of my parents, Dilshad and Pyarali. Your 'ordinariness' is far from ordinary. It is the source of my inspiration and all my endeavors.

DEDICATION

With love to my mother, Dilshad

ABSTRACT

Recent advances in data-driven knowledge discovery approaches, such as 'omics' technologies, provide enormous opportunities to uncover the multifarious determinants of several pharmaceutically relevant biological traits. This work focuses on the challenges, which include: (i) Deciphering the regulation of antibiotic production in *Streptomyces coelicolor*, and (ii) Elucidating the attributes of high recombinant protein productivity in mammalian cell culture processes.

The phenotypic complexity of Streptomycetes, which produce several clinically relevant antibiotics and other natural products, manifests in their diversity of secondary metabolism and morphological differentiation. To identify the dynamic gene regulatory networks that confer such complex phenotypes, the temporal transcriptomic characteristics of the model organism *S. coelicolor,* under more than twenty-five diverse genetic and environmental perturbations, were integrated with other functional and genomic features. A whole-genome operon map was also predicted, and a significant portion of the map was experimentally verified. Such a systems approach can reveal several insights about the functional processes relevant for antibiotics production.

The therapeutic value of recombinant proteins has brought about a continuously rising demand that is met by development of hyper-producing mammalian cell lines. However, the molecular ingredients of high productivity are not well understood. The transcriptomes of several recombinant antibody-producing NS0 cell lines with a wide productivity range were surveyed in an attempt to identify the physiological functions that are modulated in high-producing cells. Cell culture process enhancement also entails an understanding of the process parameters and their interactions, which are critical determinants of high recombinant protein productivity. The comprehensive process archives of modern production plants present vast, underutilized resources containing information that, if unearthed, can enhance process robustness. The on-line and off-line process data of several production 'trains' from a commercial manufacturing facility were investigated using kernel-based machine learning tools to elucidate predictive correlations between process parameters and the outcome.

Together, such discovery strategies based on integrative data mining hold immense potential for enhancing our understanding of industrially relevant biological processes.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1

1.1 THE PROMISE OF BIOTECHNOLOGY 1
1.1.1 The Dawn of Omics in Gene Expression 1
1.1.2 Other Large-Scale Biological Datasets 4
1.2 SCOPE OF THESIS 5
1.3 THESIS ORGANIZATION 6

CHAPTER 2 BACKGROUND 7

2.1 STREPTOMYCETES 7
2.1.1 Pharmacological Relevance 7
2.1.2 Life Cycle of Streptomycetes 7
2.1.3 Complete Genome Sequence of Model Organism – *Streptomyces coelicolor* 8
2.1.4 Antibiotics Produced by *S. coelicolor* 9
2.1.5 Regulation of Antibiotics Synthesis 10
2.1.6 Global Gene Expression Profiling to Investigate Secondary Metabolism 14
2.1.7 Antibiotic Regulation – A Synopsis 15
2.2 MAMMALIAN CELL CULTURE 17
2.2.2 The Process of Cell Line Development[93] 18
2.2.3 Transcriptome and Proteome Surveys to Understand Cell Physiology 19
2.2.4 Data Analysis for High Productivity Complex Traits 20
2.2.5 The Complex Trait of Hyperproductivity 22
2.2.6 Mammalian Cell Culture – A Synopsis 24

CHAPTER 3 MINING BIOPROCESS DATA: OPPORTUNITIES AND CHALLENGES 25

3.1 SUMMARY 25
3.2 INTRODUCTION 25
3.3 CHARACTERISTICS OF BIOPROCESS DATA 26
3.4 KNOWLEDGE DISCOVERY AND BIOPROCESSES 29
3.4.1 Data Preprocessing 31
3.4.2 Feature Selection – Dimensionality Reduction 32
3.4.3 Data Mining 33
3.4.4 Model Validation and Interpretation 40
3.5 CONCLUDING REMARKS 41

CHAPTER 4 TRANSCRIPTOME DYNAMICS-BASED OPERON PREDICTION AND VERIFICATION IN *STREPTOMYCES COELICOLOR* 43

4.1 SUMMARY 43
4.2 INTRODUCTION 43
4.3 MATERIALS AND METHODS 45
4.3.1 Microarray Data 45
4.3.2 Genome Organization 46
4.3.3 Prediction of transcription terminators 48
4.3.4 Experimental Verification of Operons 48

4.3.5 Supervised Classification 49
4.4 RESULTS 52
4.4.1 Known Operon Pairs have Shorter Intergenic Distance 52
4.4.2 Genes in Known Operons have Greater Expression Correlation 53
4.4.3 Transcription Terminators 54
4.4.4 Binary Classification Results 54
4.4.5 Identification of Transcription Units 60
4.4.6 Operons with Internal Regulation 61
4.4.7 Operon Predictions for Entire Genome 62
4.4.8 Experimental Verification 64
4.5 DISCUSSION 68
4.5.1 Dependence of Transcript Dynamics for Operon Prediction 68
4.5.2 Other Features 69
4.5.3 Prediction of Transcription Units 70
4.5.4 Using Operon Predictions for Functional Annotations 71
4.5.5 Comparison of Operon Predictions with Earlier Report 71
4.6 CONCLUDING REMARKS 73

CHAPTER 5 FURTHER REFINEMENT OF OPERON PREDICTIONS AND DISCOVERY OF REGULATORY HUBS FOR STREPTOMYCES SECONDARY METABOLISM 75

5.1 SUMMARY 75
5.2 INTRODUCTION 75
5.3 METHODS 77
5.3.1 Microarray Data 77
5.3.2 Functional Similarity 80
5.3.3 Conservation of Gene Order 81
5.3.4 Supervised Classification 81
5.3.5 Functional Network Analysis 81
5.4 RESULTS 82
5.4.1 Features used for Genome-wide Operon Prediction 83
5.4.2 Binary Classification Results 86
5.4.3 A Whole-Genome Operon Map 88
5.4.4 Identifying Transcriptional Interactions in *S. coelicolor* 89
5.5 DISCUSSION 98
5.5.1 Refining the Whole-genome Operon Map 98
5.5.2 Reverse Engineering Transcriptional Network of *S. coelicolor* 100
5.6 CONCLUDING REMARKS 102

CHAPTER 6 MINING TRANSCRIPTOME DATA FOR FUNCTION-TRAIT RELATIONSHIP OF HYPER PRODUCTIVITY OF RECOMBINANT ANTIBODY 104

6.1 SUMMARY 104
6.2 INTRODUCTION 104
6.3 MATERIALS AND METHODS 106
6.3.1 Cells and Sample Preparation 106
6.3.2 Microarray Hybridization 106

6.3.3 Microarray Data Processing 107
6.3.4 Differential Expression Analysis 107
6.3.5 Gene Selection and Support vector machines (SVM) Classification 107
6.3.6 Functional Analysis 108
6.4 RESULTS 108
6.4.1 Classification of Producers with Different Productivity 109
6.4.2 Functional Analysis 110
6.4.3 Genes enriched in High and Low Producer Classes 112
6.5 DISCUSSION 119
6.5.1 Identification of Molecular Signature for Productivity Trait 119
6.5.2 Significance Testing for Functional Analysis 120
6.6 CONCLUDING REMARKS 122

CHAPTER 7 MINING CELL CULTURE PROCESS DATA TO UNVEIL HIGH PRODUCTIVITY CHARACTERISTICS 135

7.1 SUMMARY 135
7.2 INTRODUCTION 135
7.3 METHODS 136
7.3.1 Data Preprocessing 136
7.3.2 Estimation of Similarity between Parameter Profiles 140
7.3.3 Estimation of Parameter Weight 141
7.3.4 Supervised Machine Learning 141
7.4 RESULTS 143
7.4.1 Selection and Preprocessing of Bioprocess Data 144
7.4.2 Kernel Transformation and Comparison of Process Runs 144
7.4.3 Productivity-based Approach for Parameter Weighting 145
7.4.4 Integration of all Process Parameters 146
7.4.5 Predictive Data Mining using Support Vector Regression 147
7.4.6 Stage-specific Identification of Critical Process Parameters 150
7.5 DISCUSSION 154
7.5.1 An Adaptable Framework for Mining Process Cell Culture Data 154
7.5.2 An Efficient Weighting Strategy for Integrating Heterogeneous Process Parameters 154
7.5.3 Influence of Process Parameters on Outcome 155
7.6 CONCLUDING REMARKS 157

CHAPTER 8 SUMMARY AND CONCLUDING REMARKS 158

REFERENCES 161

LIST OF TABLES

TABLE 4.1. SUMMARY OF MICROARRAY DATA USED FOR OPERON PREDICTIONS 47
TABLE 4.2. COMPARISON OF DIFFERENT CLASSIFIERS USING LEAVE-ONE-OUT CROSS-VALIDATION 57
TABLE 4.3. COMPARISON OF DIFFERENT CLASSIFIERS BY 5-FOLD CROSS-VALIDATION… 59
TABLE 4.4. DISTRIBUTION OF SCORES OF SAME-STRAND GENE PAIRS WITH UNKNOWN OPERON STATUS 63
TABLE 4.5. FUNCTIONAL ANALYSIS OF SAME-STRAND GENE PAIRS 64
TABLE 4.6. RT-PCR BASED VERIFICATION OF CO-TRANSCRIPTION OF GENE PAIRS 65
TABLE 4.7. EXTENSION OF CISTRON BOUNDARY OF KNOWN OPERONS 67
TABLE 4.8. SIZE DISTRIBUTION OF THE PREDICTED TRANSCRIPTION UNITS 70
TABLE 5.1. SUMMARY OF MICROARRAY DATA COMPILED FOR ANALYSIS 79
TABLE 5.2. COMPARISON OF THE AUC OF DIFFERENT CLASSIFIERS… 87
TABLE 5.3. MOST INTER-CONNECTED HUBS IN THE PREDICTED *S. COELICOLOR* TRANSCRIPTIONAL NETWORK 92
TABLE 6.1. SCORES FOR DIFFERENT PRODUCERS BASED ON SVM CLASSIFICATION 109
TABLE 6.2. FUNCTIONAL GENE SETS IDENTIFIED BY DIFFERENT GENE SET TESTING METHODS .. 113
TABLE 6.3. LIST OF DIFFERENTIALLY EXPRESSED GENES IN THE FUNCTIONAL CLASS 'GOLGI APPARATUS' 124
TABLE 6.4. LIST OF DIFFERENTIALLY EXPRESSED GENES INVOLVED IN CYTOSKELETON FUNCTION 127
TABLE 6.5. LIST OF DIFFERENTIALLY EXPRESSED GENES IN THE GENE SET 'CHROMATIN' 129
TABLE 6.6. LIST OF DIFFERENTIALLY EXPRESSED GENES INVOLVED IN CELL CYCLE PROGRESSION 130
TABLE 6.7. LIST OF DIFFERENTIALLY EXPRESSED GENE IN THE FUNCTIONAL CLASS 'STRUCTURAL CONSTITUENT OF RIBOSOME' 132
TABLE 6.8. LIST OF DIFFERENTIALLY EXPRESSED GENES INVOLVE IN THE FUNCTIONAL CLASS 'LIGASE ACTIVITY' 133
TABLE 6.9. DIFFERENTIALLY EXPRESSED GENES INVOLVED IN EARLY SECRETION PATHWAY AT NODES 4 AND 5 134
TABLE 7.1. SUMMARY OF PROCESS PARAMETERS AT DIFFERENT BIOREACTOR SCALES 139

LIST OF FIGURES

FIGURE 1.1. PROCESS WORKFLOW FOR A TWO-CHANNEL DNA MICROARRAY 3
FIGURE 2.1. CHEMICAL STRUCTURES OF THE FOUR KNOWN ANTIBIOTICS IN *S. COELICOLOR* 10
FIGURE 2.2. KNOWN REGULATORY MAP FOR *S. COELICOLOR* 16
FIGURE 3.1. EXAMPLE OF BIOPROCESS DATA…. 27
FIGURE 3.2. AN APPROACH FOR DATA-DRIVEN KNOWLEDGE DISCOVERY IN BIOPROCESS DATABASES…. 31
FIGURE 3.3. AN APPROACH TO DETERMINE THE SIMILARITY BETWEEN DIFFERENT PROCESS RUNS 36
FIGURE 3.4. A KERNEL-BASED LEARNING APPROACH…. 37
FIGURE 3.5. MAXIMUM MARGIN SUPPORT VECTOR CLASSIFICATION…. 40
FIGURE 4.1. DEFINITION OF KNOWN OPERON PAIRS (KOPS), NON-OPERON PAIRS (NOPS), SAME-STRAND PAIRS, AND OPPOSITE-STRAND PAIRS… 46
FIGURE 4.2: DENSITY DISTRIBUTION OF INTERGENIC DISTANCE IN KOPS AND NOPS. 53
FIGURE 4.3. COMPARISON OF PEARSON CORRELATION BETWEEN TRANSCRIPT LEVELS OF ADJACENT GENES IN KOPS AND NOPS.. 55
FIGURE 4.4. COMPARISON OF DIFFERENT CLASSIFIERS BY ROC CURVE. 59
FIGURE 4.5. EXPERIMENTAL VERIFICATION OF CO-TRANSCRIPTION OF ADJACENT GENES BY RT-PCR.. 65
FIGURE 4.6. COMPARISON OF THE PERFORMANCE OF SVM CLASSIFIER WITH THE PREDICTIONS OF PRICE *ET AL.* 73
FIGURE 5.1. COMPARISON OF PEARSON CORRELATION BETWEEN TRANSCRIPT LEVEL OF ADJACENT GENES IN KOPS AND NOPS. 83
FIGURE 5.2. COMPARISON OF FUNCTIONAL SIMILARITY BETWEEN ADJACENT GENES IN KOPS AND NOPS. 84
FIGURE 5.3. COMPARISON OF DIFFERENT SVM CLASSIFIERS BY 10-FOLD CROSS-VALIDATION AND ROC GRAPHS. 87
FIGURE 5.4. DISTRIBUTION OF CISTRON SIZES 88
FIGURE 5.5. PREDICTED TRANSCRIPTIONAL REGULATORY NETWORK FOR *S. COELICOLOR*. 91
FIGURE 5.6. GLOBAL CONNECTIVITY PROPERTIES OF PREDICTED *S. COELICOLOR* TRANSCRIPTIONAL NETWORK. 92
FIGURE 5.7. NETWORK MODULES RELATED TO SECONDARY METABOLITE SYNTHESIS. 95
FIGURE 5.8. IDENTIFYING PUTATIVE TRANSCRIPTIONAL INTERACTIONS. 96
FIGURE 5.9. OTHER BIOLOGICALLY RELEVANT MODULES. 97
FIGURE 6.1. GENES DIFFERENTIALLY EXPRESSED BETWEEN HIGH AND LOW-PRODUCING NS0 CELL LINES GROUPED ACCORDING TO INTRACELLULAR FUNCTION. 123
FIGURE 7.1. HISTOGRAM OF THE NORMALIZED PRE-HARVEST TITER OF THIRTY PRODUCTION RUNS 137
FIGURE 7.2. PREPROCESSING OF CELL CULTURE PROCESS DATA. 138
FIGURE 7.3. FLOW DIAGRAM OF THE PROPOSED METHODOLOGY FOR KNOWLEDGE DISCOVERY IN MANUFACTURING CELL CULTURE PROCESS DATASETS 143
FIGURE 7.4. A DIFFERENTIAL WEIGHTING SCHEME FOR PROCESS PARAMETERS. 146
FIGURE 7.5. EVALUATION OF SVR MODEL PERFORMANCE. 149

FIGURE 7.6. RELATIVE IMPORTANCE OF PROCESS PARAMETERS AT DIFFERENT STAGES OF THE PRODUCTION PHASE (12000L SCALE)..... 151
FIGURE 7.7. SELECTED CRITICAL PROCESS PARAMETERS AT DIFFERENT STAGES OF THE PRODUCTION PHASE..... 152
FIGURE 7.8. RELATIVE IMPORTANCE OF PROCESS PARAMETERS ACQUIRED AT 400L AND 2000L SCALE INOCULUM BIOREACTORS..... 153
FIGURE 7.9. CRITICAL PROCESS PARAMETERS MEASURED AT 400L SCALE. 153

PERMISSION TO REPRODUCE PUBLISHED MATERIAL

Authorization was granted by the publishers to reproduce in this dissertation, content contained within the following materials published during the course of my graduate studies.

1. Charaniya, S., Karypis, G., Hu, W.S. (2008). Mining transcriptome data for function-trait relationship of hyper productivity of recombinant antibody. *Biotech Bioeng* (accepted)

2. Charaniya, S., Hu, W.S., Karypis, G., (2008). Mining bioprocess data: opportunities and challenges. *Trends Biotechnol* **26** (12), 690-699.

3. Seth, G., Charaniya, S., Wlaschin, K., Hu, W.S. (2007). In pursuit of a super producer – alternative paths to high producing recombinant mammalian cells. *Curr Opin Biotechnol* **18**, 1-8

4. Charaniya, S., Mehra, S., Lian, W., Jayapal, K.P., Karypis, G., Hu, W.S. (2007). Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucl Acids Res* **35** (21), 7222-7236.

CHAPTER 1 INTRODUCTION

1.1 THE PROMISE OF BIOTECHNOLOGY

The discovery of penicillin by Alexander Fleming in 1928 and streptomycin by Selman Waksman in 1943 revolutionalized modern medicine in post-World War II twentieth century. The bioprocess advances in penicillin production provide an archetypical model for the rise and impact of biotechnology. The worldwide capacity for penicillin has increased phenomenally over the last sixty years to more than 60,000 tonnes. Much of this improvement is attributed to isolation and engineering of *Penicillium chrysogenum* strains that produce 10^5-fold higher penicillin titers compared to the first strain isolated 80 years ago[1, 2]. These colossal improvements reinforce the importance of deciphering and engineering the cellular phenotypes that induce very high titers of these drugs.

Tens of millions of people across the world have benefitted from numerous antibacterial drugs and modern protein-based therapeutics for treatment of life-threatening diseases such as cancer. The introduction of genome-scale omics technologies offers a powerful tool to comprehend and enhance these bioprocesses to provide cost-effective and affordable healthcare in decades to come.

1.1.1 THE DAWN OF OMICS IN GENE EXPRESSION

The development of technologies for dideoxy-based DNA sequencing by Frederick Sanger[3] in 1970s and polymerase chain reaction (PCR) by Kary Mullis[4] in 1980s inscribe, arguably, the most fundamental events that ushered a new era of high-throughput biology. Since the first report on the whole genome sequence of a free-living microorganism *Haemophilus Influenzae*[5], 773 prokaryotic and 23 eukaryotic genomes, including the human genome, have been completely sequenced[§]. The availability of the genetic code presents unprecedented opportunities to fathom and unlock the mysteries of complex biological phenomena.

---

§Source: National Center for Biotechnological Information (NCBI)

The genomes of living organisms constitute thousands of genes whose products (proteins) modulate the most vital cellular functions such as survival and proliferation. The outcome of many cellular functions is governed by the synergy between numerous genes and proteins. It is the co-ordination and the interdependence between these genes that determines the abundance and temporal dynamics of RNA and protein moieties, which profoundly affects physiological functions. Although the gel-based hybridization technique (northern blot) to quantitatively estimate the transcript level of a gene was introduced in late 1970s, it was largely restricted to investigations of handful of genes. In a seminal report, Schena *et al.*[6] proposed the 'DNA microarray' technology to simultaneously interrogate the expression level of thousands of genes (called transcriptomics) allowing a molecular-level snapshot of the intricate biological machinery in an organism. Since then, microarrays have been widely used to analyze and correlate gene expression patterns to delineate gene functions and their roles in biological pathways. So ubiquitous has been the use of the microarrays that the number of publications using DNA microarrays has increased exponentially from less than 100 reports in 1998 to more than 30,000 by the end of 2007.

Microarrays utilize the complementarity of nucleic acid sequences. A DNA microarray comprises thousands of 'probes', each of which is specific and complementary to the messenger RNA (mRNA) for a particular gene. Each gene probe, usually a PCR-amplified fragment of the gene (or its complementary DNA) or a short chemically-synthesized oligonucleotide, is deposited on the microarray surface by contact-spotting, ink-jet printing, or *in situ* synthesis by photolithography. The process workflow for the popularly used contact-spotted 'two-channel' microarray platform is shown in Figure 1.1. The mRNA is extracted from cells under two conditions (treatment and control). The mRNA samples are reverse-transcribed and labeled with different fluorescent dyes. Thereafter, the cDNA samples are hybridized on the microarray slide where the fluorescently-labeled cDNA species (referred to as 'target') competitively bind to the corresponding gene probes. After hybridization, the microarray is scanned with lasers to detect fluorescence intensities for target-probe interaction. The ratio of the two intensities, after normalization, represents the relative expression level of a gene in the two samples. Alternative 'single-channel' microarray platforms for absolute gene expression quantitation are also popularly used (e.g., Affymetrix, NimbleGen, and Agilent).

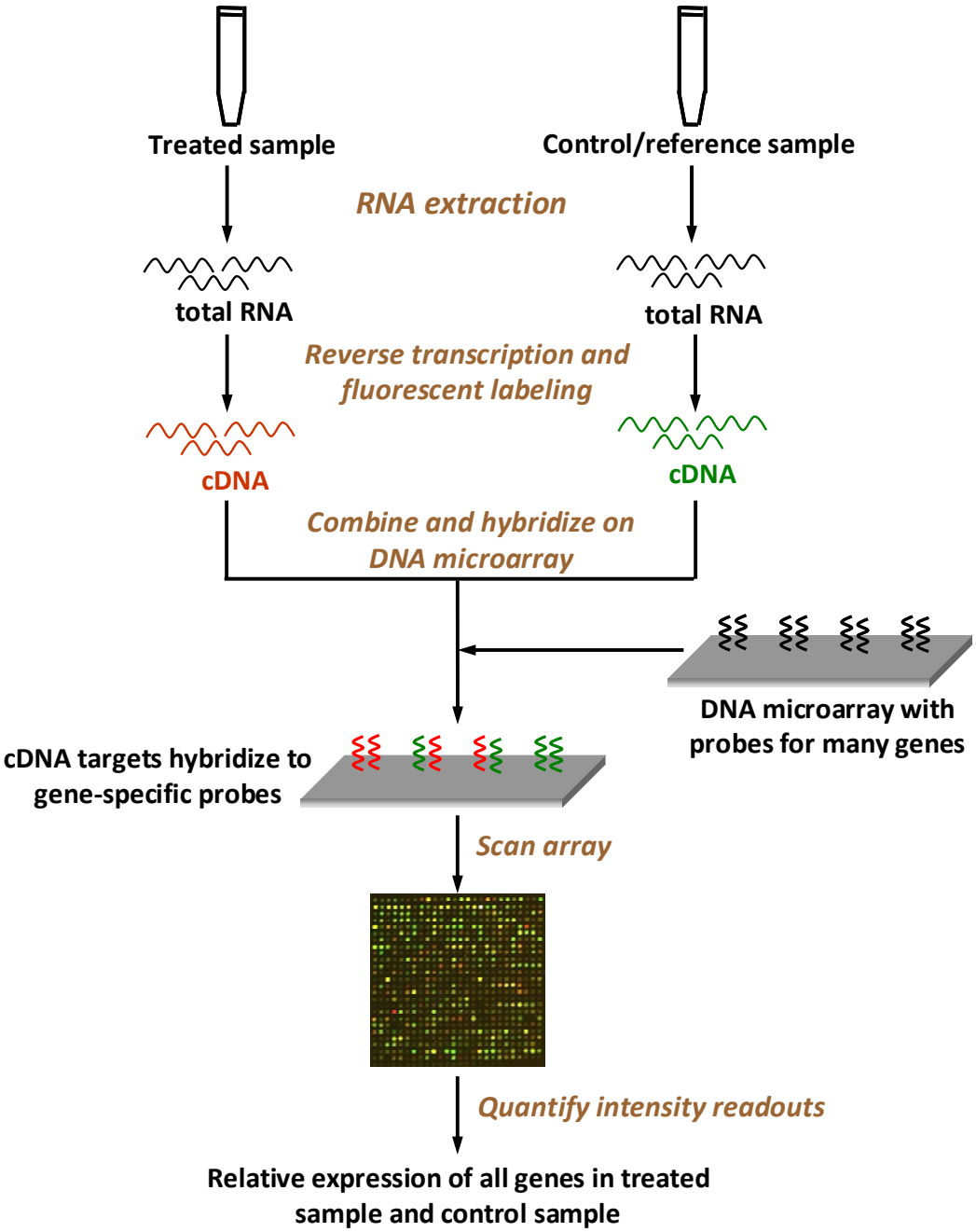

Figure 1.1. Process workflow for a two-channel DNA microarray

Microarrays provide a valuable tool in the quest to decipher the associations between diseases (such as cancer), the underlying genes' expression changes, and the drugs that modulate the activities of those gene products[7, 8]. Despite their enormous value for clinical diagnosis, concerns have been raised about their cross-platform reproducibility and reliability[9]. A comparison of three different commercial microarray platforms showed little commonality between the lists of differentially expressed genes identified by each platform[10]. These concerns prompted a MicroArray Quality Control (MAQC) project, led by US Food and Drug Administration (FDA) and involving more than fifty organizations, to compare and assess the performances of seven different microarray platforms. The results of MAQC provide compelling evidence to suggest that, with careful design and evaluation, microarray studies produce substantially similar results across different laboratories and platforms[11].

Analyzing the vast reams of microarray-based gene expression data to decipher cellular regulation is a daunting endeavor. A wide gamut of computational and bioinformatics methods have been proposed for data normalization, statistical analysis for differential expression, and pattern recognition and data mining. Despite the abundance of data analysis methods, several points of consensus have emerged that provide a set of prudent guidelines for microarray analysis[12, 13].

1.1.2 OTHER LARGE-SCALE BIOLOGICAL DATASETS

Microarray-based transcriptome technology is complemented by other genome-scale omics technologies including proteomics, metabolomics, and interactomics that seek to understand cellular functions in the context of proteins, metabolites, and the interactions between these molecular species[14]. Integrating microarray data with these functional genomic data sources, including associations discovered by mining the exhaustive biomedical literature, can augment the potential for discovering the functions and interactions of the genes associated with complex phenotypes[15, 16]. The historical process archives of modern production plants constitute another vast, underused resource for enhancing the robustness and efficiency of bioprocesses.

As we stride forward, the union of complex biological datasets derived from omics and other sources with computational data mining-based knowledge discovery approaches will synergize our pursuit of deciphering the attributes of superior bioprocesses.

1.2 SCOPE OF THESIS

The focus of this study is to develop and apply statistical and data mining tools in an effort to deconvolute the labyrinthine biological phenotypes associated with pharmaceutically relevant products. Specifically, the focus is on: (i) the genetic circuits that regulate and trigger antibiotic synthesis in *Streptomyces* spp., in particular, the model organism *Streptomyces coelicolor*, (ii) cellular and process characteristics of high recombinant protein productivity in mammalian cells.

Streptomyces spp. produce a wide spectrum of natural products of immense biotechnological importance. This study focuses on the use of large-scale temporal transcriptome data obtained using whole-genome DNA microarrays for the model species *S. coelicolor*, to discern the regulatory aspects of the synthesis of these 'secondary metabolites', which are believed to be largely non-essential for vegetative cell growth. The nuts and bolts for this work were laid by past researchers in our laboratory, particularly Dr. Lian[17] who constructed the whole-genome microarray using PCR-amplified gene segments, and Dr. Kyung who created disruption mutants for several genes with regulatory functions. Dr. Lian and Dr. Jayapal[18] established the protocol and characterized the temporal transcriptome profiles of many of these mutants. Besides this, a large amount of transcriptome data analyzed in this work was obtained from DNA microarray repositories (Stanford Microarray Database, Gene Expression Omnibus, and ArrayExpress).

Gene regulation in bacteria occurs at the level of operon, i.e. a group of chromosomally contiguous genes that is transcribed as a single mRNA unit. In this study, considerable effort was placed to construct and experimentally verify a whole-genome operon map for *S. coelicolor*. Predictive data mining tools were employed to *learn* transcriptomic, functional, and genomic attributes of a set of literature-reported known operons. All the subsequent efforts to elucidate regulatory associations were performed at the level of cistrons, i.e. monocistronic genes and polycistronic operons.

Second part of the thesis research focuses on discerning the distinguishing factors of high recombinant protein productivity in mammalian cells. There is profound interest in comprehending the molecular basis of high productivity. It is our belief that high productivity is a culmination of several physiological attributes such as high protein synthesis and secretion, and better growth and metabolic characteristics. In this study, two types of datasets were analyzed.

Transcriptome data was obtained from several high and low antibody-producing NS0 mouse myeloma cells. The early work on transcriptome-based comparison of high and low-producing cells using statistical differential expression analysis was carried out by Dr. Seth[19]. This study focused on mining the transcriptome data using tools for binary pattern classification and statistical 'gene set' analysis to unveil significant correlations between biological functions and the observed phenotype.

The second dataset comprises the comprehensive process archives of Genentech's commercial facility for recombinant protein production. The enormity and heterogeneity of bioprocess data have resulted in limited computational efforts to systematically scrutinize these historical archives. In this study, a flexible data mining framework was proposed to integrate and analyze heterogeneous bioprocess data from several production runs.

1.3 THESIS ORGANIZATION

This thesis is organized into eight chapters. Chapter 2 reviews the biological traits that were investigated in this study. Chapter 3 outlines the data mining-based knowledge discovery approach for analyzing large and heterogeneous datasets, with emphasis on bioprocess data. Chapter 4 describes a machine learning approach to build and experimentally verify a genome-scale operon map for *S. coelicolor* using a compilation of global temporal transcriptome data. This operon map was refined in chapter 5 and a 'reverse engineering' approach was used to combine transcriptomic and gene functional features to predict regulatory linkages in *S. coelicolor*. Chapter 6 highlights the application of functional investigation tools to relate physiological pathways with the observed antibody productivity phenotype in recombinant NS0 cell lines. Chapter 7 outlines the development of a data mining framework to predict recombinant protein productivity in mammalian cell-based manufacturing processes using vast archives of process data. Chapter 8 summarizes the conclusions of this study and provides suggestions for further research.

CHAPTER 2 BACKGROUND

This chapter provides an overview of the biological systems and the traits that were investigated in this study. Specifically, the emphasis is on *Streptomyces coelicolor*, a model prokaryote of the *Streptomyces* genus, which produces antibiotics. The biological mechanisms that regulate antibiotic synthesis in *S. coelicolor* are particularly underlined. The second part discusses the various aspects of recombinant protein production in mammalian cells. The composite nature of high productivity trait is highlighted by summarizing the past attempts to understand and engineer cells to achieve superior productivity characteristics.

2.1 STREPTOMYCETES

2.1.1 PHARMACOLOGICAL RELEVANCE

The species of *Streptomyces* genus produce a variety of natural products including antibiotics, anti-tumor agents and immunosuppressants. A few prominent ones include streptomycin–the first antibiotic treatment for tuberculosis, and vancomycin–the 'last resort' medication against Gram-positive pathogenic microorganisms. It is estimated that species of *Streptomyces* genus produce approximately two-thirds of the thousands of antibiotics and natural products made by bacteria and fungi[20]. Approximately 250 of these antibiotics are commercially manufactured and their worldwide market is currently more than US $25 billion[1]. Most of the bioactive molecules in *Streptomyces* species are produced by complex secondary metabolite pathways that are widely believed to play no essential role during vegetative growth, but serve as chemical weapons to provide a selective advantage to the bacteria in their competitive soil habitat[21].

2.1.2 LIFE CYCLE OF STREPTOMYCETES

Streptomycetes belong to the taxonomic order *Actinomycetales*, a branch of Gram-positive bacteria characterized by their high-GC content (> 55%). Their widespread occurrence in soil is greatly facilitated by their complex life cycle during which they differentiate into spores that are

resistant to desiccation and low nutrient availability, and can persist in a dormant state for many years. The complete life cycle of *Streptomyces* is markedly complex. When unigenomic spores are exposed to a growth-conducive environment, a germ tube emerges from spores and grows at the tip to form filamentous hyphae, which later branch off to form a mesh of substrate mycelium. Later in the growth stage, certain hyphae grow away from the substrate to form aerial mycelium. This stage corresponds to the onset of many changes, which may represent stress responses induced by nutrient limitations. These changes include production of antibiotics and secondary metabolites, the lysis of some portions of substrate mycelia, and initiation of glycogen storage metabolism. The aerial hyphae, supplied with nutrients from the substrate mycelia, continue to grow as multigenomic filaments, which finally undergo synchronous cell division to produce several unigenomic spores[22].

2.1.3 COMPLETE GENOME SEQUENCE OF MODEL ORGANISM – *STREPTOMYCES COELICOLOR*

The model organism of this genus, *Streptomyces coelicolor*, has been the focus of genetic studies for the past five decades[23]. The complete genome of *S. coelicolor* was sequenced in 2002[24]. With a linear chromosome of length 8.7 Mbp, *S. coelicolor* had the largest genome amongst all the sequenced bacteria up until 2002. The genomes of two other species, *Streptomyces avermitilis* and *Streptomyces griseus,* have also been sequenced recently[25, 26].

The genome of *S. coelicolor* has 7825 predicted open reading frames (ORFs), almost twice the number of genes predicted in the endospore-forming *Bacillus subtilis* (4099 ORFs). Strikingly, 965 gene products, constituting 12.3% of the genome, have predicted regulatory functions. This includes 65 sigma factors – an exceptionally large number in contrast to *Escherichia coli,* which has seven sigma factors. Extra-cytoplasmic function (ECF) sigma factors form the largest group, comprising 45 members. These sigma factors respond to external stimuli and subsequently transcribe the genes involved in various cellular events such as cell-wall homeostasis (sig^E)[27] and aerial mycelium formation (*bldN*)[28]. Several stress-inducible sigma factors have also been identified[29]. *S. coelicolor* also has a large number of two-component regulatory systems, including 84 sensor kinases, of which 67 were found adjacent to genes encoding corresponding response regulators[30]. Further, more than 20 gene clusters have been identified in the genome, which encode enzymes responsible for the production of secondary metabolites, such as antibiotics, siderophores (involved in iron acquisition), and pigments.

2.1.4 ANTIBIOTICS PRODUCED BY *S. COELICOLOR*

Among the secondary metabolite biosynthetic clusters in *S. coelicolor* genome, the products of four clusters are the four known antibiotics – Actinorhodin (Act), undecylprodigiosin (Red), calcium-dependent antibiotic (CDA) and methylenomycin (Mmy). Two of these antibiotics, Act and Red, are pigmented. Hence, their onset can be easily detected by visual inspection and titers measured by spectrophotometric assays. The structures of these antibiotics are shown in Figure 2.1.

Undecylprodigiosin (Red) is a red-pigmented tripyrrole antibiotic that is produced in the late exponential or early stationary phase in suspension cultures. In fact, the red-pigmented appearance of the mycelial pellets is attributed to four tripyrroles of which undecylprodigiosin and butylcycloheptylprodiginine are the most abundant components[31]. The antibiotic is synthesized from mono and bipyrrole precursors using complex, branched pathways that involve acetate, glycine, alanine, proline, methionine, and serine[32, 33]. All the genes involved in Red biosynthesis and resistance are clustered together on the chromosome as a 36 kb fragment, which contains 27 ORFs.

Actinorhodin is a dimeric benzoisochromanequinone and it belongs to the class of type II polyketides. Polyketides (PKs) are an extensively studied class of natural products, which include important antibacterial agents (erythromycin, tylosin), immunosuppressants (rapamycin), and cholesterol-lowering compounds (lovastatin)[34]. Actinorhodin was originally isolated as a red pigment from a strain of *S. coelicolor* in 1950[35]. It acts as a pH indicator turning blue above a pH of 8.5 and is red below. Although the intracellular product was identified as actinorhodin itself, the blue color of the medium has been attributed to γ-actinorhodin, a lactone form of the antibiotic. Actinorhodin is synthesized by enzymes encoded in the *act* biosynthetic cluster, which is 22 kb long and includes 23 ORFs. The cluster also contains genes encoding proteins responsible for antibiotic resistance and transport across the cell membrane.

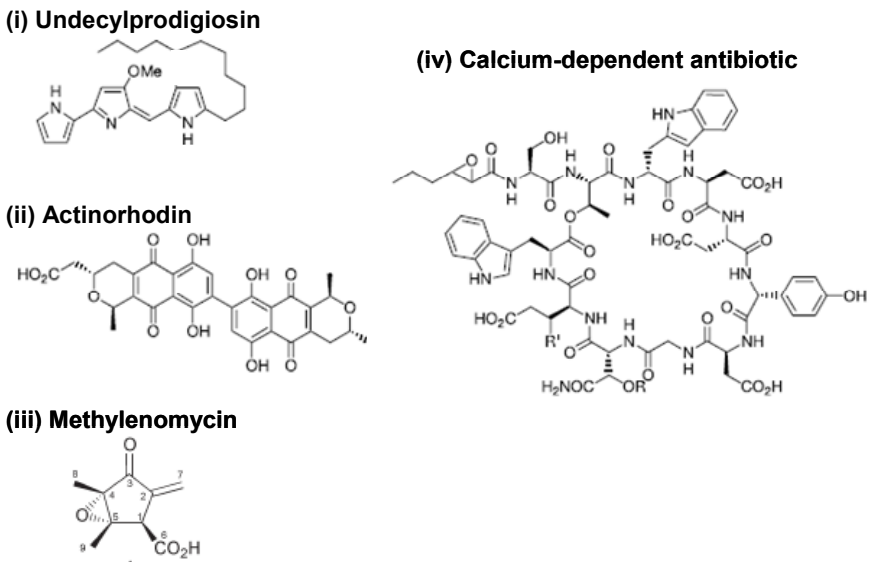

Figure 2.1. Chemical structures of the four known antibiotics in *S. coelicolor* (from ref.[24, 36])

Methylenomycin (Mmy) is one of the few known examples of antibiotic clusters in *Streptomyces* located on a plasmid instead of chromosome[37]. The *mmy* cluster comprises a 17 kb segment located on a large linear plasmid SCP1 in *S. coelicolor*[38].

Calcium-dependent antibiotic (CDA) belongs to the class of peptide antibiotics that are synthesized by non-ribosomal peptide synthases (NRPS). It is a 11-amino acid cyclic lipopeptide with an *N*-terminal fatty acid side chain[22]. Huang *et al.*[39] identified the chromosomal boundaries of the *cda* cluster by clustering the expression profiles of contiguous genes in the region, which suggested that the *cda* cluster extends approximately 83 kb and comprises 40 ORFs.

2.1.5 REGULATION OF ANTIBIOTICS SYNTHESIS

2.1.5.1 Cluster-situated regulators (CSRs)

The production of the four antibiotics (Act, Red, CDA and Mmy) is controlled by the regulatory proteins that are encoded by genes present in the respective gene clusters – ActII-ORF4[40-43], RedD[44-46], CdaR[47] and MmyR[21]. The *actII-ORF4* and *redD* mutants fail to express the *act* and *red* biosynthetic genes and consequently fail to produce Act and Red antibiotics. The accumulation of *redD* and *actII-ORF4* transcript is growth-phase dependent, with rapid increase in their mRNA level as the culture enters stationary phase[48, 49]. In addition to RedD, the *red*

cluster has an additional cluster-situated regulator, RedZ, which controls the transcription of *redD*[50].

Many of these pathway-specific regulators have defined a family of regulatory proteins called SARPs (for *Streptomyces Antibiotic Regulatory Proteins*)[51]. SARPs contain an *N*-terminal OmpR-like DNA-binding domain and they are predicted to bind to the promoter regions of the target biosynthetic genes at heptameric direct repeats. This has indeed been confirmed for ActII-ORF4 by DNase I footprinting assay[52].

The identification and characterization of these pathway specific regulators puts forth intriguing questions about the genetic circuits that regulate the expression and interplay between these cluster-situated regulators. Indeed, this is the focus of the ongoing research in several laboratories across the world.

2.1.5.2 Two-component signal transduction systems in S. coelicolor

A two-component system consists of a sensor kinase and its cognate response regulator. The sensor kinase is an integral membrane protein with an *N*-terminal sensor domain in the cytoplasmic membrane and a *C*-terminal histidine kinase/ATPase domain. Sensor kinases generally function as homodimers and when the sensor domain recognizes an environmental stimulus, each kinase domain phosphorylates its partner at a conserved histidine residue in the *C*-terminal domain. The phosphate group is then relayed to the conserved aspartate residue on the receiver domain of the cognate response regulator. The response regulator also has a helix-turn-helix DNA binding domain, and the DNA-binding activity of the activated regulator controls the transcriptional activation/repression of the target genes[53]. The presence of nearly 80 two-component systems in *S. coelicolor*, 25% more than the average number found in the genomes of non-pathogenic, free-living bacteria[30], suggests that *S. coelicolor* is well-equipped with a system to rapidly respond to a wide-range of environmental cues.

So far, three two-component systems in *S. coelicolor* have been reported to have a global regulatory effect on synthesis of multiple antibiotics. These are *afsQ1/afsQ2*[54], *cutR/cutS*[55], and *absA1/absA2*. Particularly, *absA1/absA2* has been the focus of several studies in the past two years. The *absA1/absA2* two-component system was identified by the analysis of mutants that resulted in the loss of production of all the four antibiotics in *S. coelicolor* (Abs$^-$ phenotype)[56, 57]. AbsA1 was identified as the sensor kinase and AbsA2 as the response regulator. Disruption of either *absA1* or *absA2* resulted in precocious hyperproduction of antibiotics (*Pha* phenotype)[58]. Subsequent studies have shown that phosphorylated AbsA2 can negatively regulate antibiotic

synthesis[59]. The transcript levels of the pathway specific activators, *actII-ORF4* and *redD* were elevated in the *Pha* phenotype, whereas their levels were reduced in the Abs$^-$ phenotype. However, for a long time it was unclear whether the regulatory effect of AbsA2 on antibiotics is direct or mediated via other components. A recent study showed that phosphorylated AbsA2 can bind to the promoter regions of the genes encoding the CSRs of Act, Red, and CDA thereby suggesting that the negative regulatory effect is via direct transcriptional control[60].

2.1.5.3 Eukaryotic-type serine/threonine protein kinases

The genome sequence of *S. coelicolor* encodes 44 serine/threonine protein kinases, which are commonly found only in eukaryotes. The effects of most of the members of this family on secondary metabolism are uknown. AfsK/AfsR was the first serine-threonine kinase-mediated signal transduction pathway characterized in prokaryotes[61, 62]. Disruption of *afsK* resulted in reduction, but not complete abolition of Act. Moreover, AfsR could still undergo phosphorylation at serine and threonine residues strongly suggesting the presence of an additional kinase protein[62], which was confirmed by recent studies[63]. Similar to *afsK*, disruption of *afsR* resulted in a four-fold reduction of Act, suggesting that *afsR* is not an essential gene for actinorhodin production[64]. AfsS, a 63 amino acid protein has been identified as the downstream target of AfsR. Overexpression of *afsS* leads to increased transcript levels of *actII-ORF4* resulting in overproduction of Act[65]. Furthermore, disruption of *afsS* results in complete suppression of Act[66], highlighting the role of AfsS as an activator. However, the precise mechanism of actinorhodin regulation by AfsS is unknown.

2.1.5.4 Low molecular weight signaling molecules

γ-butyrolactones are signaling molecules that are produced at threshold levels in several *Streptomyces* species. Several studies have demonstrated that these small molecules play a major role in triggering antibiotic production (reviewed in Takano[67]). Further, these molecules can diffuse across the cytoplasmic membrane barrier and act as microbial pheromones to synchronize an entire population for antibiotic production.

A recent study showed that *S. coelicolor* synthesizes at least four γ-butyrolactones that, upon exogenous addition to a culture of growing cells, result in precocious antibiotic production. One of these compounds was purified and its chemical structure was determined to be ((2R,3R,1'R)-2-(1'-hydroxy-6-methylheptyl)-3-hydroxymethyl-butanolide (SCB1)[68]. Subsequent studies identified two adjacent and divergently transcribed genes, *scbA* and *scbR*, which play a

critical role in synthesis and regulation of SCB1. There is strong evidence to suggest that *scbA* encodes an enzyme that synthesizes SCB1[69], whereas *scbR* encodes a TetR-family transcriptional repressor with an SCB1-binding site that modulates it DNA-binding properties[68]. Although deletion of *scbA* resulted in overproduction of Act and Red antibiotics, the mechanism of this genetic trigger are not well-understood. However, recent studies show that ScbR directly represses the cluster-situated activator CpkO of a cryptic type I polyketide[70].

γ-butyrolactones have also been identified in other *Streptomyces* species. The A-factor (autoregulatory factor) in *S. griseus* comprises the most characterized γ-butyrolactone system in *Streptomyces* (reviewed in Horinouchi[71]). A threshold level of A-factor triggers a wide range of responses including streptomycin production, streptomycin resistance, formation of a yellow pigment, aerial mycelium formation, and sporulation. A-factor triggers these responses by binding to its specific receptor protein ArpA and modulating its effect on transcription of key genes responsible for these physiological responses[72]. γ-butyrolactones have also been identified in other *Streptomyces* species, and in two cases their effect on antibiotics production has been characterized. Virginiae butanolides (VB) control virgniniamycin production in *S. virginiae*[73] and IM-2 controls production of showdomycin and minimycin in *S. lavendulae* FRI-5[74]. Orthologs of ScbA and ScbR have recently been found in *S. virginiae*[75], *S. fradiae*[76], *S. lavendulae*[77], *S. pristinaespiralis*[78], and *S. clavuligerus*[79].

2.1.5.5 Effect of nutritional environment and metabolites on antibiotic production

There is substantial evidence to support the notion that expression of CSRs, *actII-ORF4* and *redD,* is a pre-requisite for Act and Red synthesis, respectively. This suggests that all the pleiotropic regulators affect Act and Red synthesis via one or more regulatory cascades that culminates at the CSRs. However, the complexity of this cascade is compounded by the observation that the effect of many pleiotropic regulators is dependent on the nutrient source. For example, many *S. coelicolor bld* mutants which lack aerial mycelia and are unable to produce antibiotics, display a nutrient-dependent phenotype; they are 'bald' on glucose-minimal medium, but can form aerial mycelium in other carbon sources such as galacotose, mannitol, or glycerol[80]. Similarly, production of Act in *Streptomyces lividans,* a close relative of *S. coelicolor*, is conditionally-dependent on the carbon source[81].

The temporal correlation between a reduction in growth rate, antibiotic production, and morphological differentiation in many *Streptomyces* species strongly suggests that these transitions are in response to nutrient depletion. A recent study showed that *N*-acetylglucosamine

derived from hydrolysis of substrate mycelium cell wall, presumably an indication of nutrient famine, can trigger antibiotic synthesis and differentiation in many *Streptomyces* species[82]. Thus, it is physiologically relevant to suspect that secondary metabolism and sporulation are undesirable in a nutrient-rich environment. In agreement with this hypothesis, glucose has been observed to repress the production of actinomycin, streptomycin, tetracycline, kanamycin and tylosin production in their natural hosts[21]. Similarly, ammonium and phosphate salts have a repressive effect on actinorhodin production in *S. coelicolor*[83]. The negative effects of phosphate *S. lividans* and *S. coelicolor* are relayed by *PhoR/PhoP*, a two-component system that responds to phosphate starvation by inducing a phosphate ABC transporter[84, 85]. Studies have also shown that induction of *ppGpp* synthesis, and indication of amino acid depletion, can trigger antibiotic production and sporulation[86].

2.1.5.6 Multiple layers of control for antibiotic synthesis

The transcriptional induction of antibiotic pathway-specific genes by cluster-situated regulators provides a simple and elegant 'synthesize-when-necessary' paradigm predominant in prokaryotes. However, additional layers of regulatory control exist in *S. coelicolor* that modulate the activity of the final gene product involved in secondary metabolism. *absB* gene encodes an RNase III homolog that controls its target genes at post-transcriptional level. Loss-of-function mutations in *absB* result in a global defect in antibiotic synthesis[87]. Similarly, *bldA* gene encodes a rare tRNA codon UUA for leucine. *bldA* mutants fail to produce Red and Act due to the presence of the rare codon in *redZ* and *actII-ORF4,* respectively, indicating a mechanism for translational attenuation of antibiotic synthesis.

2.1.6 GLOBAL GENE EXPRESSION PROFILING TO INVESTIGATE SECONDARY METABOLISM

The availability of *S. coelicolor* genome has fueled much interest in exploring secondary metabolism by global gene expression profiling. In the first study reported in 2001, Huang *et al.*[39] used DNA microarrays to survey the temporal transcriptome profiles of wild-type strain and Red and Act-deficient mutants and identified a set of genes whose expression is co-ordinated with Red and Act antibiotic cluster genes. Karoonuthaisiri *et al.*[88] investigated the temporal transcriptional response of *S. coelicolor* to a variety of stress conditions, such as osmotic shock, temperature shift (upshift and downshift) to identify a set of common differentially expressed transcripts that

may represent the underlying similarities in response to these stress conditions. Using whole-genome microarrays to compare an *afsS* disruption mutant with the parent strain, studies from our laboratory have shown that the pleotropic regulator AfsS can completely abolish Act production and also modulate nutrient starvation response[66]. Hesketh *et al.*[86] recently probed the regulatory role of stringent response factor, guanosine pentaphosphate (ppGpp) in secondary metabolism using Affymetrix GeneChip arrays.

Proteomic tools have also been employed to survey the translational and post-translational facets of gene expression. A 2D-gel and matrix-assisted laser desorption ionization time-of-flight (MALDI-TOP) approach identified 770 proteins across different functional classes, many of which were associated with post-translational modifications[89]. Recent studies have examined and compared time series proteomic and transcriptomic datasets. Using singular value decomposition, a matrix factorization technique, Vohradsky *et al.*[90] alluded that the overall population dynamics are similar at mRNA and protein levels. Jayapal *et al.*[91] used isobaric stable isotope labeled peptides (iTRAQ^{TM}) to quantify more than 1100 proteins, which were compared with the corresponding transcript profiles. In contrast to Vohradsky *et al.*, it was observed nearly 30% of genes exhibit dissimilar dynamics suggesting that factors such as differences in mRNA and protein turnover, translation efficiency, and codon usage could play a substantial role in gene regulation.

2.1.7 ANTIBIOTIC REGULATION – A SYNOPSIS

Streptomyces coelicolor, with an 8.7 Mbp linear chromosome and approximately 7800 predicted ORFs, has one of the largest completely sequenced bacterial genomes. It belongs to the genus *Streptomyces* whose members are widespread in soil, have complex multicellular lifecycle, and produce nearly two-thirds of reported naturally occurring antibiotics and a variety of other natural compounds including anti-tumor agents and immunosuppressants. With its complex life cycle and the capacity to produce numerous antibiotics, *S. coelicolor* has a very dynamic life cycle of vegetative growth and sporulation, undergoing changes from primary metabolism to secondary metabolite production. Figure 2.2 summarizes the known regulatory elements that affect secondary metabolism in *S. coelicolor*. It is noteworthy that the mechanistic details of majority of these interactions are unclear. Further, except in a few cases, it is not known whether the effect of the regulatory genes on antibiotics is direct or relayed by one or more intermediate components. The availability of whole-genome sequence for three *Streptomyces* species and tools

for global gene expression profiling have considerably widened the scope for delineating the regulatory machinery which modulates the fundamental physiological processes that trigger the synthesis of the commercially relevant secondary metabolites in *S. coelicolor* and other *Streptomyces* species.

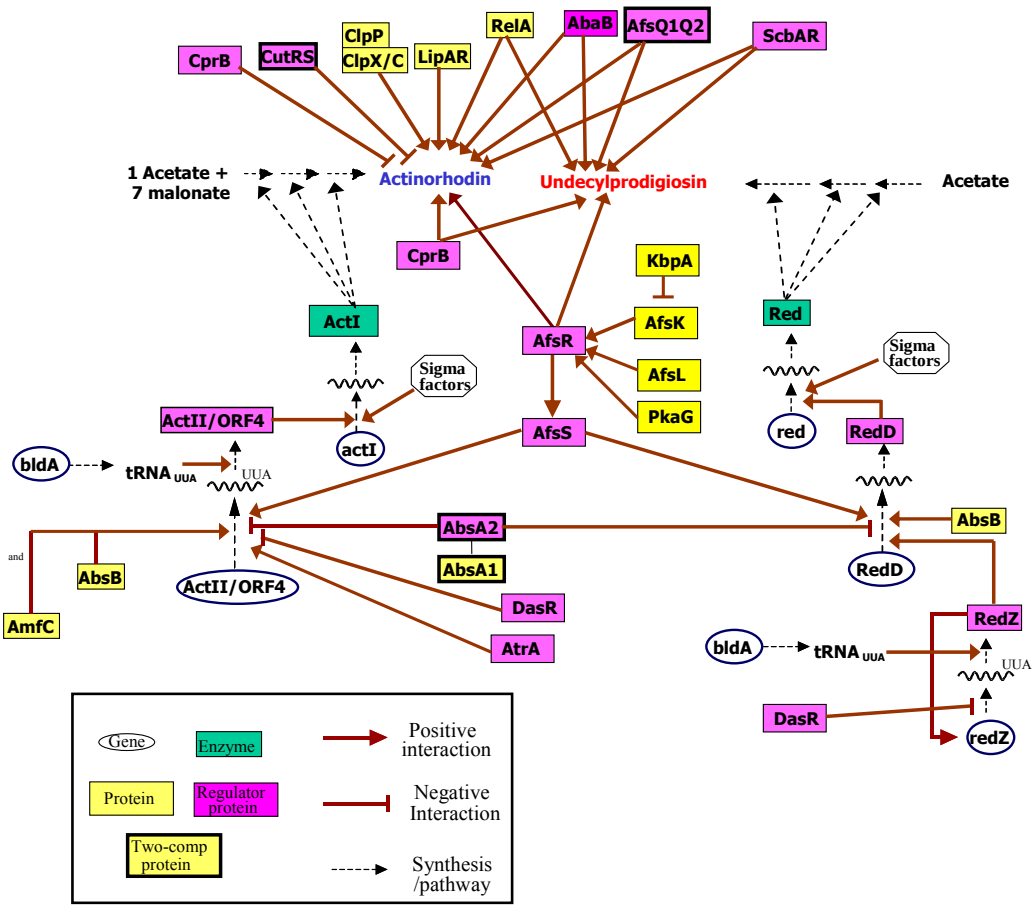

Figure 2.2. Known regulatory map for *S. coelicolor* (modified from Mehra[92])

2.2 MAMMALIAN CELL CULTURE

In the past two decades, recombinant proteins have provided life-saving therapies for several debilitating illnesses. The high therapeutic value of these recombinant proteins has caused the market for biologics to grow by 21% and 10% in 2006 and 2007, respectively. The annual market for biologics in 2007 was estimated at US $44 billion[93]. Recombinant proteins constitute the bulk majority of these biologics. Nearly 60-70% of these protein biologics are produced in mammalian cell 'hosts' to ensure proper protein-folding and post-translational modifications (e.g. glycosylation), which affect the stability and the efficacy of the drug and cannot be achieved in bacterial hosts[94]. Some of the recent US FDA-approved drugs include monoclonal antibodies such as Avastin (antibody against vascular endothelial growth factor) by Genentech for treatment of metastatic colorectal, lung and breast cancer; Humira (antibody against tumor necrosis factor) by Abbott Labs for rheumatoid arthritis treatment; Vectibix (antibody against epidermal growth factor receptor) by Amgen for treatment of metastatic colorectal cancer. With 26 FDA-approved drugs, monoclonal antibodies (mAbs) constitute the most important category of biologics. Other categories of protein biologics include growth factors, cytokines, therapeutic enzymes, blood factors, and anti-coagulants.

2.2.1.1 Host cell lines for recombinant protein production

A vast majority of recombinant proteins are commercially produced using Chinese hamster ovary (CHO) cells. CHO cells were derived from the ovary of a female Chine hamster and cultivated in tissue cultures in 1958 by Tjio and Puck[95]. Subsequent mutagenic studies led to the isolation of mutants deficient in dihydrofolate reductase (DHFR) enzyme, which reduces dihydrofolic acid to tetrahydrofolic acid, an essential precursor for thymine synthesis[96]. This folate auxotrophy allows introduction of a heterologous gene encoding the recombinant protein of interest and selection of cells that produce large quantities of the exogenous protein. Indeed, DHFR expression system is widely used in industry for development of cell lines for recombinant protein production. Apart from CHO cells, the other industrially important cell lines include mouse myeloma cells such as NS0 and Sp2/0 cells. NS0 cells were derived from plasma-cell neoplasms induced in BALB/c mice in 1962 by Potter and Boyce[97]. Although NS0 cells do not synthesize or secret the heavy and light chain immunoglobulin (Ig), their derivation from

professional antibody-secreting plasma cells suggests that NS0 cells may have retained the expanded secretion machinery for protein secretion. The glutamine synthetase (GS) NS0 system is the preferred host for expressing recombinant proteins. GS NS0 cells have little endogenous glutamine synthetase activity and, consequently, cannot produce the amino acid glutamine, which is essential for cultured cells. Hence, GS is used as a selectable marker to introduce the heterologous gene encoding the recombinant protein of interest[98].

2.2.2 THE PROCESS OF CELL LINE DEVELOPMENT[94]

The increased demand for therapeutic proteins has been met through expansion of manufacturing capacity and process intensification and development of mammalian cell lines with enhanced productivity. For each therapeutic product, a cell line with sufficient production capability is developed. Current strategies involve several labor-intensive and time-consuming steps. Recombinant DNA is introduced into mammalian cells, commonly using the DHFR or GS system. After transient transfection, a pool of cells in which recombinant DNA has integrated into the chromosome is selected. The efficiency of chromosomal integration is often lower than 0.1%. For DHFR system, the cells at this stage are cultivated in glycine, hypoxanthine, and thymidine-free medium at low levels of methotrexate (MTX), a reversible inhibitor of DHFR enzyme. Thereafter, the cells are exposed to a very high concentration of the MTX drug (upto 1μM) to co-amplify the gene of interest along with the *dhfr* gene. This step often generates clones with more than a hundred copies of the gene encoding the recombinant protein. A pool of heterogeneous cells is isolated after amplification. The heterogeneity of different clones in this pool is due to the differences in protein productivity, and the randomness associated with the site of gene integration and the copy number of the gene of interest. Individual clones are isolated from this population by limiting serial dilution and are selected based on optimal characteristics for growth and protein (product quantity and quality). It is not uncommon to screen hundreds of clones before selecting a candidate cell line for process development and volumetric scale-up. The entire process of cell line development often spans several months. Several technologies based on flow cytometry and high throughput robotics (e.g., ClonePix, Cello) have been proposed for systematic screening of large number of clones to increase the likelihood of isolating hyper producers (reviewed in Brown and Al-Rubeai[99]).

2.2.3 TRANSCRIPTOME AND PROTEOME SURVEYS TO UNDERSTAND CELL PHYSIOLOGY

The cell line development process can generate high-producing clones secreting 20-60 $pg \cdot cell^{-1} \cdot day^{-1}$. However, the process is largely empirical and the molecular determinants of high productivity are not well-understood. Hyper productivity, however, is a complex trait, which entails several essential elements such as a high transcription, translation, and secretion rate of the recombinant protein, and favorable growth and metabolic characteristics. Cells in culture typically consume large amounts of glucose and glutamine, a large proportion of which is channeled to lactic acid and ammonia. High levels of these metabolites inhibit cellular growth thereby diminishing volumetric protein productivity. Recent efforts to engineer cells with favorable production characteristics therefore include strategies to reduce lactate accumulation[100], enhance viability of the culture[101], and manipulate protein processing and secretion pathways inside cells[102].

Most cell engineering efforts to date are confined to adding or suppressing one or two factor(s) to alter a localized pathway. However, physiological behavior of cells (such as protein secretion, cell cycle) likely manifests through interaction of complex networks. The effect of localized perturbations are difficult to predict and will require a better understanding of the regulation and interaction of networks. With the availability of global transcriptome and proteome survey tools, attempts to elucidate the genes conferring the complex traits of hyperproductivity are underway (reviewed in Griffin *et al.*[103]).

Until recently, little genomic resources were available for this work in CHO cell lines. Although there has been some effort towards EST (Expressed sequence tag) sequencing[104], the information available in the public domain is still scarce compared to that for mouse, rat and human. Resorting to the use of a mouse DNA microarray for transcriptome analysis of CHO cells has been reported to give reasonable results for more highly conserved genes with a 60-mer oligonucleotide-based DNA microarrays[105]. In contrast, another study concluded that while cross-species microarrays gave roughly similar expression profiles for many genes, the data lacked sensitivity and generally failed to satisfy statistically significant differential expression criterion[106].

Due to lack of genomic resources for CHO cells, most transcriptome studies investigating recombinant protein production in mammalian cells have been carried out in mouse myeloma cells[107] or mouse hybridoma cells[108]. However, due to higher sequence similarity among species

at the protein sequence level, satisfactory results have been reported for proteomic investigations in CHO cells[109]. Reports employing transcriptome or proteome analysis to study cell physiology related to bioprocessing include surveys of cells exposed to high-producing conditions such as temperature shift[110], sodium butyrate treatment[106], and hyperosmotic stress[109].

Regulation of protein expression at transcriptional and translational levels might not always correspond. Thus, it is prudent to combine DNA microarray data with proteomics data in large scale gene expression surveys to understand the mechanism underlying a given phenotype under investigation. Limited reports comparing proteome and transcriptome data have largely shown consistent results[107, 108, 111]. Ability to survey only a limited number of proteins using 2D-gel based assay[112] is only recently being augmented by the use of isotope tagging and LC-MS/MS approaches[107].

2.2.4 DATA ANALYSIS FOR HIGH PRODUCTIVITY COMPLEX TRAITS

2.2.4.1 Gene-level differential expression analysis

Methods for statistical analysis based on inferential hypothesis testing are commonly used to identify genes that are differentially expressed between two more classes such as high- and low-producing cells, or productivity-enhancing conditions and reference conditions. A wide-array of statistical tests have been employed to identify differentially expressed transcripts (reviewed in Jeffery *et al.*[113]). Although there is no clear winner, a consensus has emerged that, for microarray studies with few samples per class (typically, $n < 5$), methods which borrow information across genes for estimating within-class variance (called 'shrinkage') perform better than simple *t*-tests[12]. Shrinkage-based methods assert that the inaccuracy in estimating the variance of a gene (due to small sample size) can be alleviated by pooling the variances of several genes. Further, resampling-based methods that estimate statistical significance (i.e., a *p*-value or the probability that the significance statistic was observed by chance) of a gene by random permutation are preferred over those methods that infer differential expression by assuming a normal distribution of gene expression level (e.g. *t*-test). A typical microarray experiment compares the expression of thousands of genes. Consequently, a *p*-value threshold of 0.05 on a 10,000-probe microarray would, on average, result in identification of 500 (0.05×10000) genes as differentially expressed by chance. Controlling the false discovery rate (i.e., the fraction of the genes that satisfy a significance threshold by chance and hence falsely called differentially expressed) has emerged as a popular choice for limiting the number of false-positives identified in a microarray dataset[12].

Several methods have been proposed for estimating false discovery rate (e.g., significance analysis of microarrays (SAM)[114], extraction of differential gene expression (EDGE)[115])

In transcriptome and proteome profiling of mammalian cell lines, whether comparing cells of different productivities or seemingly harsh culture treatments, the number of differentially expressed genes and extent of differential expression are low compared to the changes observed in other biological systems such as development or differentiation. Although very small fold changes (1.2-1.4 fold) can be physiological significant, one should be mindful that the conclusion of such small changes may be sensitive to data normalization methods or the basis of comparison. The assumption of cells having comparable amounts of total mRNA, or protein content may be subject to question, especially for cases where cell samples differ in cell volume or cytoplasmic complexity. 'Spiking' of RNA samples with external RNA controls provides an alternative method for normalization in scenarios where the assumption of equal mRNA per cell is not satisfied[116].

2.2.4.2 Gene set analysis

Biological interpretation of individual genes in a long list of differentially expressed genes presents a challenging task. Gene set analysis provides a statistical framework to assess whether a particular class of functionally related genes has been enriched with differentially expressed genes. Several methods have been proposed using different statistical tests and functional annotation schemes (reviewed in Curtis *et al.*[117]) In particular, gene set enrichment analysis (GSEA) aims to reveal gene classes that have a significant correlation to a particular phenotype. Instead of pre-selecting genes based on differential expression criteria, GSEA ranks all genes surveyed based on a phenotype-correlation metric and uses a Kolmogorov-Smirnov-like statistic to determine whether genes in a functional class are preferentially located at the extremes of the rank-ordered list[118, 119]. Preferential localization of relevant genes at the top or bottom of the list suggests that the function has been altered. Similar methods that compare the *distribution* of gene ranks (instead of the *number* of differentially expressed genes) to an overall distribution have been proposed[120, 121]. Commercial tools are also available to facilitate interpretation of differentially expressed genes in the context of biological functions (e.g., Ingenuity Pathway Analysis, http://www.ingenuity.com/)

2.2.4.3 Pattern recognition methods

To date, reports on transcriptome and proteome data on recombinant cells have largely employed statistical tools for differential expression analysis. These approaches have focused on

identifying genes conferring the hyperproductivity trait. As the studies expand and data accumulates, what will become important is the discernment of patterns or signatures of gene expression that can be used to distinguish populations of cells with different productivities. For such applications, techniques such as supervised (predictive) and unsupervised (descriptive) classification are commonly used (described in Chapter 3). The results of these data-driven, machine learning techniques often facilitate further biological interpretation of the trait and hold promise for establishing molecular signatures for hyperproductivity.

To date the applications of these pattern recognition techniques have largely been to omics datasets from disease phenotypes, especially cancer[122] (reviewed in Nevins *et al.*[8]). Given the availability of limited number of clones of varied lineage and productivity, finding the signature genes of hyperproductivity may prove to be harder than identifying marker genes for tumors, where often hundreds or even thousands of patients of a broader genetic makeup are available. Additionally, the application of these techniques to cell culture processing data will require more critical analysis and adaptation of methods to overcome uncertainties arising from the need to measure small, but physiologically significant, changes in gene expression.

2.2.5 THE COMPLEX TRAIT OF HYPERPRODUCTIVITY

In spite of the complexity, common features underlying the hyperproductivity trait are beginning to emerge. Transcriptome and proteome analysis will likely continue to be applied to two complementary studies: surveying cells with varying productivity, and examining process conditions that enhance productivity. The common gene sets that emerge from comparative studies of different conditions that enhance productivity and/or clones with varying productivity will most likely encompass functional classes or genes that potentially contribute to the complex trait of hyperproductivity.

Proteomic and genomic tools have been used to examine cells with varying levels of productivity. 2D-gel-based proteomics analysis[112, 123] revealed significant changes in the abundance (~1.3-1.6 fold) of non-endoplasmic reticulum chaperones, cytoskeletal, and metabolic proteins in high-producing NS0 cells. In another report, microarrays, 2D-gel and isotope tagging (iTRAQ™)-based proteomics were used in parallel to compare eleven NS0 cell lines classified as high or low producers[107]. Similar gene expression changes in functional classes such as protein folding, cytoskeleton organization, and carbon metabolism were reported between high and low producing NS0 cells. Differential expression of chaperones and disulfide isomerases, such as

endoplasmin precursor (ENPL), protein disulfide isomerase member 6 (Pdia6) and prolyl 4-hydroxylase (P4hb) as revealed in these studies suggests an increased protein folding load in high-producing cells. Genes involved in cytoskeleton reorganization (such as actin and tubulin alpha-1 chain)[108, 112], and redox balance (genes such as thioredoxin and peroxiredoxin) were also found to be differentially expressed[107, 108, 123].

A recent report compared antibody productivity response of CHO and mouse hybridoma cells (MAK) to hypothermic treatment and butyrate exposure[124]. Among the two cell lines and two treatments, except hypothermic treatment of MAK cells, three combinations resulted in enhanced productivity. Comparative studies of genes differentially expressed only under conditions positively affecting the productivity but not under hypothermic treatment of MAK cells provides further cue for the involvement of protein secretion and cytoskeleton-related elements in enhanced recombinant protein productivity. Of particular interest is the suggested involvement of protein trafficking molecules in the secretory pathway. Transport of protein among endoplasmic reticulum (ER), Golgi compartments and plasma membrane is mediated by trafficking of vesicles that overlap with endocytotic pathway. Therefore, a producer with high protein secretion rate is likely to have a higher vesicle trafficking and membrane recycling activity to maintain membrane and organelle homeostasis.

While examining the complex trait of hyperproductivity one may draw parallels to the differentiation of B cells to plasma cells[125]. Transcriptome and proteome studies[126, 127] have implicated the involvement of organelle biogenesis (akin to unfolded protein response) and enhanced energy metabolism in this transformation of non-secretory B cell to a professional secretor. A known Xbp1 target – vesicle docking protein p115 (Vdp), which is upregulated up to two-fold during B cell differentiation[128] was also upregulated in CHO cells upon butyrate treatment[108]. Considering that hyperproducers secrete a substantial fraction of the synthesized protein, it may not be surprising that hyperproductivity also requires higher levels of energy. No evidence that high-producing cells consume more energy (e.g. through higher specific oxygen consumption on per cell or per cell mass basis) has been presented; however, an unusually high isolation frequency of ESTs from mitochondrial genome of CHO cell line producing recombinant immunoglobulin G (IgG) was reported[104]. An increase in mitochondrial mass by 1.5 fold in cells with high specific productivity of the expressed antibody was also observed by Dinnis *et al.*[112].

2.2.6 MAMMALIAN CELL CULTURE – A SYNOPSIS

In the past two decades, recombinant proteins have provided life-saving medicines to thousands of ailing patient worldwide. Mammalian cells that synthesize these drugs in large quantities are the pivotal components of these biotechnological innovations. Studies on global expression profiling at transcriptome and proteome level suggest that high productivity entails a wide-range of alterations in multiple cellular functions such as protein processing and secretion, cell growth and death, and cytoskeletal organization. As we move forward, genomic survey tools combined with statistical and data mining methods will play a valuable role in our quest to decipher the molecular attributes of this complex hyperproductivity trait.

CHAPTER 3 MINING BIOPROCESS DATA: OPPORTUNITIES AND CHALLENGES

3.1 SUMMARY

Modern biotechnology production plants are equipped with sophisticated control, data logging and archiving systems. These data hold a wealth of information that can shed light on the causes of process outcome fluctuations, whether the outcome of concern is productivity or product quality. These data might also provide clues on means to further improve process outcome. Data-driven knowledge discovery approaches can potentially unveil hidden information, predict process outcome, and provide insights on implementing robust processes. Here we describe the steps involved in process data mining with an emphasis on recent advances in data mining methods pertinent to the unique characteristics of biological process data.

Although bioprocess data is the focus of the discovery approach outlined in this chapter, the data mining techniques discussed in this chapter provide a toolbox for analyzing large-scale datasets acquired from other sources, particularly from omics technologies. In the subsequent chapters, these data mining tools were employed to: (i) predict a whole-genome operon map and the transcriptional regulation network of antibiotic-producing *S. coelicolor* using genomic datasets (Chapters 4 and 5), (ii) decipher the characteristics of high recombinant protein productivity in mammalian cells using transcriptome (Chapter 6) as well as process data (Chapter 7).

3.2 INTRODUCTION

In the past two decades we have witnessed a major transformation of bioprocess manufacturing. Protein-based therapeutics have overtaken natural product-based drugs as the major biologics. A majority of the protein therapeutics are produced using recombinant mammalian cells. They are manufactured in modern production plants equipped with systems for automated control as well as comprehensive data collection and archiving. These archives represent an enormous opportunity for data mining in that they might unearth a wealth of

information for enhancing the robustness and efficiency of manufacturing processes. However, despite the stringent process control strategies employed, variations in the final process outcome are commonly observed. With each production run valued at millions of dollars and every manufacturing plant costing hundred million dollars and upwards, there is a great potential for cost saving through mining process databases in order to uncover the distinguishing characteristics of a good process.

This chapter discusses the challenges associated with investigating bioprocess data and the techniques that have been previously proposed to mine process data. We describe a scheme to systematically analyze a complex bioprocess dataset, and also highlight the recent advances in data mining, which are applicable for analyzing bioprocess data.

3.3 CHARACTERISTICS OF BIOPROCESS DATA

Any modern bioprocess plant maintains electronic records of material input (quantity, quality control records, lot number), process output (cell density, product concentration and quality, etc.) control actions (base addition, CO_2, O_2 flow rate, etc.) as well as physical parameters (agitation rates, temperature, etc.), from the frozen cell vial to the production scale bioreactors. Based on the frequency of measurements, bioprocess parameters can be categorized into different types. A vast majority of the process data is acquired on-line. However, a few key measurements, such as viable cell density and concentrations of product and some metabolite and nutrients are measured off-line (Figure 3.1). While the off-line parameters are measured periodically, many on-line parameters are measured continuously with respect to the time scale of the production cycle. Additionally, the information about some process parameters may be available at a single time point only. For example, product concentration and quality index might be measured at the final time point, before or after product recovery. Bioprocess data is thus heterogeneous with respect to time scale. Process data is also heterogeneous in terms of data types. Some parameters are continuous, such as cell and product concentrations, pH, while others are discrete or even binary, such as the valve settings for nutrient feeding and gas sparging, which can only be in the ON or OFF state. Even quality-related parameters for either raw material or product can be discrete. For example, the glycosylation profile as a measure for the quality of a glycoprotein is often evaluated by the discrete distribution of different glycans. Due to these heterogeneities in time scales and data types, bioprocess data is significantly different from the data arising in other

application areas in which data mining methods have been used (e.g., retail records). These heterogeneities should be taken into consideration when data mining methods are devised.

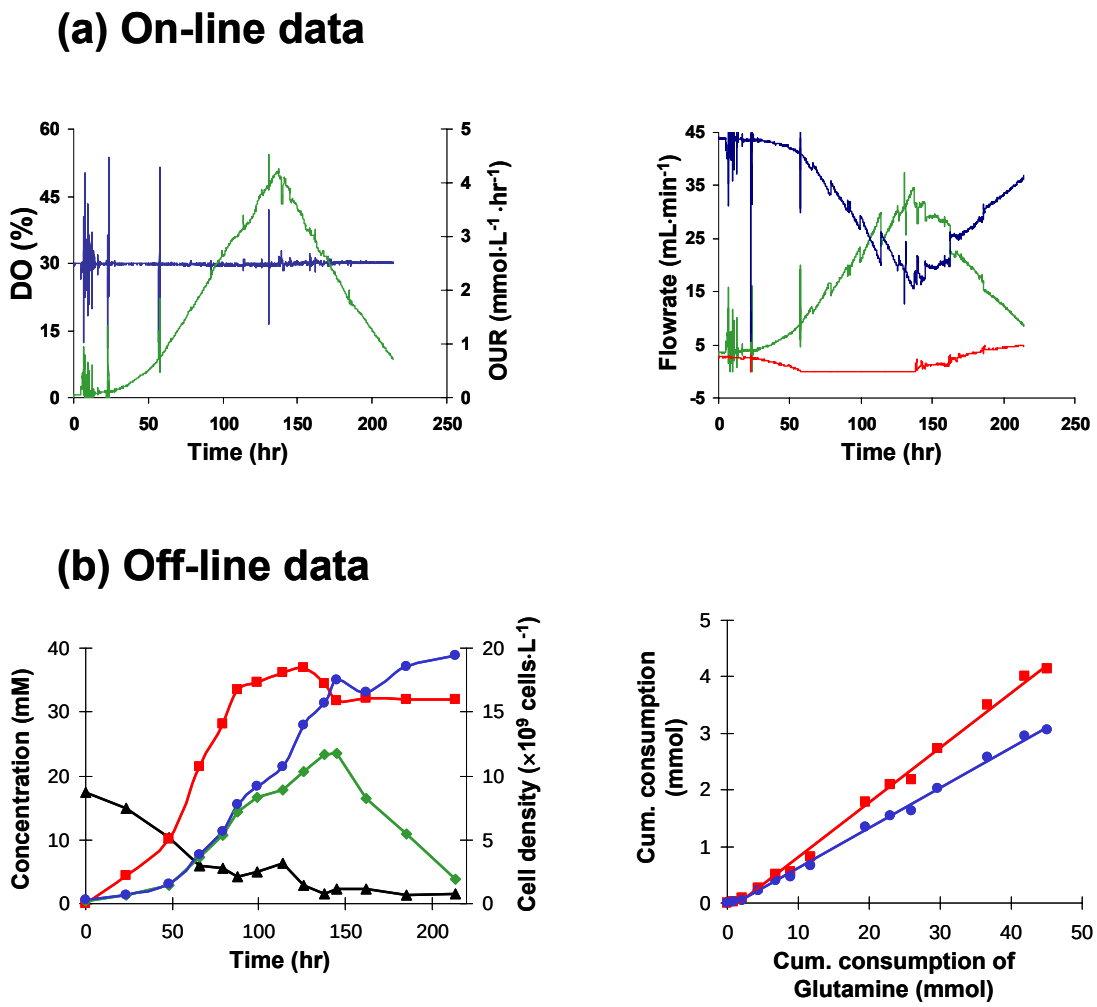

Figure 3.1. Example of bioprocess data. (a) Representative online data are shown, which are recorded every few minutes during the entire culture duration. The left panel shows the profile of typical reactor state parameters such as percent air saturation of dissolved oxygen (shown in blue) and the oxygen uptake rate (in green). The right panel shows the profile of gas flow rates as common control action parameters. Observed curves are for nitrogen in blue, oxygen in green and carbon dioxide in red. (b) Typical off-line data for a process are shown. The left panel illustrates the raw data containing biochemical parameter profiles for the total cell density (in blue), viable cell density (in green), glucose (in black), and lactic acid (in red). The right panel shows the profile for parameters that have been derived form the raw data and that are physiologically relevant, such as the cumulative consumption or production of nutrients and metabolites. Shown here are the consumption of threonine (in red) and phenylalanine (in blue) with respect to the consumption of the key nutrient glutamine. The slope of the linear regression provides the stoichiometric ratio of

threonine and phenylalanine with respect to glutamine. Representative data obtained from experiments performed by Dr. Lee[129].

3.4 KNOWLEDGE DISCOVERY AND BIOPROCESSES

The aim of mining bioprocess data is to uncover knowledge hidden within the enormous amounts of data associated with different process runs that can be used to improve and enhance the robustness and efficiency of production processes. This is achieved by analyzing different types of process runs in order to identify novel and useful relations and patterns that associate various aspects of the production process with different measures of process outcome, such as product titer and product quality. These process outcome measures are often used to categorize process runs into different classes. For example, if product titer is the outcome of interest, the different runs can be classified as 'high' or 'low' producing runs. Similarly, process runs can be grouped as 'good' or 'bad' using product quality as the metric of process outcome. The notion of gaining knowledge by scrutinizing large volumes of data has been applied to a wide array of problems ranging from from image classification in astronomy to identifying fraudulent activities in financial transactions[130].

A typical knowledge discovery process entails several iterative steps (Figure 3.2). These steps include: data preprocessing, feature selection and/or dimensionality reduction, data mining, and expert analysis for interpretation of the results. The data acquired in a bioprocess typically include some parameters that are not readily amenable for analysis. The data preprocessing step transforms these data into a form (called feature) that is suitable for the subsequent steps. This usually involves various steps including data cleaning, normalization, transformation, denoising, and missing value imputation. In a subsequent step of feature selection or dimensionality reduction, the obtained features are analyzed in order to obtain the set of features that is best suited for data mining. This often involves the selection of those features that correlate most with process outcome, and the combination of highly correlated features. The data mining step applies various computational methods, such as pattern recognition and machine learning to discover any significant trends within the data. These trends are useful for describing any correlations between process parameters and for developing *models* to predict the process outcome. Finally, during the expert evaluation step, the validity of the produced results is assessed by those knowledgeable of the process (domain experts) to discern the effect of the discovered correlations on cellular physiology and process outcome.

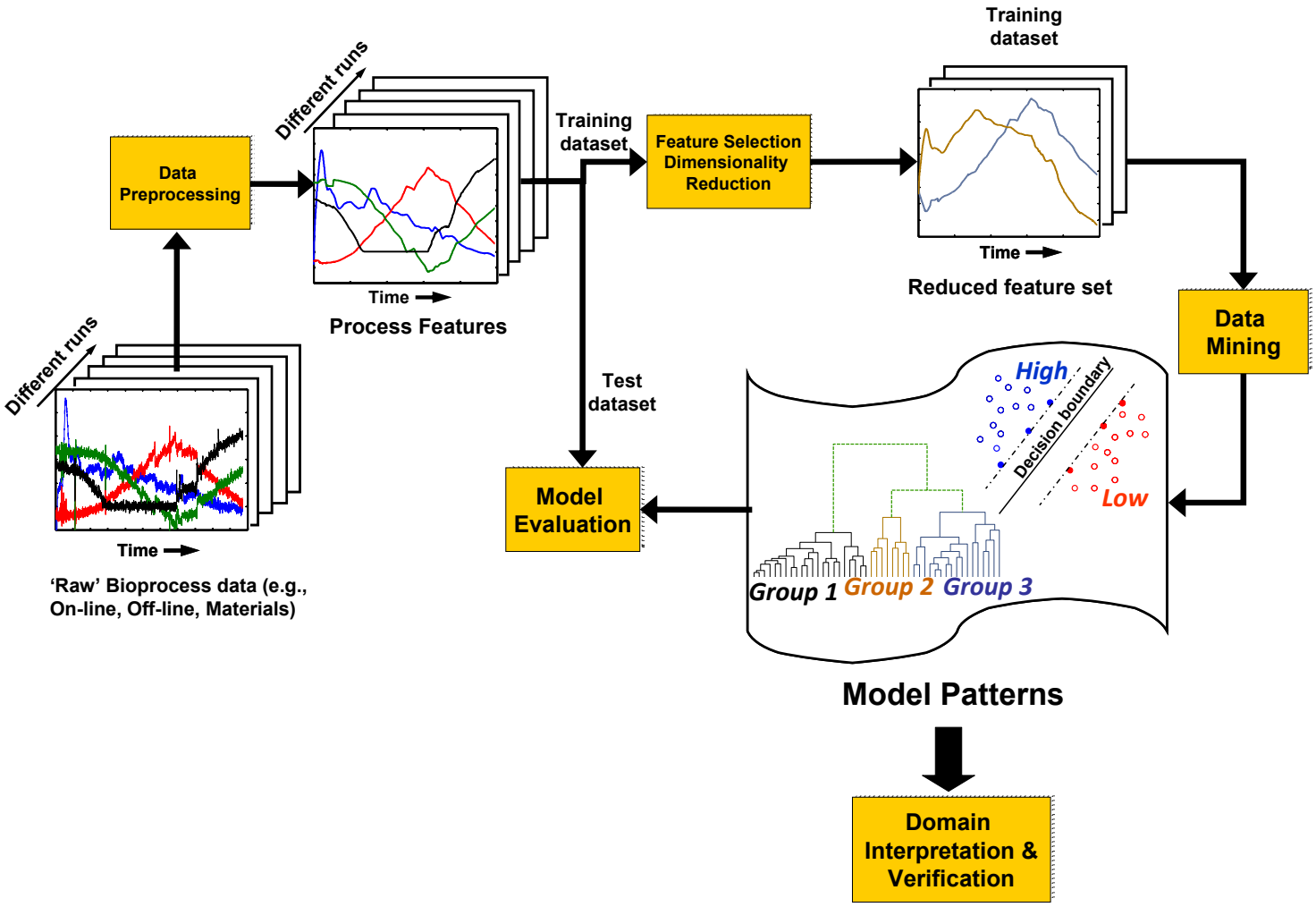
Different runs
Time
'Raw' Bioprocess data (e.g., On-line, Off-line, Materials)
Data Preprocessing
Different runs
Time
Process Features
Training dataset
Test dataset
Feature Selection Dimensionality Reduction
Training dataset
Time
Reduced feature set
Data Mining
Model Evaluation
High
Decision boundary
Low
Group 1
Group 2
Group 3
Model Patterns
Domain Interpretation & Verification

Figure 3.2. An approach for data-driven knowledge discovery in bioprocess databases. Process data includes off-line and on-line parameters, as well as raw material logs. Representative raw profiles from four temporal process parameters of a single run are shown. Process data from several runs are preprocessed in order to extract compact and smoothened features that depict the underlying process signals. The entire dataset is then split into a *training* subset, which is used for model construction, and a *test* subset, which is used for model assessment. Feature selection or dimensionality reduction is implemented on the training dataset. For example, principal component analysis can be used to identify two dominant patterns in the dataset shown here and thereby reducing the number of the initial features by half. Data mining methods are applied on the reduced feature set with the aim to discover model patterns, which are subsequently evaluated on the test dataset. The training and evaluation procedure can be repeated multiple times for further refinement of the model. Thereafter, the model patterns can be interpreted and verified by process experts, and the gained knowledge can be used for process enhancement.

3.4.1 DATA PREPROCESSING

Modern production plants are electronically supervised and create process records that are well-characterized and less prone to human errors, which significantly reduce some of the preprocessing requirements that are often associated with data cleaning and missing values imputation. However, the temporal nature of the data obtained from fermentation, cell culture, and downstream processes creates some unique challenges that need to be addressed with data preprocessing methods.

In particular, on-line parameters are often recorded every few minutes for the entire culture period that can last from a couple of days to two weeks. The culture period may even extend to a few months for some continuous or perfusion-based processes. The resulting long time series need to be preprocessed in order to extract the features that compactly and smoothly represent the underlying process signals. In addition, preprocessing is also important to eliminate the noise that may be present in process measurements due to instrument limitations and sampling artifacts. The work of Cheung *et al.*[131, 132] and Bakshi *et al.*[133, 134] laid the framework for extracting useful information from temporal process parameters. Cheung *et al.* proposed a triangular representation method in which a parameter profile was segmented into different time intervals. Within each interval, the first and second order derivatives of the profile were used to represent an increasing or decreasing trend. Bakshi *et al.* on the other hand, proposed the use of wavelet decomposition to deduce temporal features. Besides these two approaches, several other approaches can be used, such as discrete Fourier transform, methods for piecewise approximation (such as piecewise linear approximation, adaptive piecewise constant approximation), and symbolic aggregate approximation (SAX). Among these, SAX leads to a string-based representation of a parameter profile. This representation is directly amenable to several string manipulations and data mining

methods that have been developed for the analysis of protein and DNA sequences, including methods for protein structure predictions[135] and discovery of *cis*-regulatory elements[136].

In addition, due to the occurrence of a lag phase or due to variations in the growth rate, the time series obtained from different runs may not be temporally aligned. As a result, identical time points might not represent similar process states. Ignoring such time scale differences and directly comparing identical time points across different runs, for example by mean hypothesis testing methods[137, 138], can lead to incorrect results. This problem can be addressed by aligning the time series of different runs during the preprocessing step. A dynamic time warping strategy, originally developed for speech recognition[139], can be used to align the time profiles, or their approximate representations[140, 141].

3.4.2 FEATURE SELECTION – DIMENSIONALITY REDUCTION

The feature selection step is used to identify features which are significantly correlated to the process outcome. A large number of feature selection approaches have been developed that can be categorized into filter and wrapper approaches. These methods are useful for constructing models to predict the process outcome (discussed in the following section). Filter methods select relevant features independently of the data mining step. For example, features that discriminate process runs from two or more outcome-derived classes can be identified using hypothesis testing methods, such as a *t*-test (e.g., selection of genes for expression-based tumor classification[142]). Huang *et al.*[137] and Kamimura *et al.*[138] used filter approaches that were based on hypothesis testing to select relevant features. In contrast, wrappers are iterative approaches, where feature selection relies on the results of the subsequent data mining step. Thus, for example, a subset of features is selected and its suitability is evaluated from the error rate of the predictive classifier learned from that subset. Approaches in which features are progressively added (forward selection) or removed (backward elimination) can be applied for the selection of an optimal feature subset. However, these approaches are computationally expensive and potentially suboptimal for large datasets. Alternatively, change in an objective function upon addition or removal of a feature can also be used as a feature selection strategy. Wrapper approaches based on decision trees have been previously used to identify the key parameters that allow one to differentiate process runs into high and low productivity classes[133, 143-145]. These studies identified specific time points, or time windows, during which one or more features could discriminate between runs in different outcome classes.

Due to the temporal nature of process data, feature selection methods must take into account the sequence of events. To this end, statistical methods can be used to assess the significance of a feature, i.e., to assess its ability to distinguish the process runs from different classes. In bioinformatics applications, several hypothesis testing methods have been proposed with the aim of identifying genes that are temporally differentially expressed between two or more phenotypes[115, 146, 147]. Such methods can also be used to evaluate the relative importance of temporal process features in discriminating runs from different groups.

The temporal profiles of some features within individual runs may be correlated. For example, oxygen uptake rate and cell density are often correlated, at least in the exponential growth stage of the culture. Hence, such features provide information that is often redundant. Dimensionality reduction techniques are commonly used to obtain a set of features independent from each other using methods, such as principle component analysis (PCA)[148]. PCA determines the linear correlation structure of multivariate process data as a set of patterns, called principal components (PCs). The first few PCs, which highlight the most dominant correlation patterns among the process parameters, are typically used for dimensionality reduction. The profile of any temporal parameter can be regenerated as a weighted, linear combination of the PCs. Non-negative matrix factorization (NMF)[149] is another dimensionality reduction method used to identify linear correlations between process parameters. Kamimura *et al.*[150] used a PCA-based approach to approximate multiple time-dependent process features of each run as a single temporal pattern, the so-called first principal component (PC1). This reduced feature was subsequently used to cluster process runs into different groups, which corroborated with their known classes.

3.4.3 DATA MINING

Data mining approaches can be broadly categorized as either descriptive or predictive. Descriptive approaches aim to discover patterns that characterize the data, whereas predictive approaches aim to construct models (e.g., functions) to predict the outcome of a future run by learning from the observed parameters.

3.4.3.1 Descriptive Approaches

The descriptive approaches fall into two categories: discovering interesting patterns in the data and clustering the data into meaningful groups.

3.4.3.1.1 Pattern discovery

Algorithms for finding patterns in very large datasets have been one of the key success stories of data mining research. These methods aim to analyze the features of various runs to identify a pattern that is observed in a large number of runs. A pattern can correspond to specific values of a subset of features or a specific temporal profile of a particular feature. Any pattern must occur frequently across different process runs to be considered statistically significant and interesting[151, 152]. Patterns discovered from process data can provide insights into the relationship between different features, and can also be used to discover association rules. For example, specific (on a per cell basis) glucose consumption and lactate production rates of Chinese hamster ovary cells may vary under different growth conditions. However, a switch from lactate production to lactate consumption occurs only within a small window of low specific glucose consumption rate (feature 1) and low specific growth rate (feature 2). Analyzing process data from a large number of runs can reveal the values of the specific rates at which this metabolic change is likely to occur.

3.4.3.1.2 Clustering[153]

Clustering methods can be used to group different process runs into subsets (groups) of runs according to the similarity in the behavior of some features. For example, in some process runs the time profiles of cell density and metabolite concentrations are more similar to one another than in the remaining runs being studied and these can be clustered into one group. Clustering can thus provide insights into different types of runs. In addition, by using various cluster visualization tools (e.g., Spotfire[154]), these methods can also identify those features that distinguishes the clusters. Clustering tools are extensively used in the analysis of large scale gene expression datasets[155]. For example, use of hierarchical clustering to group gene expression profiles of several prostate cancer and normal prostate samples identified clinically relevant tumor subtypes that could be correlated with increased disease recurrence[156].

Clustering methods can be differentiated along multiple dimensions, one of them being the top-down (partitional) or bottom-up (agglomerative) nature of the algorithm. Partitional methods initiate with all process runs (or object/record) belonging to one cluster and they are divided into designated number of clusters. *K*-means, partitioning around medoids (PAM), self-organizing maps (SOM), and graph-based clustering methods are popular examples of partitional algorithms. In contrast, agglomerative methods start with each run belonging to a separate cluster and the clusters are merged, based on the similarities of their feature profiles, until the runs have been

grouped into a pre-specified number of clusters. Hierarchical agglomerative clustering is the most commonly used agglomerative method.

A critical element of clustering methods is the approach used to estimate the similarity between any two runs based on their set of temporal features. To account for the heterogeneity of the temporal features associated with each run, the similarity between two runs is often assessed in two steps. First, the similarity between the corresponding temporal features of a pair of runs is determined and second, the overall similarity between the runs is established by aggregating the individual feature-wise similarities (Figure 3.3). The feature-wise similarity can be computed using various approaches[157]. The most commonly used are Euclidean distance, cosine similarity, and the Pearson's correlation coefficient. Other measures that are based on information theory, such as mutual information can also be used[158]. Mutual information estimates the general dependency between the profiles of two (or more) features, but can only be used for features that have discrete values (e.g., a SAX-represented profile). Note that these methods for assessing similarity can be applied for comparing the same feature across different runs (for pattern recognition), as well as comparing different features of the same run (for dimensionality reduction).

3.4.3.2 Predictive Approaches

Predictive approaches can be used to analyze a set of process runs that exhibit different outcomes (e.g., final product concentration) in order to identify the relationship between process features and the outcome. The discovered relationships (called model or classifier) can be used to predict the process outcome and provide key insights into how the predicted outcome might affect other features of the run, thereby allowing for an intelligent outcome-driven refinement of the process parameters. Commonly used predictive methods include regression, decision trees (DT), artificial neural networks (ANN), and support vector machines (SVM). These methods have been designed for problems that arise when process runs are divided into discrete classes. Often, the process outcome (such as product titer) is a value within a certain range, rather than a discrete variable (such as high- or low-producing runs). In such cases, one can divide the outcome into a number of classes. Alternatively, regression-based methods can be used that are able to predict an outcome variable which is continuous.

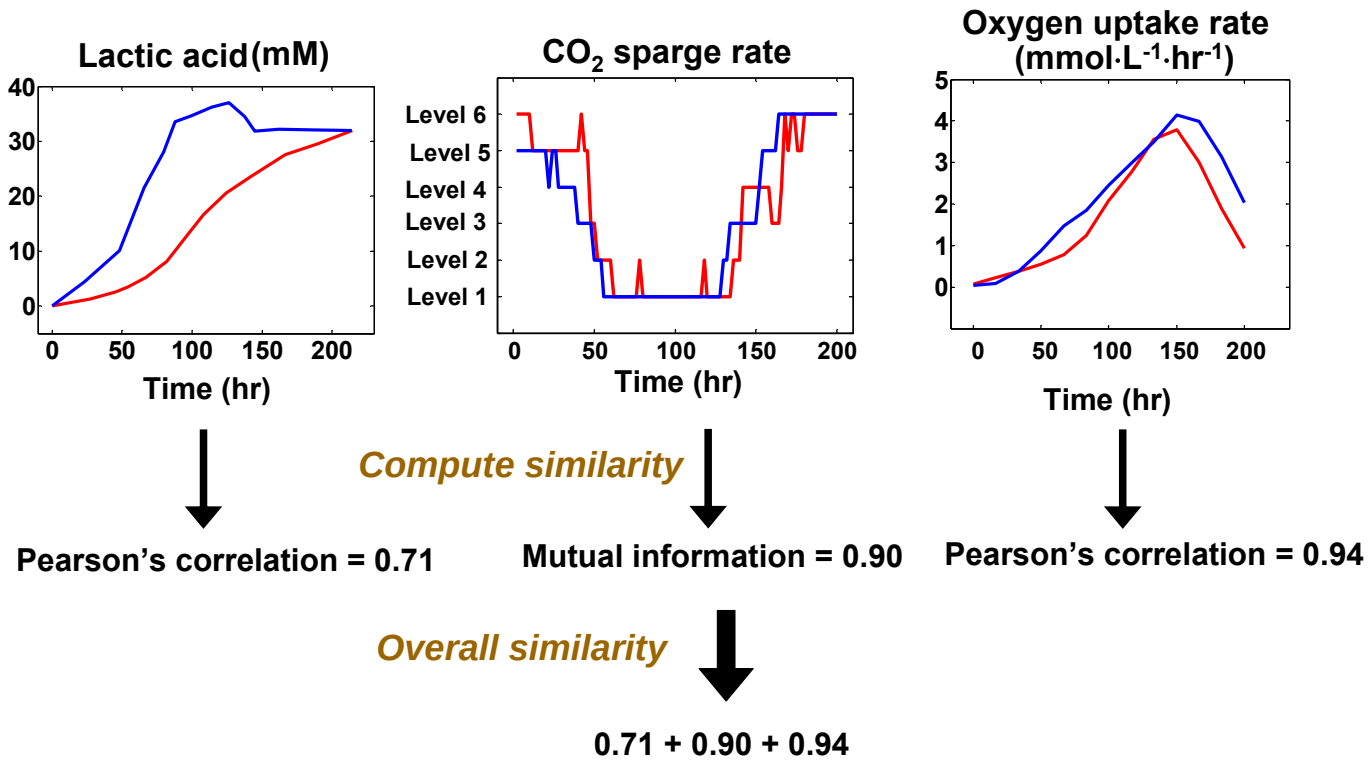

Figure 3.3. An approach to determine the similarity between different process runs. The profiles of different run features, i.e., lactic acid concentration, CO_2 sparge rate, and oxygen uptake rate (OUR) are shown for two runs (in red and blue). The obtained continuous profiles of lactic acid and OUR were compared using a Pearson's correlation[157]. The noisy and long raw profiles of CO_2 sparge rates were discretized into six levels using symbolic aggregate approximation (SAX) method[159]. The levels 1 through 6 represent increasing intervals of CO_2 sparge rates. The discrete profiles of CO_2 sparge rates were compared to each other by estimating their mutual information. The overall similarity between the two runs can then be estimated as an aggregate of these similarities. Prior to aggregation, the similarity metrics should be normalized to ensure that they have the same range. When prior knowledge is available, the aggregation of the feature-wise similarities can be done in a weighted fashion to give greater importance to some of the features.

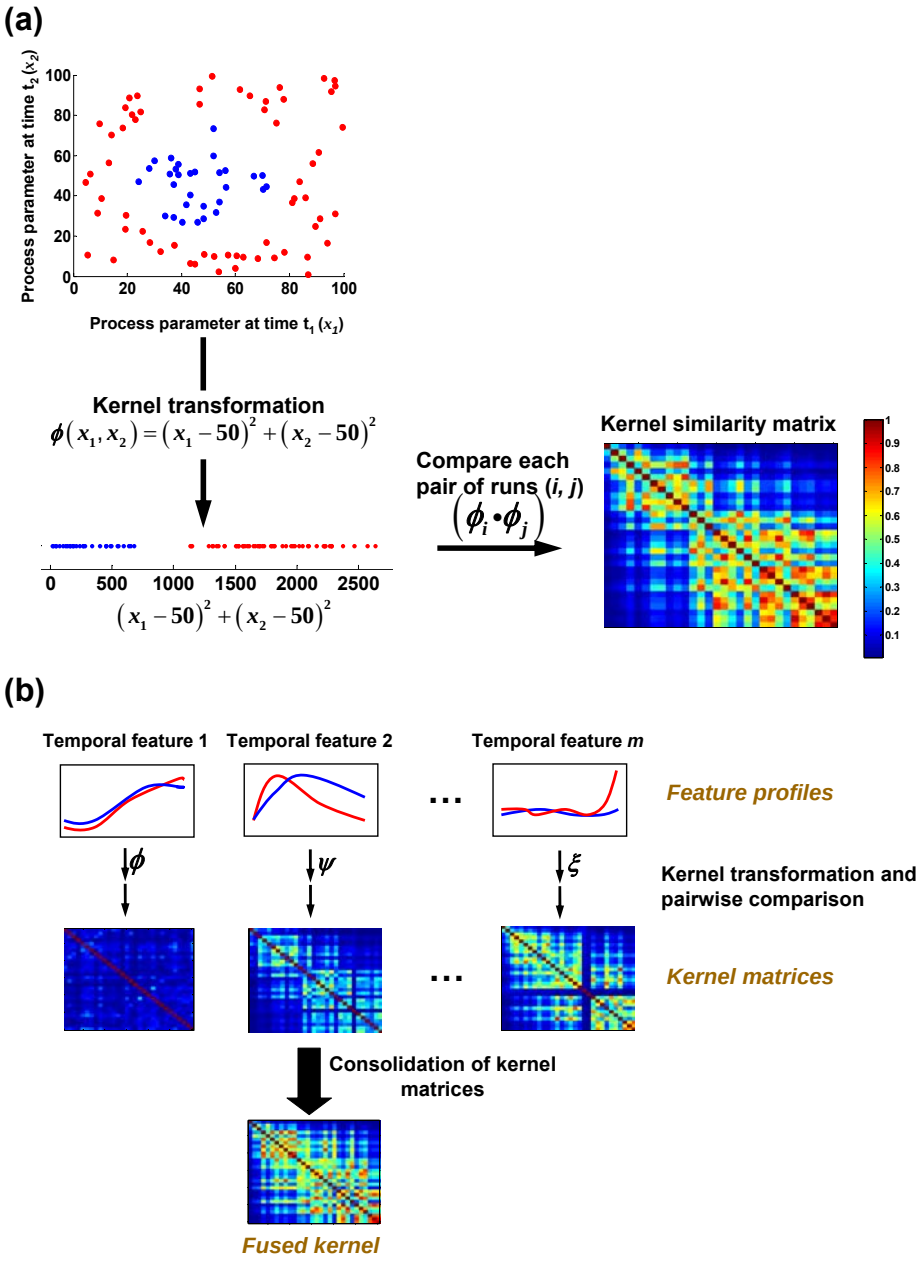

Figure 3.4. A kernel-based learning approach. (a) A simplified scheme of the approach is illustrated. Process data of a single parameter at two different time points is shown for a set of runs categorized into two classes based on process outcome: high (in blue), or low (in red). The distinction between the two classes is immediately obvious after the data have been transformed using a specifically designed kernel function (ϕ) that, for this example, results in a visible 'separation' of the runs. Thereafter, a kernel matrix is obtained by computing the similarity between each pair of run parameters on a scale from dissimilar (0) to identical (1). Note that the diagonal entries in the kernel matrix are 1, i.e. a run is identical to itself. (b) A number of different kernel transformations can be performed to compare different temporal features. The resulting kernel matrices for individual features can then be combined to obtain a fused kernel that can be used for model construction.

Predictive approaches have been extensively used to analyze bioprocess data. Several studies have employed ANNs to predict the output of a fermentation process as a non-linear function of the process inputs[160-163]. ANN models can also be used in conjunction with optimization methods to identify the combination of process inputs that are able to maximize the desired output[144, 164]. Decision trees have also been beneficial for identifying the process trends that allow to discriminate between runs with high and low productivity[133, 145]. For example, a low glucose feed rate was identified as the most discerning process feature for a high productivity run[145]. More recently, a regression method based on partial least squares (PLS) has been used to identify predictive correlations between output parameters and process parameters in order to characterize the process and detect process abnormalities. Furthermore, PLS-based assessment of the similarity of the temporal parameter profiles for process runs at two different reactor scales (2L and 2000L) suggested process comparability at different scales[165].

Recent advances in predictive methods have significantly enhanced their applicability for process data mining. The development of the Vapnik-Chervonenkis theory has laid the foundations of the structural risk minimization (SRM) principle[166, 167], which derives the upper limit on the generalization error of a classifier (i.e., the error incurred by a classifier in predicting the outcome of a new instance (e.g., a new process run)). This upper limit is optimized by classifiers that maximize the separation (called margin) between instances from two (or more) classes. Due to its strong mathematical foundations and intuitive appeal, the idea of maximizing the separation between two groups has gained immense popularity and has been successfully used to improve the predictive robustness of several well-known classification methods, such as ANN[168], *k*-nearest neighbors[169], and regression.

Another major development was the introduction of kernel-based learning that decouples the optimization step in many classification approaches from any data modeling aspects. Kernel-based methods employ a kernel function, which measures the similarity between each pair of runs (Figure 3.4a). A pair-wise comparison of all the runs results in a kernel matrix, which is then used to construct the model. Kernels also provide an elegant solution for addressing the heterogeneity of process data. Multiple kernels can be used, in which each kernel serves to compare one temporal process feature (e.g., oxygen uptake rate, osmolality) over different runs. Kernel functions that quantify linear or non-linear relationships, or even empirically defined functions based on process knowledge and/or historical data can be used to compute the pair-wise

similarities of a particular process feature across different runs. Individual kernels can then be compiled into a 'fused' kernel (Figure 3.4b). Furthermore, the individual features (or their kernels) can be differentially weighted in such a way that the features that are more predictive of the process outcome have higher contribution to the final fused kernel. This step of sorting different features according to their relative importance can be incorporated in the process of model construction. The weights of different features can be 'learned' from the data such that the predictability of the model is maximal[170, 171]. The SRM principle and kernel-based learning also form the basis of support vector machines (SVM), a relatively novel method that has already been widely used to analyze several data-rich applications, such as gene expression analysis[172], text classification[173], and image retrieval[174].

3.4.3.2.1 Support vector machines (SVM)[175]

Based on the structural risk minimization principle, SVMs learn a 'decision boundary' that maximizes the separation between runs from the two groups. The training runs that are closest to the decision boundary and hence most informative are called support vectors. The decision function is calculated based on these support vectors only; the runs distant from the boundary are ignored. The formulation of a linear SVM problem for binary classification of N runs is described below [166, 176].

The i^{th} run is described by a feature vector $\mathbf{x_i}$ $\left(\mathbf{x_i} \in \mathbb{R}^{\boldsymbol{d}}\right)$, which comprises the information about all the parameter profiles for that run. The outcome of the run is discretized as $\boldsymbol{y_i} \in \{+\mathbf{1}, -\mathbf{1}\}$ (high (+1) or low (-1)). SVM identifies a decision boundary $\left(\mathbf{w} \cdot \mathbf{x} + \boldsymbol{b} = \mathbf{0}\right)$ that discriminates the runs in the feature space with maximum margin, i.e., the distance between the two parallel hyperplanes on either side of, and parallel to, the separating hyperplane. These two hyperplanes 'touch' the runs (called support vectors) from the two classes that are nearest to the decision boundary (Figure 3.5). A soft-margin approach, where violations $\left(\boldsymbol{\xi_i}\right)$ of the decision boundary are penalized with a cost function $\left(C\right)$, generally provides a more robust solution. The constrained optimization problem is formulated as:

$$\min_{\mathbf{w},\boldsymbol{b}} \left(\frac{\|\mathbf{w}\|^2}{\mathbf{2}} + \boldsymbol{C} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{\xi_i} \right) \text{ such that } \boldsymbol{y_i}\left(\mathbf{w} \cdot \mathbf{x} + \boldsymbol{b}\right) \geq \mathbf{1} - \boldsymbol{\xi_i} \;;\quad \boldsymbol{\xi_i} \geq \mathbf{0} \quad \forall \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \ldots, \boldsymbol{N}$$

The cost function (C) and the slack variable (ξ_i) provide a trade-off between the violations of the decision boundary and the width of the margin. Quadratic programming techniques are typically used to solve for $\mathbf{w}$ and b [177]. For datasets that cannot be separated with linear models, kernel functions can be employed to project the input data onto higher dimensional feature vectors. Recent one-class[178] and multi-class[179] extensions of SVMs have considerably broadened their applications.

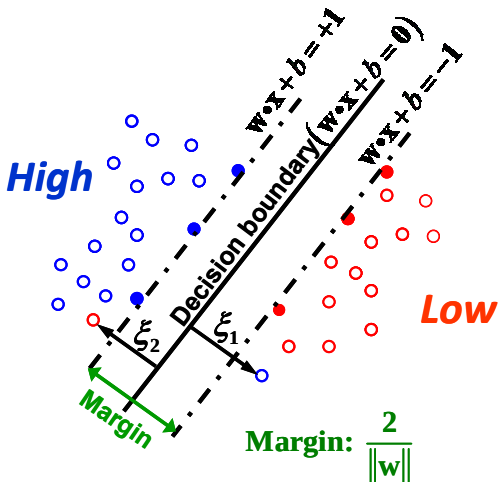

Figure 3.5. Maximum margin support vector classification. Several runs categorized as high (blue circles) and low (red circles) are shown in a high-dimensional feature space $\mathbb{R}^d$. The support vectors, i.e. runs closest to the decision boundary, are depicted as solid circles – high (solid blue), low (solid red). The decision boundary maximizes the margin between parallel hyperplanes (shown as dash dotted lines) that maximally separate runs from the two classes. Two violations of the decision boundary are also shown (ξ_1, ξ_2) in this illustration.

3.4.4 Model Validation and Interpretation

Discovery of a model pattern or trend must be followed by subsequent evaluation and expert interpretation. In descriptive methods, it is important to examine whether a pattern or a cluster represents a genuine relationship between the performances of different process runs or is simply the outcome of a spurious trend. In addition, noise in process measurements can obscure the interpretation of a discovered pattern. Furthermore, many clustering algorithms are designed to find a set of clusters that are only locally optimized. For example, the initial assignment of the runs to clusters (which is often random) may have an effect on the final clustering, and different

initial assignments may lead to different groupings of the runs. Resampling-based approaches have been proposed to evaluate the reproducibility of a set of clusters[180, 181]. In these procedures, a subset of runs can be sampled from the original dataset and clustering performed. This process is repeated multiple times and the agreement of the resulting clusters is compared across all the subsets and is used to assign a confidence term for the clustering.

Predictive methods run the risk of constructing an 'overfitted' model. Datasets where the number of process features is much higher than the number of runs used for model construction are particularly vulnerable to overfitting. To avoid this, it is essential to assess the predictive ability of a model for new runs. A subset of runs (training set) is used for model construction and the remaining runs (test set) are used for model evaluation. Error rates are calculated based on the number of runs initially misclassified by the model. For datasets with finite or few runs, cross-validation and resampling schemes (e.g. bootstrap) can be used, where the dataset is divided into multiple training and test subsets to obtain an average estimate of the error[182].

The introduction of a 'selection bias' is another relevant issue for generating models based on a subset of features (selected from the entire feature set). This bias is introduced if all runs (including test set runs) are involved in the feature selection process, and the test set is used merely to validate the model build on the pre-selected features. Both feature selection and model construction must be implemented on the training subset only, without any input from the test set[183].

3.5 CONCLUDING REMARKS

Modern production plants are equipped with sophisticated control systems to ensure high consistency and robustness of production. Nevertheless, fluctuations in process performance invariably occur. Understanding the cause of these fluctuations can greatly enhance process outcome and help to achieve higher performance levels. Given the vast amount of archived process data in a typical modern production plant, the opportunities for unveiling any hidden patterns within the data and recognizing the key characteristics for process enhancement are enormous. The ultimate aim of mining bioprocess data is to gain insights for process advancement or even process innovation. Interpretation by process experts is essential to relate the discovered patterns to cellular physiology, which in turn can generate hypotheses for experimental verification. In a bioreactor operation, ultimately it is the physiological state of the

cells that determines the process outcome. The benefits to be gained from mining bioprocess data will be immense. These opportunities are met with major advances in data mining tools that have become available in the past decade. The application of these tools to explore bioprocess data will be highly rewarding in the near future.

CHAPTER 4 TRANSCRIPTOME DYNAMICS-BASED OPERON PREDICTION AND VERIFICATION IN *STREPTOMYCES COELICOLOR*

4.1 SUMMARY

Streptomyces spp. produce a variety of valuable secondary metabolites, which are regulated in a spatio-temporal manner by a complex network of inter-connected gene products. Using a compilation of genome-scale temporal transcriptome data for the model organism, *Streptomyces coelicolor*, under different environmental and genetic perturbations, we have developed a supervised machine-learning method for operon prediction in this microorganism. We demonstrate that, using features dependent on transcriptome dynamics and genome sequence, a support vector machines-based classification algorithm can accurately classify greater than 90% of gene pairs in a set of known operons. Based on model predictions for the entire genome, we verified the co-transcription of more than 250 gene pairs by RT-PCR. These results vastly increase the database of known operons in *S. coelicolor* and provide valuable information for exploring gene function and regulation to harness the potential of this differentiating microorganism for synthesis of natural products.

4.2 INTRODUCTION

Transcriptional regulation is perhaps the most fundamental control in gene expression. Many functionally related genes are often co-regulated, meaning that their expression is coordinated temporally or even spatially in response to the need of the organism in a given environmental condition. In prokaryotes these co-regulated genes are often organized in their genomes into physical clusters called operons. An operon thus consists of more than one adjacent gene expressed as a transcription unit. Operons allow an organism to simultaneously express the genes that are needed for cell survival under the same condition, providing a control circuit that is both simple and economical. In some cases, however, there is also a need to fine tune the expression of individual genes in an operon under some circumstances. This is accomplished by

alternative regulation of genes, which are normally co-regulated in one operon[184]. Transcription of a unit encoding a single gene or an operon is controlled by a promoter and a terminator. Alternative regulation in an operon is accomplished by one or more alternative promoters or internal transcription terminator.

In the past few years numerous bacterial genomes have been completely sequenced and the number is steadily increasing. Identifying potential operons in those genomes facilitates the functional annotation of the genes involved and is important in elucidating the regulation of those genes. Several approaches have been previously used for operon predictions. Most methods rely on features based on the genome structure or the functional similarity of genes of interest. Since adjacent genes in an operon often are physically closer to each other than those not in the same operon, intergenic distance provides information about the likelihood that two adjacent genes may be on the same operon[185]. The conservation of gene order in multiple organisms is also taken into account[186]. Additionally, the similarity of codon usage (the frequency with which synonymous codons encode amino acids in neighboring genes) is also used for operon predictions [187]. Since genes on the same operon are co-regulated, at least under the conditions when alternative regulation is not in play, their transcription profiles are likely to be well correlated. Identifying adjacent genes whose transcription levels are well correlated also provides much information on the likelihood of their being in the same operon[188].

Unsupervised Bayesian methods using features based on genome sequence and functional similarity have been reported for operon prediction in all sequenced prokaryotes[186, 187, 189]. An empirical scoring method has also been reported previously[190]. Since these methods do not require a training set, they are advantageous for organisms where little or no information about known operons is available. Alternatively, machine learning approaches have also been used to train models based on databases of known operons. Studies have shown that log-likelihoods derived from distribution of intergenic distance in a set of known operons can be used for operon prediction in several prokaryotes[185, 191]. Naïve Bayesian classifier as well as C5.0, a decision tree-based algorithm, have been reported for predicting operons in *Escherichia coli*[192, 193]. A support vector machine-based model has recently been reported for operon prediction in *Escherichia coli* and *Bacillus subtilis*[194]. Few reported methods have combined transcriptome data and genome sequence for predicting operon structure. A hidden Markov model based on expression data alone has been reported for *E. coli*[195]. Bayesian methods that combine similarity of transcript profiles

with information based on genome sequence have been previously used for operon prediction in *E. coli* and *B. subtilis*[188, 196, 197].

In this study, we employed genome-wide temporal transcriptome data from several strains and culture conditions, information about intergenic distance and transcription terminator predictions, and applied a support vector machines (SVM)-based model for operon prediction in the entire genome of *Streptomyces coelicolor*. The model predicts more than 2000 gene pairs as being co-transcribed, of which 250 were subsequently experimentally verified.

4.3 MATERIALS AND METHODS

4.3.1 MICROARRAY DATA

4.3.1.1 Strains and culture conditions

S. coelicolor A3(2) strain M145 (prototroph, SCP1$^-$, SCP2$^-$) and mutant strains of two regulatory genes were used – YSK3225 (M145 Δ*absA1::apr*), and YSK4425 (M145 Δ*afsS::apr*). The strains were grown in batch culture in liquid medium as described elsewhere[198].

4.3.1.2 Probe preparation and microarray hybridization

Temporal transcriptome profiling was performed using a whole genome DNA microarray of *S. coelicolor* that has probes for 7579 genes[198]. Cell samples were taken at different time points along the culture for transcriptome profiling. RNA extraction, cDNA synthesis, microarrary hybridization, washing, scanning and image analysis was performed as described elsewhere[198]. Genomic DNA (gDNA) was used as a common reference for all the hybridizations. Details for all the protocols are available at http://hugroup.cems.umn.edu/Protocols/protocol.htm.

4.3.1.3 Microarray data compilation and processing

The time series microarray data comprise 67 cell samples from three different strains – 27 samples from wild-type (M145) and 40 samples from two mutant strains (YSK3225, YSK4425). The data was arranged as 3 sets (set 1-3) as shown in Table 4.1. All hybridizations were performed using genomic DNA as a reference (cDNA:gDNA). The data was normalized by quantile normalization method, which assumes that the overall distribution of total mRNA is the same for different RNA samples[198, 199].

Transcriptome data publicly available in the Stanford Microarray Database (SMD), comprising time series experiments reported by Karoonuthaisiri *et al.*[88] on two *S. coelicolor* strains – M145 and M600, under different stress conditions was also compiled. In these

experiments hybridization was performed by pairing two cDNA samples (cDNA:cDNA) with one being t = 0 hr cDNA sample, used as a reference in most cases. The data from 61 samples was arranged as three different sets (set 4-6) depending on the type of experiment, strain, and growth medium used (Table 4.1). Additionally temporal transcriptome data reported by Huang *et al.* [200] on *S. coelicolor* A3(2) M145, J1501 and several mutant strains was also compiled. This dataset, which included 48 cDNA:cDNA measurements and 30 cDNA:gDNA measurements was arranged as three sets (set 7-9) as shown in Table 4.1. The experiments with genomic DNA as reference were quantile normalized. Several genes in each array sample were flagged 'absent' due to low intensity, small spot diameter, or low spot regression coefficient. Samples with greater than 25% genes flagged, were discarded before further processing.

Similarity between the transcript levels of genes in every pair was calculated by the Pearson correlation coefficient (r). For calculation of Shannon entropy, the samples with cDNA:gDNA measurements were standardized by dividing the cDNA/gDNA ratio for every gene by the cDNA/gDNA ratio of that gene in the first time point of M145 in set 1.

4.3.2 GENOME ORGANIZATION

The genome sequence of *S. coelicolor* and the annotation files were obtained from The Sanger Institute (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/). The leading and lagging strands were scanned and pairs of genes were grouped based on whether they were transcribed in the same directions (same-strand gene pairs) or in different directions (opposite-strand gene pairs). The 7825 genes in the linear chromosome were binned into 4965 same-strand pairs and 2859 opposite-strand pairs as shown in Figure 4.1.

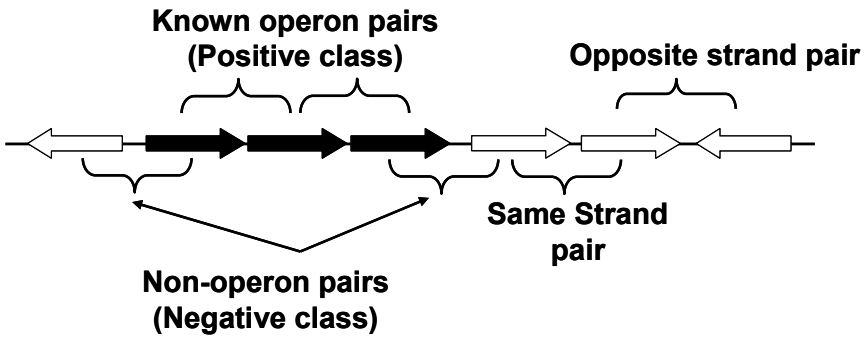

Figure 4.1. Definition of known operon pairs (KOPs), non-operon pairs (NOPs), same-strand pairs, and opposite-strand pairs. Closed-block arrows indicate genes in a known operon. Open-block arrows represent genes with unknown operon status

Table 4.1. Summary of Microarray data used for operon predictions

Set	Strain	Description	Array[a]	Growth medium	No. of samples	Reference
1	M145	Kinetics of growth, biological triplicates	cDNA: gDNA	R5$^-$, liquid	27	This study
2	YSK3225	Kinetics of growth, biological duplicates	cDNA: gDNA	R5$^-$, liquid	18	This study
3	YSK4425	Kinetics of growth, biological duplicates	cDNA: gDNA	R5$^-$, liquid	22	This study
4	M600, M145	Kinetics of growth	cDNA: cDNA	R5$^-$, liquid	17	[88]
5	M600	Response to stress (temperature upshift, ethanol upshift, sucrose downshift)	cDNA: cDNA	R5$^-$, liquid	19	[88]
6	M600	Response to stress (temperature upshift, temperature downshift, phosphate upshift)	cDNA: cDNA	SMM, liquid	25	[88]
7	J152, J153, J154, J155	Kinetics of growth	cDNA: cDNA	R5$^-$, solid	24	[200]
8	C122, C123, C124, C125	Kinetics of growth	cDNA: cDNA	R5$^-$, solid	24	[200]
9	J151, C121, M145, M512, M550	Kinetics of growth	cDNA: gDNA	R5$^-$, solid	30	[200]

(a) cDNA:cDNA indicates that cDNA from two RNA samples was used in the assay. cDNA:gDNA indicates that genomic DNA (gDNA) from the parent strain was used as the reference sample in the assay

4.3.2.1 Intergenic distance calculation

Intergenic distance in base pairs between the genes in every gene pair (gene *I* – gene *II*) was calculated as $\mathbf{distance}_{I-II} = \mathbf{gene}_{II}\_start - \mathbf{gene}_{I}\_end - 1$. Negative intergenic distance implies an overlap of the translated region of the two genes.

4.3.3 PREDICTION OF TRANSCRIPTION TERMINATORS

The presence of rho-independent transcription terminator in the intergenic region of every gene pair was predicted by the TransTerm algorithm[201]. The algorithm searches for mRNA motifs that potentially form a hairpin structure and are followed by a short uracil-rich region both within and between the genes. The stability of the hairpin structure and the presence of the U-rich region are characterized by a score that is used to estimate a confidence score/probability of the presence of terminator at a particular site in the genome. Using a confidence level of 0.9 that has been reported to identify 95% of known terminators in *E. coli*[201], we searched all the gene pairs in *S. coelicolor* for which the probability of the presence of terminator in the intergenic region is 0.9 or higher.

4.3.4 EXPERIMENTAL VERIFICATION OF OPERONS

4.3.4.1 Culture condition, RNA extraction and cDNA synthesis

S. coelicolor M145 wild-type spores were grown in batch culture in modified R5 liquid medium [39], as described elsewhere[198], and samples were withdrawn periodically for RNA extractions. The mycelia was fragmented in liquid nitrogen using mortar and pestle and total RNA was extracted using RNeasy Mini Kit (Qiagen, Valencia, CA) according to the manufacturer's protocol. Residual genomic DNA was digested using Turbo DNA-free™ kit (Ambion, Austin, TX) according to the protocol suggested by the manufacture for rigorous DNase treatment. Total RNA was suspended in 50 µl of nuclease-free water and stored at -80°C until further use.

Equal amounts of RNA from four samples corresponding to exponential, late exponential, transition, and stationary phase, were pooled and reverse transcribed using random hexamers and Superscript™ III (Invitrogen, Carlsbad, CA) at 50°C for 1 hr according to manufacturer's protocol. 50 ng of random hexamer was used for every 5 µg of total RNA. A negative control was also done without the addition of the reverse transcriptase enzyme. Thereafter, the RNA was

digested by addition of RNase H (Invitrogen) and incubation at 37°C for 20 min. cDNA was stored at -20°C until further use.

4.3.4.2 PCR

Gene-specific primers used for whole genome microarray construction[198], were used for RT-PCR based verification of transcripts. To confirm that a pair of adjacent genes is on the same mRNA transcript, the 5' primer of the first gene and the 3' primer of the second gene were combined to form a primer pair at a working concentration of 5 μM for each primer. The length of the amplicon for this primer pair was obtained from the chromosomal location of the primers, obtained by blasting the primer sequences against a database of *S. coelicolor* genome.

cDNA from 100 ng of pooled RNA was used as template for every PCR reaction. PCR was also performed on an equivalent amount of negative control from cDNA synthesis to check for any residual genomic DNA contamination in the RNA samples. The PCR conditions were as follows: 5 min of initial denaturation at 95°C, 40 cycles of amplification - denaturation for 30 sec at 94°C, annealing for 30 sec at a temperature between 60 – 64°C depending on the melting temperature of the primers, and extension at 72°C for 150 seconds. The final extension was done at 72°C for 5 min. The total reaction volume was 50 μl and 20 μl was analyzed on 1% (w/v) agarose gel.

4.3.5 SUPERVISED CLASSIFICATION

4.3.5.1 Support vector machines (SVM)

SVMs are a class of kernel-based machine learning methods that use the principle of structural risk minimization to identify a decision function that separates objects from two classes with maximum margin[166, 176]. SVM*light*, an implementation of SVMs in C was used for model training and evaluation[202]. Two of the various kernel functions, linear and radial basis function (RBF), were used for classification. A linear kernel (k) measures the similarity between two training objects ($\mathbf{x}_1$ and $\mathbf{x}_2$) as a dot product in the input feature space, $k\left(\mathbf{x_1}, \mathbf{x_2}\right) = \mathbf{x_1} \bullet \mathbf{x}_2^{/}$. The radial kernel function transforms the data using the non-linear function, $k\left(\mathbf{x_1}, \mathbf{x_2}\right) = \exp\left(-\gamma \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2\right)$, where γ determines RBF width. For radial kernel function, the parameters γ (`-g`) and the cost function (`-c`) were selected using the leave-one-out model selection (*looms*) procedure[203]. The algorithm calculates the leave-one-out error rates for a range of parameters and outputs the one with minimum error rates.

4.3.5.2 Training set – Positive and negative classes

The training set consists of 49 known operons compiled from literature. An additional 6 known operons in *S. lividans* were also included in the training set. *S. lividans* and *S. coelicolor* are close relatives with 99.6% similarity in their 16s rRNA sequences [204], and common structural and genetic organization[205]. To increase the number of known operons in the training set, an additional 8 operons reported in other *Streptomyces* spp. (*S. griseus, S. antibioticus, S. ambofaciens, S. ramocissimus, S. thermoviolaceus,* and *Streptomyces* sp. NRRL 5331) with a conserved gene order in *S. coelicolor* were also included in the training set. 27, 12, 17, 2, and 5 of the known operons have 2, 3, 4, 5, and 6 or more genes, respectively.

The gene pairs formed by consecutive genes in the known operons were referred to as *known operon pairs* (KOPs), as shown in Figure 4.1. The resulting 149 KOPs constitute the positive class of the training set. The set of gene pairs that comprise the negative class was created as follows. The first gene of every known operon and the gene immediately upstream, as well as the last gene in every known operon and the gene immediately downstream form *non-operon pairs* (NOPs) (Figure 4.1). The resulting set of 122 NOPs constitute the negative class. Nine of the known operons have internal regulation with one or more internal promoters or a transcriptional terminator. For these operons, the pair of genes on either side of the internal control element was not considered as a KOP.

4.3.5.3 Model training and selection

Binary SVM classifiers were trained for operon prediction using three different features – intergenic distance, correlation of transcript profiles, and transcription terminator predictions. Intergenic distance is measured in base pairs and varies from -26 to 811 bp in the training set, whereas Pearson correlation coefficient is bound between -1 and 1. Due to the large difference in the range of these features, scaling was performed by discretizing the intergenic distances into 7 bins corresponding to $d \leq 0$, $0 < d \leq 20$, $20 < d \leq 50$, $50 < d \leq 100$, $100 < d \leq 200$, $200 < d \leq 300$, and $d > 300$ bp.

The discrimination rule established during training can result in *overfitting* whereby the classifier cannot accurately discriminate test/unseen data. Leave-one-out and k-fold cross-validation was thus performed to estimate the performance of the model in classifying an independent dataset that was not used for training (i.e., assess its generalizability)[206].

4.3.5.3.1 Leave-one-out approach

Leave-one-out cross-validation is an iterative approach where each gene pair in the training set of '*n*' gene pairs is left out in one iteration. The model is trained with (*n-1*) gene pairs and used to classify the n^{th} gene pair. In each iteration, the true class of the pair (whether it is a KOP or NOP) is compared with the predicted class. The performance of the model is then evaluated using different metrics.

4.3.5.3.2 Evaluation Metrics

The following metrics were used to compare the performance of different classifiers.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}}$$

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}}$$

$$\textbf{False positive rate (FPR)} = \frac{\textbf{FP}}{\textbf{FP+TN}}$$

$$\textbf{Total error rate} = \frac{\textbf{FP+FN}}{\textbf{TP+FN+TN+FP}}$$

$$\textbf{F-factor} = \frac{2 \times \textbf{Recall} \times \textbf{Precision}}{\textbf{Recall + Precision}}$$

where, TP (true positives) = Number of KOPs accurately classified as operon pairs by the model

FN (false negatives) = Number of KOPs falsely classified as non-operon pairs by the model

FP (false positives) = Number of NOPs falsely classified as operon pairs

TN (true negatives) = Number of NOPs accurately classified as non-operon pairs

Recall quantifies the sensitivity of the model – how many KOPs can be predicted as operon pairs by the model and precision quantifies the specificity of the model – how many of the operon pairs predicted from the training set (KOPs and NOPs) are KOPs. F-factor combines the two metrics to quantify the overall performance of the model. The F-factor can range from 0 to 1 with 1 corresponding to an ideal classifier.

4.3.5.3.3 K-fold cross-validation

A stratified 5-fold cross-validation procedure was implemented to compare the performance of classifiers with different features. In this procedure, the training set was randomly

divided into 5 subsets, where each subset was stratified such that it contains the same proportion of KOPs and NOPs as the original training set. Four subsets were used for training the model which was then used to assign a score (*s*) to every gene pair in the 5th test subset. The procedure was repeated five times. This 5-fold cross-validation was performed five times (5 x 5) and the true class of the gene pairs in each of the 25 test subsets and their scores were then used to generate Receiver Operating Characteristics (ROC) graphs. An ROC curve is a plot of recall as a function of FPR. Using the test subsets, 25 ROC graphs were generated for each classifier. Instead of merging the 25 ROC graphs to one large set and calculating a single ROC curve for each classifier, we used the Vertical averaging procedure described by Fawcett (2004) [207]. This procedure combines the 25 ROC graphs to estimate the average recall and its standard deviation at different FPRs. Briefly, for a fixed value of FPR, each of the 25 ROC graphs are scanned and the maximum recall or true positive rate at that FPR is chosen, using interpolation if necessary. These values are used to compute the average recall and draw confidence intervals (± standard deviation) at the fixed FPR. The FPR can be increased from 0 to 1 in small step sizes to get the average ROC curve. *Area under ROC curve* (AUC) was used as a scalar measure for comparing the performance of different classifiers – the AUC for a random classifier is 0.5 and that of an ideal classifier is 1.

4.4 RESULTS

4.4.1 KNOWN OPERON PAIRS HAVE SHORTER INTERGENIC DISTANCE

As described in Materials and Methods, from a set of known operons we obtained 149 known operons pairs (KOPs) and 122 non-operon pairs (NOPs). The density distribution of the intergenic distances in KOPs and NOPs is shown in Figure 4.2. For KOPs, the distribution has a sharp peak around intergenic distance of 0 bp. Sixty-seven (45%) KOPs have an intergenic distance less than 0 bp indicative of a translational overlap between the genes. Fifty-seven of these gene pairs have an overlap of 4 bp. Among them 35 have ATGA as the overlapping sequence, where ATG corresponds to start codon for the second gene and TGA is the stop codon for the first gene. The overlapping sequence in other 22 pairs is GTGA. Since *S. coelicolor* has 72% GC content, GTG is also a commonly observed translational start codon. An overlap of 1 bp between the start and the stop codons of adjacent genes was also observed among 5 of the KOPs. In contrast, only six (5%) NOPs have an overlap in the intergenic distance.

Although a short intergenic distance is a strong indication of co-transcription, a significant fraction of genes in KOPs are separated by intermediate to large intergenic distance. In the training set, 33 (22%) and 21 (14%) KOPs have an intergenic distance that is greater than 50 bp and 100 bp, respectively. If only intergenic distance is used for operon prediction based on this training set, using a distance threshold of 50 bp, 116 (78%) KOPs can be classified accurately. However, 19 (16%) NOPs will also be falsely classified as being co-transcribed. If the threshold is increased to 100 bp, 128 (86%) KOPs can be correctly classified with a large false positive rate of 32%.

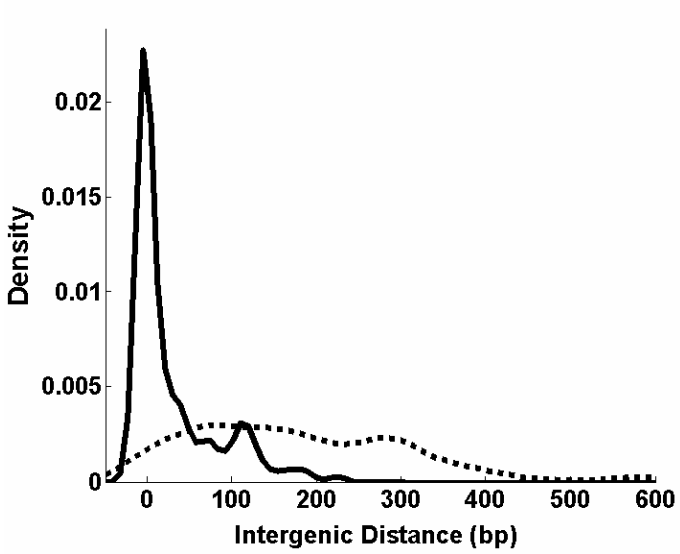

Figure 4.2: Density distribution of intergenic distance in KOPs and NOPs. (——) KOPs; (---) NOPs.

4.4.2 GENES IN KNOWN OPERONS HAVE GREATER EXPRESSION CORRELATION

Genes within an operon are likely to have a higher similarity in their transcript levels compared to genes that are not co-transcribed. Similarity of transcript profiles can be measured using several metrics such as Euclidean distance, cosine function, and Pearson correlation coefficient (r). Among these metrics, it has been previously observed that Pearson correlation achieves the best separation between KOPs and NOPs[196].

Temporal transcriptome data obtained from 206 cell samples were divided into nine different sets depending on the experimental design, strains and culture conditions used (Table 4.1). For every KOP, the Pearson correlation between the transcript levels of the adjacent genes was calculated for each of the nine sets, and the number of sets in which the correlation exceeds 0.7 was counted. The KOPs were divided into 10 groups according to the number of sets

(0,1,2,…,9) in which transcript correlation exceeds 0.7. Figure 4.3a shows the distribution of the KOPs in different groups. Only one out of 149 KOPs has transcript correlation $r > 0.7$ in all the nine sets. The error in measurement of transcript level due to noise, may have contributed to the relatively low correlation between genes in KOPs. The presence of as yet-unidentified site for internal regulation (internal promoter or transcription terminator), or differential mRNA degradation could also potentially reduce the similarity in transcript level of genes in a KOP. Nonetheless, 58 (39%) KOPs have transcript correlation $r > 0.7$ in four or more sets. In contrast, only six (5%) NOPs have transcript correlation $r > 0.7$ in four or more sets (Figure 4.3b). Further, 78 (64%) NOPs do not satisfy the correlation threshold of 0.7 in any of the nine sets, in contrast to only 18 (12%) KOPs.

The separation between KOPs and NOPs is evident even at higher correlation thresholds. 32 KOPs have transcript correlation $r > 0.8$ in four or more sets in contrast to only one NOP. To confirm that the higher Pearson correlation in KOPs is not by chance, the correlation between the transcript levels of genes in 20000 randomly selected pairs was also calculated for all the nine sets. Only 5% of randomly selected gene pairs have $r > 0.7$ in four or more sets (Figure 4.3c). This indicates that the higher degree of correlation between the transcript levels of genes in KOPs can be used for operon prediction.

4.4.3 TRANSCRIPTION TERMINATORS

Using TransTerm which identifies rho-independent transcription terminators, none of the KOPs were found to have a transcription terminator predicted in the intergenic region with a probability of 90% or higher. In contrast, 16 NOPs have a predicted transcription terminator with 90% or greater likelihood, of which nine have a probability greater than 99%. Thus, the probability of presence of a transcription terminator in the intergenic region of gene pairs can also be used as a discriminatory feature for operon prediction.

4.4.4 BINARY CLASSIFICATION RESULTS

Using support vector machines (SVM) as a supervised classification tool, binary classifiers were designed to discriminate KOPs and NOPs using different combinations of features. As described in Materials and Methods, leave-one-out and *k*-fold cross-validation was used for evaluation and selection of the *best* classifier.

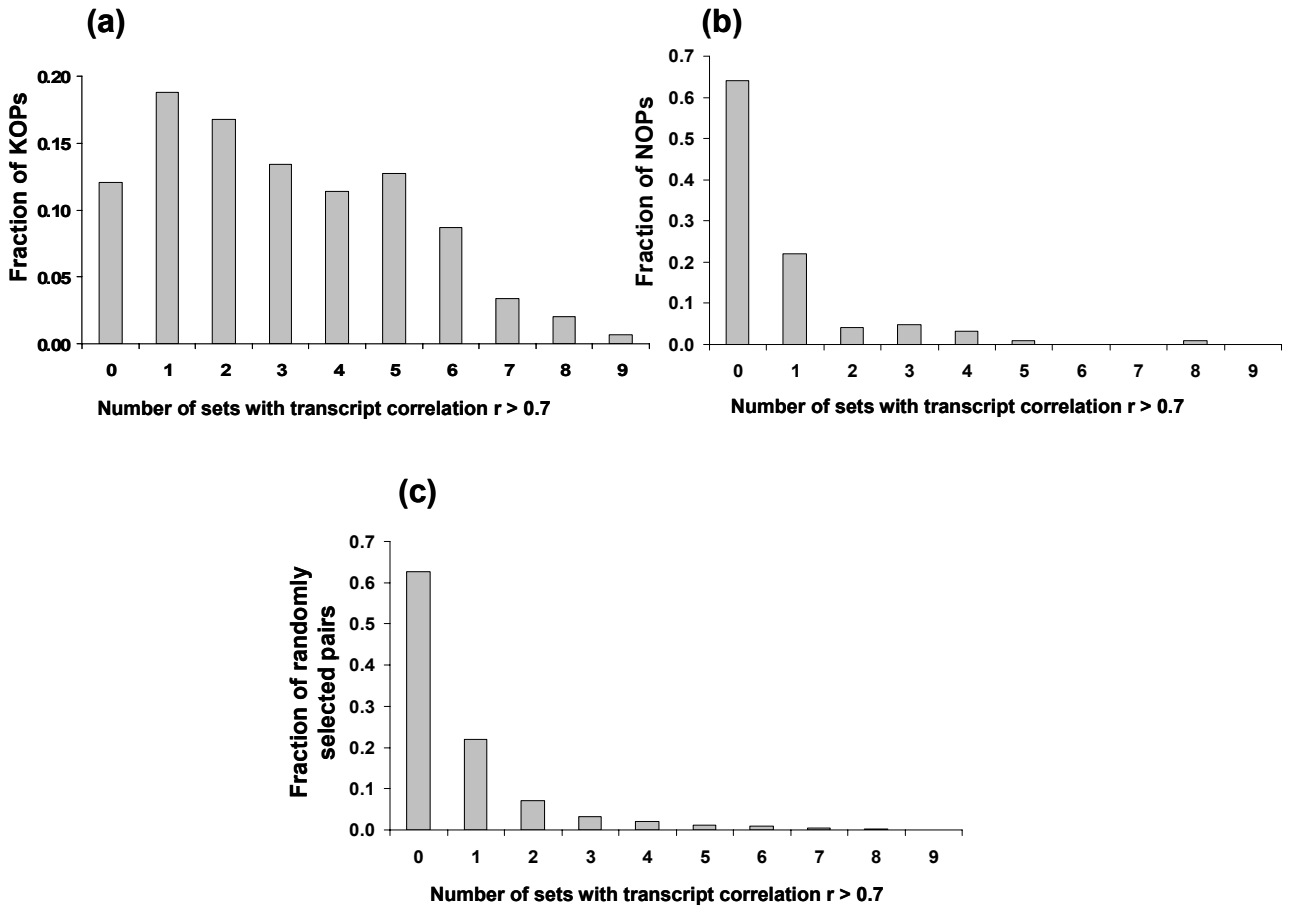

Figure 4.3. Comparison of Pearson correlation between transcript levels of adjacent genes in KOPs and NOPs. The microarray experiments were divided into 9 sets and correlation between transcript levels of adjacent genes in every pair was calculated for each set. The histogram of the number of sets in which correlation exceeds 0.7 in (a) Known operon pairs (KOPs); (b) Non-operon pairs (NOPs); (c) Randomly selected pairs.

4.4.4.1 Leave-one-out cross-validation results

The performance of different classifiers is shown in Table 4.2. If only intergenic distance is used for classification of training set (classifier I), 82% of KOPs can be accurately classified as operon pairs with a precision of 86%. However, 16% of NOPs are misclassified as operon pairs. Discretization of distance (classifiers II and III) results in a small reduction in recall (78%) with comparable precision and false positive rates (FPR). If only transcriptome data is used for classification, a radial SVM model (classifier V) with recall and precision of 80% and 82%, respectively, performs marginally better than linear SVM model (classifier IV). However, with an F-factor of 0.838 the performance of distance-based classifier I is slightly better than the transcriptome-based classifier V (F-factor = 0.810). Terminator predictions alone can differentiate only 16 (13%) NOPs due to the presence of a predicted terminator site in their intergenic region. However, the remaining 87% NOPs cannot be differentiated from KOPs resulting in a large FPR (classifier VI in Table 4.2).

When intergenic distance and transcriptome data are combined, the performance of the linear (classifier VII) as well as the radial SVM classifier (classifier VIII) improves significantly with recall and precision of 90% and 88%, respectively. With an F-factor of 0.89, the classifiers VII and VIII that combine transcriptome data with intergenic distance are better than any of the classifiers that use only one feature (classifier I – VI). The radial model based on all the three features (classifier X) has a marginal improvement in recall (92%) and precision (89%) compared to classifier VII and VIII. Among the various combinations of feature sets and kernel functions, the radial classifier X has the highest recall and precision (Table 4.2).

4.4.4.1.1 Increasing transcriptome data improves prediction accuracy

The performance of a classifier based on transcriptome data is profoundly affected by the diversity of experimental conditions under which microarray experiments are performed. To demonstrate this we trained an SVM classifier based on transcriptome data from the time course experiment of M145 wild-type in $R5^-$ liquid medium only (set 1 in Table 4.2). The classifier has a recall and precision of 60% and 71%, respectively. In contrast, the radial classifier (classifier V) based on all transcriptome data, has a significantly higher recall and prediction of 80% and 82%, respectively. Thus, addition of microarray experiments performed with different strains and culture conditions can improve the accuracy of operon predictions significantly.

Table 4.2. Comparison of different classifiers using leave-one-out cross-validation

Classifier	Kernel function	Feature(s)	Recall (%)	Precision (%)	Total error rate (%)	False positive rate (%)	F-factor
I	Radial ($\gamma = 0.01$)	Distance	82	86	17	16	0.838
II	Linear	Distance (discretized)	78	86	19	16	0.817
III	Radial ($\gamma = 0.0025$)		78	86	19	16	0.817
IV	Linear	Transcriptome	78	82	22	21	0.798
V	Radial ($\gamma = 0.02$)		80	82	21	21	0.810
VI	Linear	Terminator prediction	100	58	39	87	0.734
VII	Linear	Distance and transcriptome	90	88	12	15	0.890
VIII	Radial ($\gamma = 0.25$)		90	88	12	15	0.890
IX	Linear	Distance, transcriptome, and terminator prediction	90	88	12	15	0.887
X	Radial ($\gamma = 0.25$)		92	89	11	14	0.904

4.4.4.2 K-fold cross-validation results

In order to compare different feature sets and their combinations, a 5-fold cross-validation (see Materials and Methods) was performed on classifiers I (intergenic distance), V (transcriptome data), VIII (intergenic distance and transcriptome data), and X (all features). Since the classifier VI based on terminator predictions alone has a large FPR, we did not include it in this comparative study.

ROC graphs were generated for each classifier, as described in Materials and Methods. As shown in Figure 4.4, the classifier V based on transcriptome data results in significant improvement compared to a random classifier (depicted by a diagonal 45^{o} line). Sixty percent of KOPs can be accurately classified with a FPR of 10% indicating that correlation between transcript profiles of adjacent genes can indeed be used for operon prediction. The radial SVM classifier I based on intergenic distance alone has similar recall and FPR as classifier V based on transcriptome data. Combination of these two features in classifier VIII results in a sharp increase in recall. At a FPR of 10% it can classify 75% of KOPs compared to 60% by classifier I. Addition of terminator predictions to intergenic distance and transcriptome data results in a small but noticeable improvement in classification accuracy (classifier X).

A comparison of the AUC of the four classifiers is shown in Table 4.3. With an AUC of 0.81, there is no significant difference between the distance-based classifier I and the transcriptome-based classifier V (p-value = 0.65, Wilcoxon signed rank test). Discretization of intergenic distance did not result in any decrease or increase in the AUC (data not shown). The radial SVM classifier VIII combining intergenic distance and transcriptome data has an AUC of 0.89, which is significantly greater than the AUC of distance-based classifier I (p-value = 1.1×10^{-4}). The radial classifier X combining all the three features has the largest AUC of 0.91 and is marginally better than classifier VIII (p-value = 6.7×10^{-3}).

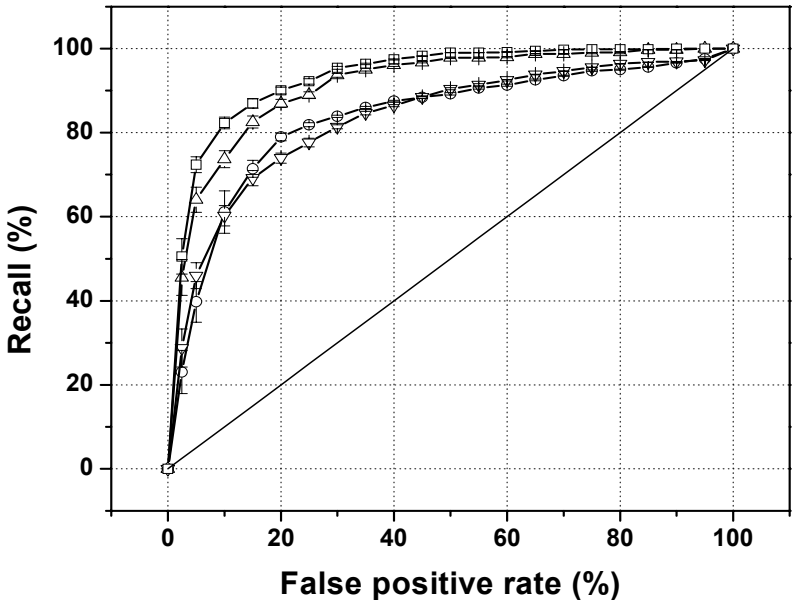

Figure 4.4. Comparison of different classifiers by ROC curve. False positive rate is the percentage of non-operon pairs (NOPs) misclassified as operon pairs and recall is the percentage of known operon pairs (KOPs) correctly classified as operon pairs. The ROC curves were generated for each classifier by a 5-fold cross-validation as described in the text. (o) classifier I; (▽) classifier V; (Δ) classifier VIII; (□) classifier X. Error bars indicate ± 1 standard deviation (*n* = 25).

Table 4.3. Comparison of different classifiers by 5-fold cross-validation. The null hypothesis was tested by comparing the AUC of the 25 ROC graphs for each classifier by Wilcoxon signed rank test

Classifier	Feature(s)	Average AUC	*p*-value	Null hypothesis
I	Distance	0.81	-	
V	Transcriptome	0.81	6.5×10^{-1}	$AUC_V - AUC_I = 0$
VIII	Distance and transcriptome	0.89	1.1×10^{-4}	$AUC_{VIII} - AUC_I = 0$
X	Distance, transcriptome, and terminator prediction	0.91	1.2×10^{-5}	$AUC_X - AUC_I = 0$
			6.7×10^{-3}	$AUC_X - AUC_{VIII} = 0$

The radial SVM classifier X was used to assign a score (*s*) to every gene pair in the training set. A positive score suggests a high likelihood that the adjacent genes are co-transcribed. At a score threshold of zero, 140 (94%) KOPs with a positive score were correctly classified. These 140 KOPs were divided into groups according to different range of scores, corresponding to increasing level of confidence – Group 1 ($0 < s < 1$), Group 2 ($1 \leq s < 1.2$), and Group 3 ($s \geq$

1.2). 37, 51, and 52 KOPs fall into the three groups, respectively. At higher score threshold in group 3, almost all the KOPs have transcript correlation $r > 0.7$ in at least one set of transcriptome data. Further, 32 (62%) gene pairs in group 3 have $r > 0.7$ in four or more sets. However the transcript correlation between adjacent genes reduces at lower scores in group 1 and group 2. Only seven (19%) of the 37 gene pairs in group 1 have correlation $r > 0.7$ in four or more sets. Thus, the score of a KOP reflects the degree of correlation between the transcript levels of adjacent genes – higher score indicative of a stronger correlation.

We also examined whether the extent of perturbation of a gene is important for determining its operon status. Using Shannon entropy of the expression level of a gene across all the microarray experiments as a measure of its perturbation, we found that the average entropy of a pair of genes in group 3 is greater than that in group 1 (p-value = 0.04, Kolmogorov-Smirnov test). This suggests that the operon status of adjacent genes with a higher degree of temporal variation in their transcript profiles can, in many cases, be determined with greater confidence.

4.4.5 IDENTIFICATION OF TRANSCRIPTION UNITS

Prediction of an entire transcription unit requires identification of intracistron genes as well as the genes at the cistron boundary. The genes in a pair with negative score have a low probability of co-transcription and are hence likely to have a cistron boundary between them. To examine the accuracy of our model to predict complete operons, we compared our classification results with known operons in the training set. Twenty-three known operons are dicistronic. All of them have a positive score indicating that they were successfully identified. Moreover, for all these 23 dicistrons, the identified cistron size is two, implying that the cistron boundaries were identified correctly.

Since many operons have more than two genes, it is important to identify adjacent gene pairs that are expressed as one transcription unit. Thirty-one known operons have more than two genes. Of these 9, 15, 2, and 5 operons have 3, 4, 5, and 6 or more genes, respectively. We examined adjacent KOPs and grouped them into operons if their score was greater than zero. Nineteen of those identified polycistrons have the same number of genes as that of the known operon, indicating that all the internal gene pairs as well as the cistron boundaries were correctly identified. Among these 19 operons, 5, 11, 1, and 2 have 3, 4, 5, and 6 or more genes, respectively. Interestingly, among the 31 operons with more than two genes, four have a larger number of genes than the size that has been reported suggesting that additional genes at the cistron boundaries are potentially co-transcribed. As will be described in the experimental

verification section, the prediction of those additional genes was verified for three operons by RT-PCR.

For eight of the known operons that have been reported to have more than two genes, the identified cistron size was less than the number of genes reported. They were incorrectly predicted as each consisting of two transcription units because one of the internal KOPs has a negative score.

4.4.6 OPERONS WITH INTERNAL REGULATION

The prediction of operons with internal control elements such as internal promoters, transcription factor binding sites, and transcription terminators is a challenging task [187, 208]. In the training set, nine known operons have been suggested to have internal regulation. Among these, four operons, *litQR*[209], *rsbB-rsbA-sigB*[210, 211], *trpCXBA*[212], *ushY-ushX-sigH*[213-215], have internal promoters. Additionally, the *rspO-pnp* operon has an intergenic transcription terminator[216, 217]. Another dicistronic operon *SCO3661-SCO3660* is induced by heat shock although constitutive expression of *SCO3660* has also been observed[218]. The *galTEK* galactose operon and the *recAX* operon involved in SOS response have been characterized in *S. lividans*. The galactose operon has two promoters, one upstream of *galT*, which is induced by galactose and another upstream of *galE* that is constitutively expressed[219]. In the *recAX* operon, the *recA* gene is expressed constitutively at a basal level whereas *recA-recX* transcript is observed in response to DNA damage[220]. The *rpsL-rpsG-fus-tuf1* operon has an internal promoter upstream of *tuf1* gene in *S. ramocissimus.* However, this promoter sequence is highly conserved among various *Streptomyces* spp. including *S. coelicolor* suggesting the possibility of a common regulatory mechanism[221, 222].

The transcript level of the genes in these operons may not be correlated due to internal regulation. We examined the features of the adjacent genes in these operons. In particular, for each of these operons, we examined the transcript correlation and intergenic distance of the pair of genes on either side of the internal regulation site. In three of these nine operons, the gene pair flanking the regulation site has correlation $r < 0.7$ in all the nine sets of transcriptome data. A notable exception is the genes in the dicistron, *rpsO-pnp*, which are strongly correlated ($r > 0.7$) in two of the nine sets of transcriptome data despite a recent report that identified the presence of an intergenic stem-loop structure, which acts as a site for RNase III processing and cleavage[217]. Also, the *rsbA-sigB* gene pair in *rsbB-rsbA-sigB* operon has transcript correlation $r > 0.7$ in two sets, although a developmentally-regulated internal promoter has been reported in the *rsbA-sigB* intergenic region[210, 211].

Interestingly, six of these nine gene pairs have a large intergenic distance (d > 150 bp). Adjacent genes in the same operon are rarely separated by intergenic distance exceeding 200 bp[186, 223]. Therefore, a high degree of transcript correlation, orthology, or functional similarity or combination of all these features is essential to predict the presence of a read-through transcript across these gene pairs. In only two of the nine operons, *galTEK* and *trpCXBA*, the gene with an upstream internal promoter is separated from its upstream neighboring gene by short intergenic distance (d < 25 bp).

4.4.7 OPERON PREDICTIONS FOR ENTIRE GENOME

Using a combination of transcriptome data obtained from several strains and culture conditions, and other features from the genome sequence, the SVM model was successful in classifying 94% of KOPs at a score threshold of zero. None of the features could achieve such a high degree of accuracy when used alone. The SVM classifier with all the features was therefore used for predicting the operon status of all same-strand pairs in *S. coelicolor* genome.

4.4.7.1 Overall analysis

The entire genome was arranged into 4965 pairs of genes in the same orientation (same-strand pairs) and 2859 pairs of genes in opposite orientation. Excluding the 149 KOPs, the 4816 same-strand gene pairs were further analyzed for co-transcription. The features, intergenic distance, correlation of transcript profiles, and the likelihood of a transcription terminator were calculated for every one of those pairs. TransTerm was used for prediction of transcription terminators [201]. Among the 2498 transcription terminators predicted in *S. coelicolor* genome, only 169 in the intergenic region of same-strand pairs with probabilities greater than 0.9 were retained.

The radial SVM classifier X was used to identify the same-strand gene pairs that have a high likelihood of co-transcription. Based on the features, the classifier predicts a score for every gene pair. The score distribution of these gene pairs is shown in Table 4.4. A total of 2012 of the 4816 same-strand pairs with unknown operon status have a positive score suggesting a high probability of co-transcription. Among these, 1369, 301, and 342 gene pairs fall into groups 1 ($0 < s < 1$), 2 ($1 \leq s < 1.2$), and 3 ($s \geq 1.2$), respectively. Both transcript correlation and intergenic distance play an important role in predicting a positive score for these gene pairs. At higher threshold in group 3, almost all the gene pairs have a transcript correlation $r > 0.7$ in at least 1 set, and 173 (51%) have $r > 0.7$ in four or more sets. As expected, the transcript correlations are somewhat lower among the gene pairs in group 1 and group 2 with lower scores (Table 4.4).

Moreover, the percentage of gene pairs with short intergenic distance (d < 25 bp) is higher in group 3 compared to group 1.

Table 4.4. Distribution of scores of same-strand gene pairs with unknown operon status

Score	No. of gene pairs	No. of pairs with r > 0.7[¶] in at least 1 set	No. of pairs with short intergenic distance (d < 25 bp)
s < -1	1452	161 (11%)	3 (< 1%)
-1 ≤ s < 0	1352	597 (44%)	123 (9%)
0 ≤ s < 1	1369	658 (48%)	1074 (78%)
1 ≤ s < 1.2	301	230 (76%)	264 (88%)
s ≥ 1.2	342	329 (96%)	307 (90%)
Total	4816	1975	1771

[¶] r is correlation between transcript profiles of the adjacent genes in a same-strand pair

Among the 4816 same-strand pairs, 1452 pairs have score less than -1. The transcript correlations among these 1452 gene pairs is significantly lower than the gene pairs with positive score, as shown in Table 4.4. Further, less than 1% of these 1452 pairs have short intergenic distance (d < 25 bp). Thus, the likelihood that adjacent genes in these pairs are co-transcribed is low; in other words, a cistron boundary is likely to exist in the intergenic region of those gene pairs.

4.4.7.2 Functional analysis

Genes involved in the same biochemical pathway/function tend to cluster together in prokaryotic genomes[224, 225], and are regulated similarly at transcription level. We therefore performed functional analysis to test if the genes in same-strand pairs with high score are functionally related. The protein classification scheme originally described by Monica Riley[226], and subsequently adapted for *S. coelicolor* was used (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml). Among the 7825 genes in the genome, 2371 (30.3%) encode hypothetical proteins without any known function and an additional 565 (7.2%) genes have putative assignments and do not belong to any functional class. Similarly 3264 (41.7%) genes are not categorized in any of the Gene Ontology classes. Thus,

4889 (62.5%) genes were assigned to 175 functional classes according to the scheme of Monica Riley[226].

Among the 149 KOPs in the training set, the adjacent genes in 121 pairs are functionally annotated. Ninety-two (76%) of these 121 pairs of adjacent genes belong to the same functional class. In contrast, out of the 72 NOPs in which both genes are annotated, only eight (11%) share the same functional class. We examined the functional relatedness of genes in same-strand pairs grouped according to their scores. At higher score threshold in group 3, the genes in 67% of the pairs belong to the same functional class. However, the functional similarity between adjacent genes decreases at lower score thresholds in group 1 and 2, as shown in Table 4.5. This trend of decreasing functional similarity is more vivid when we examine gene pairs with negative score. Only 106 (18%) pairs of adjacent genes with score less than -1 share the same functional class. This sharp difference in functional similarity of gene pairs with positive and negative score is consistent with our operon predictions. Among the functional classes shared by pairs of adjacent genes with positive score, the class of transport/binding proteins is the most abundant followed by the genes involved in secondary metabolism and its subclass polyketide synthases (PKS).

Table 4.5. Functional analysis of same-strand gene pairs

Score	No. of gene pairs	No. of annotated gene pairs	No. of pairs in same functional class
$s < -1$	1452	605	106 (18%)
$-1 \leq s < 0$	1352	521	117 (22%)
$0 \leq s < 1$	1369	667	317 (48%)
$1 \leq s < 1.2$	301	169	100 (59%)
$s \geq 1.2$	342	206	137 (67%)

4.4.8 EXPERIMENTAL VERIFICATION

To confirm the co-transcription of predicted gene pairs, RT-PCR was performed on some pairs using primers that amplify across their intergenic region. To allow for a large number of primer sets to be used readily, we employed the primers previously used for construction of whole-genome *S. coelicolor* microarray. Based on the success rate of amplification in preliminary

RT-PCR experiments, the verification was limited to only those gene pairs whose amplicon size was no larger than 2.5 kb. This list of gene pairs was further constrained by considering only those whose transcript profiles have a correlation $r > 0.7$ in at least one of the nine sets. With those criteria 114, 91, and 163 gene pairs in group 1 ($0 \leq s < 1$), 2 ($1 \leq s < 1.2$), and 3 ($s \geq 1.2$), respectively were selected for verification. A number of examples of gene pairs verified by RT-PCR are shown in Figure 4.5.

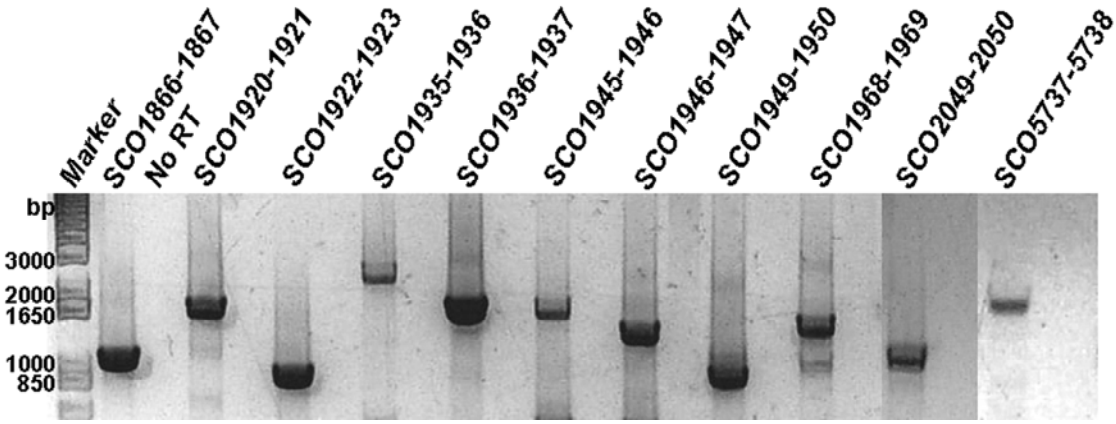

Figure 4.5. Experimental verification of co-transcription of adjacent genes by RT-PCR. RNA isolation and RT-PCR was performed as described in Materials and Methods. Primers were used to amplify across adjacent genes and the products were analyzed by gel electrophoresis. The expected sizes of the amplicons in bp are: *SCO1866-1867* – 1035; *SCO1920-1921* – 1637; *SCO1922-1923* – 885; *SCO1935-1936* – 2298; *SCO1936-1937* – 1633; *SCO1945-1946* – 1659; *SCO1946-1947* – 1353; *SCO1949-1950* – 964; *SCO1968-1969* – 1426; *SCO2049-2050* – 980; *SCO5737-5738* - 1615. For every gene pair, a negative control (No RT) in which the RT enzyme was not added was also performed. The negative control is shown next to each RT reaction.

Table 4.6. RT-PCR based verification of co-transcription of gene pairs

Score	No. of gene pairs tested	No. of gene pairs verified	Range of Intergenic distance (bp)
$s \geq 1.2$	163	122 (75%)	-32 to 131
$1 \leq s < 1.2$	91	61 (67%)	-8 to 178
$0 \leq s < 1$	114	67 (59%)	-13 to 178

Overall 250 (68%) of the 368 gene pairs tested were verified to be on the same operon. The distribution of gene pairs according to different range of scores is listed in Table 4.6. At lower scores in group 1, a transcript was detected in 59% of the gene pairs tested. The percentage of gene pairs verified to be on the same operon increases to 67 and 75 at higher scores in group 2 and 3, respectively. The range of intergenic distance of the gene pairs that were verified by RT-PCR is also shown in Table 4.6. 106 (87%) of the 122 verified pairs with $s \geq 1.2$ have short intergenic distance ($d < 25$ bp). At lower threshold of $s < 1.2$, 13 of the 128 verified pairs have intergenic distance exceeding 100 bp.

It is possible that some of the tested gene pairs that were not positively verified to be on the same operon are false positive predictions. However, a conclusion on those gene pairs cannot be drawn easily. The sample RNA used for RT-PCR was pooled from wild-type M145 cells at different culture stages. In contrast, the transcript profiles were obtained also from different mutants across a range of culture conditions. It is also possible that some of the tested operons were not transcribed in any of the cell samples collected.

4.4.8.1 Verification of false negative predictions

From the results of binary classification, 6% of the KOPs have a negative score suggesting that the SVM model did not accurately classify them to be on the same operon. Also the 92% recall of SVM classifier X (Table 4.2) indicates that the model will misclassify 8% of the KOPs. Therefore, it is likely that some of the 2804 same-strand gene pairs with $s < 0$ are in fact co-transcribed. To identify some of these gene pairs, we compared the scores of these 2804 gene pairs with the operon predictions of Price *et al.* [187]. The authors used a Bayesian approach with features derived from the genome sequence to predict the probability (*pOp*) that two adjacent genes in *S. coelicolor* are in the same operon. We performed RT-PCR on a restricted subset of 60 gene pairs which have $s < 0$ and $pOp > 0.6$. A PCR product was observed in 16 of these gene pairs. A closer examination of their transcript profiles revealed that 10 of these 16 pairs have correlation $r < 0.7$ in all the nine sets of transcriptome data and only two pairs have $r > 0.7$ in more than one set. The weak correlation between transcript levels in these pairs is a potential cause for their misclassification by our model.

4.4.8.2 Extension of boundaries of known operons

In our analysis, we sought to identify groups of adjacent genes that are expressed as a single transcription unit. When we grouped consecutive gene pairs in the training set with score

greater than zero, we observed in four cases that the identified polycistron size was greater than the size of the known operon. This indicates that additional gene pairs at cistron boundaries with positive score were predicted to be co-transcribed. One of these is the *rspO-pnp* (*SCO5736-5737*) dicistron [216, 217]. The transcript profile of *SCO5738*, encoding a putative protease downstream of this operon is strongly correlated with *SCO5737*. Moreover, the gene pair, *SCO5737-5738*, has a high score of 1.1. We verified the presence of a transcript across their intergenic region indicating that the operon has more than two genes (Figure 4.5). Further, the two downstream genes *SCO5739* and *SCO5740* encoding putative dihydrodipicolinate reductase and putative membrane protein, respectively are strongly correlated with *SCO5738* and both the genes are predicted to be part of the same operon. We have also verified the co-transcription of the genes *SCO5739* and *SCO5740* by RT-PCR.

Table 4.7 lists examples of two other operons for which pairs of adjacent genes at cistron boundaries with positive score were verified to be on the same operon. It is important to note that the reports characterizing these operons did not exclude the possibility of additional adjacent genes being part of the same transcription unit. By definition, the gene pairs at these cistron boundaries were used as NOPs in the training set and they were consistently classified as false-positives due to their positive scores. We have shown that the genes in these pairs are indeed co-transcribed in agreement with our predictions. Hence, the leave-one-out false positive rate (FPR) from our predictions is likely to be lower than 14%.

Table 4.7. Extension of cistron boundary of known operons

No.	Known operon	Known size	Predicted operon	Gene pairs verified by RT-PCR	Reference
1	*SCO5736-5737 (rspO-pnp)* (Protein synthesis)	2	*SCO5737-5740*	*SCO5737-5738, SCO5739-5740*	216, 217
2	*SCO2050-2054 (hisAHBCD* (Histidine biosynthesis)	5	*SCO2048-2054*	*SCO2049-2050*	227, 228
3	*SCO5583-5585 (amtB-glnK-glnD)* (Nitrogen metabolism)	3	*SCO5583-5586*	*SCO5585-5586*	229

4.5 DISCUSSION

Operon is the unit of transcriptional regulation in prokaryotes. Identifying operon structure in a genome is important to the study of gene expression regulation. The estimation of transcript level of a gene using microarrays can also be improved by using information about the transcriptional activity of the genes that are co-transcribed with it. This can also translate into an improvement in identification of differentially expressed genes[230]. In our study of *S. coelicolor* A3(2) and disruption mutants of regulatory genes, we have compiled a series of time profiles of transcriptome data. Such dynamic transcriptome profiles can be valuable in elucidating operon structure. Combining with transcriptome data on several *S. coelicolor* strains and culture conditions from two other studies[88, 200], the dynamic behavior of adjacent genes in the genome were used to assess the likelihood of their being in the same operon. In principle, the expression profiles of genes on the same operon should be well correlated, at least under conditions of no alternative regulation. However, in reality transcriptome data are often riddled with noise especially when the transcription level is low, rendering microarray assay insensitive to dynamic changes. In our analysis, we thus incorporated other features characteristic of genes in the same operon.

4.5.1 DEPENDENCE OF TRANSCRIPT DYNAMICS FOR OPERON PREDICTION

An essential condition for accurate calculation of correlation between genes is that they are expressed above noise level and exhibit sufficient dynamics across different experiments[188, 231]. Among the KOPs in the training set, gene pairs which have a higher score tend to exhibit more dynamics in their temporal transcript profiles. Consistent with this notion the Shannon entropy, a measure of transcript variation, of same-strand pairs with high score ($s \geq 1.2$) was found to be greater than that of same-strand pairs with score in the range $0 \leq s < 1$ (p-value = 1.3×10^{-73}, Kolmogorov-Smirnov test).

The operon prediction improves significantly when transcriptome data from diverse experimental conditions are incorporated in the model. Using transcript profiles from only M145 strain in modified R5 medium, merely 60% of the KOPs could be accurately classified with a high FPR of 35%. This study thus included temporal transcriptome data from several *S. coelicolor* strains under very different culture conditions and from different sources including ours. Using all those transcriptome data, 80% of the KOPs could be classified at a considerably lower FPR of 21%. Moreover, the performance of the transcriptome-based classifier was

comparable to the intergenic distance-based classifier (Figure 4.4, Table 4.3). As more transcriptome data become available in the future, especially when conditions under which data are acquired increase, the classification framework can be used to further expand the repertoire of operons identified.

4.5.2 OTHER FEATURES

4.5.2.1 Intergenic distance feature

The intergenic distance between two adjacent genes in an operon is shorter on an average compared to that of same-strand pairs, which are not in the same operon. This feature was first used for operon prediction in *E. coli*[185, 232]. Several studies have subsequently showed that intergenic distance can be effectively used for operon prediction in other prokaryotes[187, 191], for which the genome sequence is available. Using a log-likelihood function based on intergenic distance distribution 75% of transcription units in *E. coli* could be predicted[185]. It has been reported that operon predictions using intergenic distance has the highest recall and lowest FPR among the various features derived from genome sequence including codon usage, promoter predictions, and terminator predictions[197]. In this study using intergenic distance a FPR of 20% was seen at a recall of 80% (Figure 4.4). However, the FPR increased sharply to 35% as recall increased to 85%, indicating that intergenic distance cannot alone be used to achieve high recall at acceptable error rates.

4.5.2.2 Transcription Terminator feature

Transcripts of operons, particularly those without any internal regulation, are likely to terminate at a single transcription terminator. Therefore, the likelihood of a terminator in the intergenic region of intra-operonic genes is low. Several studies have used this feature for operon prediction in prokaryotes[190, 193, 197, 208]. Due to the highly degenerate nature of the binding site of Rho-factor (called 'rut' site)[233], identification of rho-dependent terminator site is difficult. Most terminator prediction algorithms identify rho-independent transcription terminators, which have a characteristic hairpin structure[201, 234-236]. Among the same-strand pairs in the *S. coelicolor* genome, less than 5% have high confidence (probability > 0.9) terminator predictions in their intergenic region. Hence this feature cannot be used to infer the operon status of a large fraction of the same-strand pairs.

4.5.3 PREDICTION OF TRANSCRIPTION UNITS

In this study every same-strand gene pair was assigned a score, and a score threshold of zero was used to group consecutive gene pairs into operons. A total of 5664 transcription units were predicted of which 1278 are polycistronic with two of more genes. The distribution of cistrons of different sizes is summarized in Table 4.8.

Table 4.8. Size distribution of the predicted transcription units

Cistron size	No. of cistrons	No. of cistrons with $s > 1$ in all gene pairs	No. of cistrons with $s > 1.2$ in all gene pairs
1	4386	-	-
2	839	203	85
3	235	33	13
4	111	17	6
5	46	5	3
> 5	47	2	0

4.5.3.1 Large Operons

Among the polycistrons, 47 (3.7%) have more than five genes of which 11 cistrons have 10 or more genes. This includes a large 27.5 kb, 21 gene operon *SCO0381 – SCO0401* comprising a secondary metabolite gene cluster. The genes in this operon encode deoxysugar synthases involved in the synthesis of an unknown secondary metabolite. Eight of the 20 gene pairs in this operon have $s \geq 1.2$ implying strong predictions of co-transcription of these genes. Another 16.5 kb long, 15 gene operon *SCO3235 – SCO3249* encodes for genes involved in the synthesis of Calcium dependent antibiotic (CDA) – a peptide antibiotic synthesized by non-ribosomal peptide synthases. Due to strong correlation between their transcript profiles, six of the 14 gene pairs in this operon have scores greater than 1.2.

4.5.3.2 Operons with internal regulation

Organization of genes into operons allows for an efficient way of coordinated response by the organism to environmental changes. However, there might also be circumstances in which an organism may need the product of the genes on an operon differently, either stoichiometrically or temporally, than the way they are normally prescribed. Thus, some flexibility to allow for an

escape from the coordinated expression in an operon is necessary in some cases. A promoter or transcription terminator within an operon allows for differential expression of genes in the same operon. Among the known operons in the training set, many have an internal regulation site in the intergenic region of intra-operonic gene pairs. Several of these gene pairs have wide intergenic spacing. Interestingly, wide spacing between adjacent genes in operons has been reported to indicate complex regulation[184]. Based on transcriptome data and prediction of transcription factor binding sites and terminator sites, nearly 20% of operonic genes in *S. coelicolor* are thought likely to be internally regulated[222]. The existence of internal regulation in a same-strand gene pair may reduce the degree of correlation of their transcript profiles. A better prediction of internal regulation will certainly improve the operon predictions.

4.5.4 USING OPERON PREDICTIONS FOR FUNCTIONAL ANNOTATIONS

Chromosomal proximity of adjacent genes in multiple prokaryotic genomes and their co-transcription is often an indication of their functional relatedness. The information can be used to infer their functional annotation[224, 225, 237]. In this study we observed an increasing trend of functional similarity between pairs of adjacent genes with increasing scores. A significant number of pairs are comprised of genes that do not share the same functional class, many of which have a high score ($s \geq 1.2$) and/or have been verified by RT-PCR as being on the same operon. Of these, 113 pairs have only one gene functionally annotated, with the other gene is either hypothetical or unclassified. The transcript profiles in almost all these gene pairs are strongly correlated ($r > 0.7$) in at least one set of microarray data. Potential relatedness of the physiological function may be inferred from the well annotated neighboring gene.

4.5.5 COMPARISON OF OPERON PREDICTIONS WITH EARLIER REPORT

In this study, we employed SVMs as a classification tool for operon prediction. SVMs have been recently used to study several classification problems in bioinformatics[172, 238]. These studies have demonstrated that SVM-based classifiers produce results that are better than or at least as good as those obtained by other supervised methods, as the classification models that they generate tend to better generalize on unseen instances (i.e., instances that were not used during training).

Other methods have also been used successfully for operon prediction. A Bayesian approach was used to predict operons in all sequenced prokaryotes including *S. coelicolor*[187]. The method relies on features based on genome sequence alone and uses intergenic distance,

conservation of gene clusters across different organisms, codon usage, and functional similarity to predict a probability (*pOp*) that two adjacent genes with the same orientation are co-transcribed. The authors used *S. coelicolor* genome annotation from The Institute of Genomic Research (TIGR), whereas the primary annotation from Sanger Institute (http://streptomyces.org.uk/) was used for this study. Out of 4965 same-strand pairs, 4549 pairs can be compared with their predictions.

A 5-fold cross-validation of the training set was used to compare the two methods. A reduced training set of 139 KOPs and 58 NOPs for which operon predictions were available from both studies was used for cross-validation. A comparison of the vertically averaged ROC curves is shown in Figure 4.6. The difference between the predictions from the two methods is most evident at recall greater than 70%, where the FPR from our method is significantly less than the FPR from the predictions of Price and co-workers. Consequently, the AUC for our SVM classifier is greater than the latter study (p-value = 5.2×10^{-4}, Wilcoxon signed rank test). However it is important to note that the size of training set used in this comparison is smaller than the total number of KOPs and NOPs used for comparison of different SVM classifiers described earlier. For small sample sizes the estimate of error obtained from cross-validation (leave-one-out as well as k-fold) can be highly variable. Classifiers based on small sample sizes are particularly vulnerable to situations where the perturbations introduced by k-fold partitioning results in an unstable classifier – a classifier with unreliable accuracy estimates[206, 239].

We also compared the predictions of the two methods on all the same-strand gene pairs in the genome. Using a threshold of 0 and 0.5 for score and *pOp*, respectively, 3730 of the 4410 same-strand pairs with unknown operon status have the same predictions. The gene pairs for which the predictions of the two methods do not match include 125 pairs, which have $s > 0.5$ and $pOp < 0.4$. Adjacent genes in these pairs are predicted to be co-transcribed by our SVM model but not the Bayesian approach. More than 70% of these gene pairs have a transcript correlation $r > 0.7$ in at least two of the nine sets of microarray data suggesting that our method emphasizes expression correlation to predict the likelihood of co-transcription. On the other hand 151 gene pairs have $s < -0.5$ and $pOp > 0.6$ indicating that these pairs were predicted to be co-transcribed by the Bayesian model but not by our SVM model. Among these pairs, 68% do not have a transcript correlation $r > 0.7$ in any of the nine sets and only 7% have correlation $r > 0.7$ in two or more sets. These results suggest that the information based on similarity of transcript profiles of

adjacent genes plays an important role in our model predictions and thereby enhances the performance of operon prediction models that rely on genome sequence-based features alone.

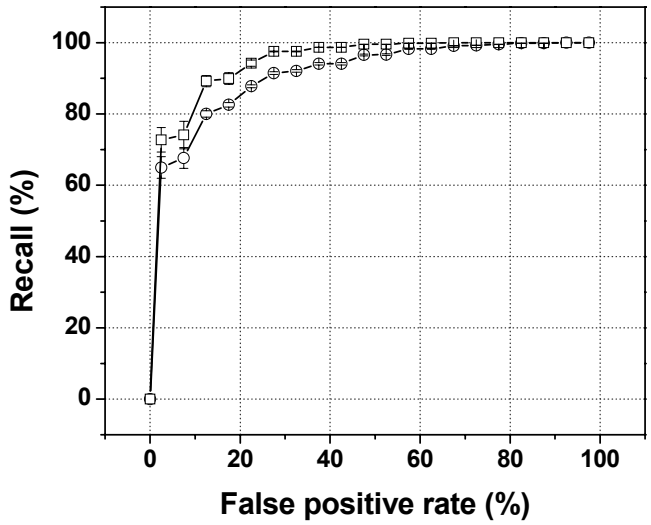

Figure 4.6. Comparison of the performance of SVM classifier with the predictions of Price *et al.*[187] using 5-fold cross-validation. Average and standard deviation of the 25 ROC curves was obtained by vertical averaging described in the main text. (□) SVM classifier; (o) Prediction from Price *et al.*

4.6 CONCLUDING REMARKS

S. coelicolor has a rich genome that encodes more genes than even the eukaryote, *Saccharomyces cerevisiae*. Its versatility to undergo differentiation with a complex life cycle and produce secondary metabolites after the cessation of growth is reflected in its high dynamic temporal transcriptome profile. A large number of genes on the genome are not well annotated, with many annotated as involved in, or hypothesized to be involved in, regulation. A better understanding of its operon structure will be valuable in gene annotation and large-scale gene expression studies for elucidating its regulatory networks that control differentiation and antibiotics production. In this study we used a support vector machines-based supervised classification approach to predict operon structure for this organism. In the past few years the transcriptome data of this organism has become a valuable resource for gaining physiological insights. The use of time series transcriptome data enhanced the predictive capability of the classifier that employed genome sequence-based features including intergenic distance and

transcription terminator predictions. The experimental verification of a large set of those predicted by the classifier further demonstrates the utility of the method. As more transcriptome data becomes available and the conditions under which they are obtained diversify, the framework established in this study will also become more versatile in further enhancing the operon predictions.

CHAPTER 5 FURTHER REFINEMENT OF OPERON PREDICTIONS AND DISCOVERY OF REGULATORY HUBS FOR *STREPTOMYCES* SECONDARY METABOLISM

5.1 SUMMARY

The onset of antibiotics production in *Streptomyces* species is co-ordinated with morphological differentiation. An understanding of the genetic circuits that regulate these coupled biological phenomena is essential to discover and engineer the pharmacologically relevant natural products made by these species. Despite the enormous experimental as well as computational challenges, a systems approach of integrating diverse genome-scale datasets to elucidate these complex networks is beginning to emerge. In this study, more than 500 samples of genome-wide temporal transcriptome data compiled from our laboratory and public repositories, comprising wild-type and more than 25 regulatory gene mutants of *S. coelicolor* probed across multiple stress and medium conditions, were examined. Information based on transcript and functional similarity was used to predict transcriptional networks constituting modules enriched in diverse functions such as polyketide synthesis and electron transport. Further, the previously-predicted whole-genome operon map was refined using a larger feature set. These efforts to tap omics resources will aid in unearthing the regulatory cascades of functional processes relevant for secondary metabolism in *S. coelicolor*.

5.2 INTRODUCTION

The biosynthesis of several pharmacologically important secondary metabolites produced in Streptomycetes is controlled by complex and interconnected cascades of gene products. Deciphering these regulatory cascades is a challenging task. The network of secondary metabolism in *S. coelicolor* is likely to involve several (>100) genes many of which are currently not known or their precise functions are not well understood. Construction of gene knockouts to understand these regulatory mechanisms is widely practiced. However, perturbing gene networks by large-scale construction and characterization of single and double gene knockouts is an

enormous task, which has been accomplished only for model organisms such as *Escherichia coli*[240, 241]. With approximately 965 regulatory genes (genes encoding proteins with predicted regulatory roles) in *S. coelicolor,* the task of systematic genetic perturbation is formidable. Nevertheless, recent advances in techniques for genetic manipulation of *S. coelicolor*, such as PCR-targeted gene replacement[242] and the availability of tools for genome-scale probing with DNA microarrays[39] and 2D LC-based proteomics[91] will facilitate system-wide analyses of *S. coelicolor* secondary metabolism. Global expression profiling allows us to examine the temporal changes in gene expression that rapidly occur as the critical regulatory events unfold. These changes provide valuable cues about the dynamical associations between different genetic elements in these networks.

However, analyzing omics datasets to decipher transcriptional interactions between network components (termed 'reverse engineering') is another computational challenge. Limited success has been achieved by methods based on Boolean networks[243], Bayesian networks[244, 245], non-linear ordinary and partial differential equations[246], and stochastic equations[247, 248]. A Boolean approach approximates continuous gene expression information as a binary state (ON or OFF). Further, gene-gene interactions can be modeled as Boolean logic functions (e.g., AND, OR, NOT). These simplifications are particularly suitable for analyzing large networks. A number of different Boolean algorithms based on time-series as well as steady-state gene expression data have been proposed to identify the topology of gene networks[243, 249, 250]. Bayesian (belief) networks (BN), where gene-gene relations are modeled as probabilistic edges, represent a generalization of Boolean networks[251]. BN are well suited for networks where the structure is only partially known and can be 'learned' from the data[251]. Dynamic Bayesian networks (DBN) have been proposed as an alternative to BN to allow feedback regulation between genes in a time-dependent manner[251, 252]. Although many studies on gene networks models employ large-scale gene expression datasets, several recent reports are based on additional data types, such as information based on protein-DNA interactions[253, 254] and protein-protein interactions[255].

The architecture of a transcriptional network also plays a crucial role in determining the stability and adaptability of an organism to a variety of genetic and environmental perturbations[256]. Large networks can be dissected into basic building blocks called 'network motifs' or 'network modules' that are structurally stable and recur throughout the network (reviewed in Alon, U[257]). These motifs include: (i) single-gene autoregulation (positive and negative) (ii) three-gene feedforward loops where a regulator X controls Y, and Y in turn

regulates Z. An additional interaction between X and Z completes the feedforward loop (iii) single input module where a single regulator X regulates multiple genes P, Q, R, and (iv) dense overlapping regulon where multiple regulators control the expression of multiple targets in a many-to-many format. These motifs can explain a variety of dynamic and robust responses that are commonly observed in transcriptional circuits. For instance, the positive and negative feedback loops in a two-gene ScbA/ScbR system in *S. coelicolor* play an important role in regulating antibiotic production[70, 258]. Through mathematical analysis, it was demonstrated that the system can behave as a bistable switch to trigger the expression of genes in a cryptic type I polyketide cluster[259]. In one of the early reports, Shen-Orr et al[260] demonstrated that network motifs are observed in disproportionately high numbers in transcriptional networks in *Escherichia coli* compared to randomly generated networks. Similar network motifs have also been examined in other biological systems, such as the transcriptional network of *Saccharomyces cerevisiae,* developmental network of *Drosophila melanogaster,* and the synaptic network of *Caenorhabditis elegans* further highlighting the fundamental importance of network motifs that impart dynamical stability in biological networks[261].

The transcriptional network that regulates primary and secondary metabolism in *S. coelicolor* is not well-understood. With increasing availability of genome-wide expression data, large-scale systematic studies to explore and decipher these transcriptional programs can provide valuable cues and hypotheses for experimental verification. In this chapter, the operon predictions (from the previous chapter) were refined using a more extensive catalogue of gene expression profiles and additional operon predictive features. Using the improved operon map, the transcriptomic and functional characteristics of more than 4000 cistrons were integrated in an attempt to reconstruct the transcriptional regulatory network of *S. coelicolor*.

5.3 METHODS

5.3.1 MICROARRAY DATA

5.3.1.1 Data compilation and processing

The microarray dataset (Table 5.1) was compiled from data that is publicly available in Stanford Microarray Database, Gene Expression Omnibus (GEO), and ArrayExpress. Additionally, the microarray datasets for characterization of deletion mutants of *scbA* (M751) and *scbR* (M752) were obtained from Takano *et al.* (unpublished results).

Several genes in each sample (cDNA:gDNA and cDNA:cDNA) were flagged 'absent' due to low regression coefficient ($R^2 < 0.1$) or due to low signal-to-noise ratio (ratio of mean intensity and median background intensity less than 2.5). Samples with absent flags for more than 30% genes were discarded before further analysis. The resulting database comprised 263 cDNA:gDNA samples (dataset 1), 156 cDNA:cDNA samples (dataset 2) and 105 samples hybridized on Affymetrix diS_div712a GeneChips (dataset 3). All experiments with genomic DNA as reference (dataset 1) were normalized using a quantile normalization method[198]. The experiments performed with Affymetrix diS_div712a GeneChips were normalized using the 'affy' package[262] in the R statistical computing environment. The following options were used for normalization, (i) Background correction was performed using Affymetrix Microarray Suite (MAS 5.0) algorithm[263], (ii) Probe pair level intensities were normalized by linear scaling ('constant' option), (iii) Perfect match (PM) correction was performed using MAS 5.0 algorithm, (iv) Median polish method was used to calculate expression summary at probe set level.

The ARACNe (algorithm for accurate reconstruction of accurate cellular networks) (see details below) requires a complete gene expression matrix, i.e., a gene must not have any absent flags (missing values). The expression dataset was therefore further processed to eliminate genes with absent flags in several samples and those with low expression dynamics. The following parameters were used for gene selection: (a) fraction of absent flags in dataset 1, (b) Standard deviation in dataset 1 (25^{th} percentile of the standard deviations is 0.50), (c) fraction of absent flags in dataset 2, (d) Standard deviation in dataset 2 (25^{th} percentile of the standard deviations is 0.43), (e) presence of a probeset on Affymetrix GeneChip. The following Boolean logical criterion was used for gene selection:

$$\{(a \le \mathbf{0.20})\, AND\, (b \ge \mathbf{0.50})\, AND\, (c \le \mathbf{0.50})\, AND\, (e = TRUE)\} \text{ OR } \{(c \le \mathbf{0.20})\, AND\, (d \ge \mathbf{0.43})\, AND\, (a \le \mathbf{0.50})\, AND\, (e = TRUE)\}$$

Missing values for the selected genes were estimated using the *k*-nearest neighbor method[264]. In each dataset (1, 2, and 3), the expression of every gene was *z*-standardized to an average and standard deviation of 0 and 1, respectively.

5.3.1.2 Similarity estimation

Similarity between the transcript levels of genes in every pair was calculated by two different metrics: (i) Pearson correlation coefficient (ii) A robust correlation described by Hardin *et al.*[265]. The R code provided by the authors was used for estimating the robust correlation.

Table 5.1. Summary of microarray data compiled for analysis

No.	Strain	Description	Array[a]	No. of samples	Reference
1	M145	Growth kinetics of wild-type M145	cDNA: gDNA	14	[66]
2	M145	Growth kinetics of wild-type M145	cDNA: gDNA	8	[266]
3	M145, M751, M752	Response to deletion of *scbA* and *scbR*	cDNA: gDNA	52	Takano *et al.* (unpublished results)
4	LY2002, YSK360, WL6268	Response to disruption of *SCO2517, SCO3654, SCO6268*	cDNA: gDNA	35	[17, 18]
6	M145, MT1110	Response to disruption of *cdaR*	cDNA: gDNA	57	[267]
7	J1501, YD2108	Response to osmotic stress	cDNA:cDNA	13	[268]
8	J1915, J2177	Response to *bldN* null mutant	cDNA:cDNA	11	[269]
9	CH999 derivatives	Kinetics of growth	cDNA:cDNA	12	[270]
10	M145	Growth kinetics of wild-type M145	cDNA:cDNA	11	[39]
11	J1501, BZ5	Response to disruption and constitutive over-expression of *ramR*	cDNA:cDNA	8	[271]
12	M600, M570, M653, M667	Response to *relA* disruption and inducible expression	Affymetrix[¶]	105	[86]

(a) cDNA:cDNA indicates that cDNA from two RNA samples was used in the assay. cDNA:gDNA indicates that genomic DNA (gDNA) from the parent strain was used as the reference sample. (¶) Affymetrix diS_div712a GeneChips were used in the study.

5.3.2 FUNCTIONAL SIMILARITY

Similarity between the biological functions of two genes in every pair was estimated by different methods.

5.3.2.1 Similarity based on protein classification scheme

The protein classification scheme proposed by Riley[226], which was adapted for *S. coelicolor* (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml) was used to estimate the similarity between adjacent genes. All the protein encoding ORFs were classified into 140 functional classes. The similarity score between two adjacent genes is 1 if both the genes belong to the same functional class, whereas the score is -1 if the genes belong to different functional classes. Similarity score was not calculated if either of the two genes was categorized as 'unclassified' or 'unknown function'.

5.3.2.2 Similarity based on gene ontologies (GO)

5.3.2.2.1 Czekanowski-Dice metric

The Czekanowski-Dice similarity metric for comparing two genes based on their ontological classes was proposed by Martin *et al.*[272]. It is estimated as follows:

$$\text{Czekanowski-Dice score} = \frac{2c}{a+b}$$

where, a and b are the number of GO terms associated with gene 1 and gene 2, respectively. The number of GO terms shared by both is equal to c. The score ranges from 0 (dissimilar) for genes that do not share any GO term, to 1 (identical) for genes that share all their GO terms.

5.3.2.2.2 Similarity metric derived from Information theory

Information theoretic similarity measures introduced by Resnik[273], Lin[274], and Jiang and Conrath[275] were used to estimate the functional similarity between genes based on their GO terms. The similarities were estimated using the GOSim[276] package in the R statistical computing environment.

5.3.3 CONSERVATION OF GENE ORDER

The number of bacterial genomes in which the orthologs of a pair of adjacent genes are present in the same chromosomal order was used for operon prediction. This information, was obtained from OperonDB[186, 277].

5.3.4 SUPERVISED CLASSIFICATION

5.3.4.1 Model training and selection

SVM^{light}, an implementation of support vector machines was used to construct and evaluate supervised classification models for operon predictions[202]. The performance of classifiers with different features was compared by a 10-fold cross-validation scheme. The classifiers were evaluated based on recall, false positive rate, and area under ROC curves (AUC), as described in the previous chapter.

5.3.4.2 Training set – Positive and negative classes

Positive and negative classes were defined as known operon pairs (KOP) and non-operon pairs (NOP), respectively, as described in the previous chapter. The 149 KOPs used in our earlier study were included here in the positive training set. Additionally, adjacent genes in 266 gene pairs, which were positively verified to be co-transcribed in our earlier study, were included in the positive training set. Further, eleven gene pairs from six recently reportedly verified operons in *S. coelicolor* were also included in the positive training set. These operons are *nikABCDE*[278], *devAB*[279], *nrdABS*[280], *nrdRJ*[280], *znuACB*[281], and *rpmG3-rpmJ2*[282]. This constitutes a positive training set of 426 KOPs. The negative training set comprises 131 NOPs. Starting with the 122 NOPs that were used as the negative set in the previous study, the three NOPs that were positively verified to be co-transcribed (Table 4.7) were removed from the negative training set. Additionally, twelve NOPs from the six recently reported operons were added to the training set.

5.3.5 FUNCTIONAL NETWORK ANALYSIS

5.3.5.1 Mutual information-based transcriptional network prediction using ARACNe

Transcriptional networks were predicted on genome-scale using ARACNe[283, 284]. ARACNe uses a two-step approach for predicting gene-gene interactions.

(i) The statistical dependency between a pair of genes is computed by estimating the mutual information between their gene expression profiles. Mutual information between any two discrete

random variables (*a* and *b*) is defined as $I(A,B) = S(A) + S(B) - S(A,B) \geq \mathbf{0}$. $S(X)$ corresponds to the entropy of any variable *X*. The entropy of a discrete variable *X* is calculated as $S(X) = -\sum_i p(x_i)\log(p(x_i))$, where $p(x_i)$ is the probability of observing the variable *X* in any one of the *i* discrete states.

(ii) Indirect interactions are eliminated using the data processing inequality (DPI). According to DPI, if a set of three genes (*A, B, C*) are interacting such that *A* directly interacts with *B*, which in turn directly interacts with *C*, then the indirect interaction between *A* and *C* can be eliminated based on the assertion that $I(A,C) \leq \min(S(A,B), S(B,C))$.

A *p*-value threshold of 1.0×10^{-9} was used as the statistical criteria for identifying significant interactions. A DPI tolerance of 0.05 (5%) was used. Resulting transcriptional networks were visualized in Cytoscape version 2.6.0[285].

5.3.5.2 Networks with functional enrichment

Network modules comprising multiple cistrons predicted as direct targets of one transcription factor were extracted from the global network. Fisher's exact test was used to identify motifs where a significant fraction of genes were involved in the same biological pathway or function. Modules with functional enrichment *p*-value less than 1.0×10^{-4} were considered significantly enriched. Functional annotation was based on two sources – 137 classes based on protein classification scheme and 173 classes based on Gene Ontologies.

5.4 RESULTS

In the previous chapter, a SVM-based machine learning method was employed to predict a whole genome operon map using transcriptome data, intergenic distance, and transcription terminator predictions. Further, adjacent genes in 266 pairs were experimentally verified based on model predictions. Here, we refine the predictions of the previous model using a larger repertoire of temporal transcriptome data and additional operon prediction features based on functional similarity of adjacent genes as well as conservation of gene order in multiple prokaryotic genomes.

5.4.1 FEATURES USED FOR GENOME-WIDE OPERON PREDICTION

5.4.1.1 Gene expression similarity

Temporal transcriptome data compiled from different sources (Table 5.1) was preprocessed to eliminate samples with high percentage of absent gene flags (see Methods). The resulting database comprised a total of 524 cell samples. This is significantly more than the 206 samples used in the previous study. Figure 5.1 shows the distribution of Pearson correlation (r) between the pairs of adjacent genes in the positive and the negative training sets. A correlation $r > 0.7$ was observed between adjacent genes in 148 (35%) of KOPs. In contrast, adjacent genes in only 3 (2%) NOPs have a correlation $r > 0.7$. The sharp discrimination between the two classes strongly indicates the importance of temporal transcriptome data for predicting operons.

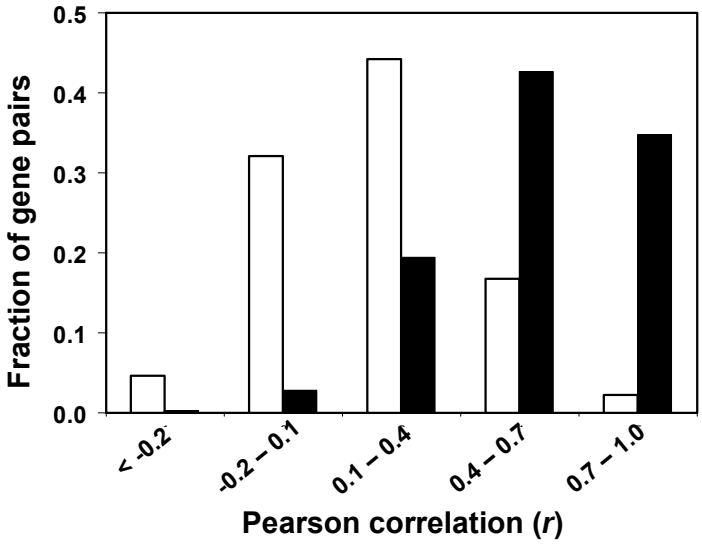

Figure 5.1. Comparison of Pearson correlation between transcript level of adjacent genes in KOPs and NOPs. (■) KOPs, (□) NOPs.

5.4.1.2 Functional similarity

Genes co-transcribed on an operon are often involved in the same biological pathway or function. Several metrics were used to evaluate the functional similarity between adjacent genes (see Methods). Pairs of adjacent genes were compared based on whether or not they belong to the same protein class. Among the 426 KOPs, both the adjacent genes in 291 pairs are assigned to a

functional class. Adjacent genes in 209 (72%) of these 291 pairs belong to the same class. In contrast, only 10% of adjacent genes in NOPs belong to the same class indicating the relevance of functional similarity as a characteristic for operon prediction (Figure 5.2a).

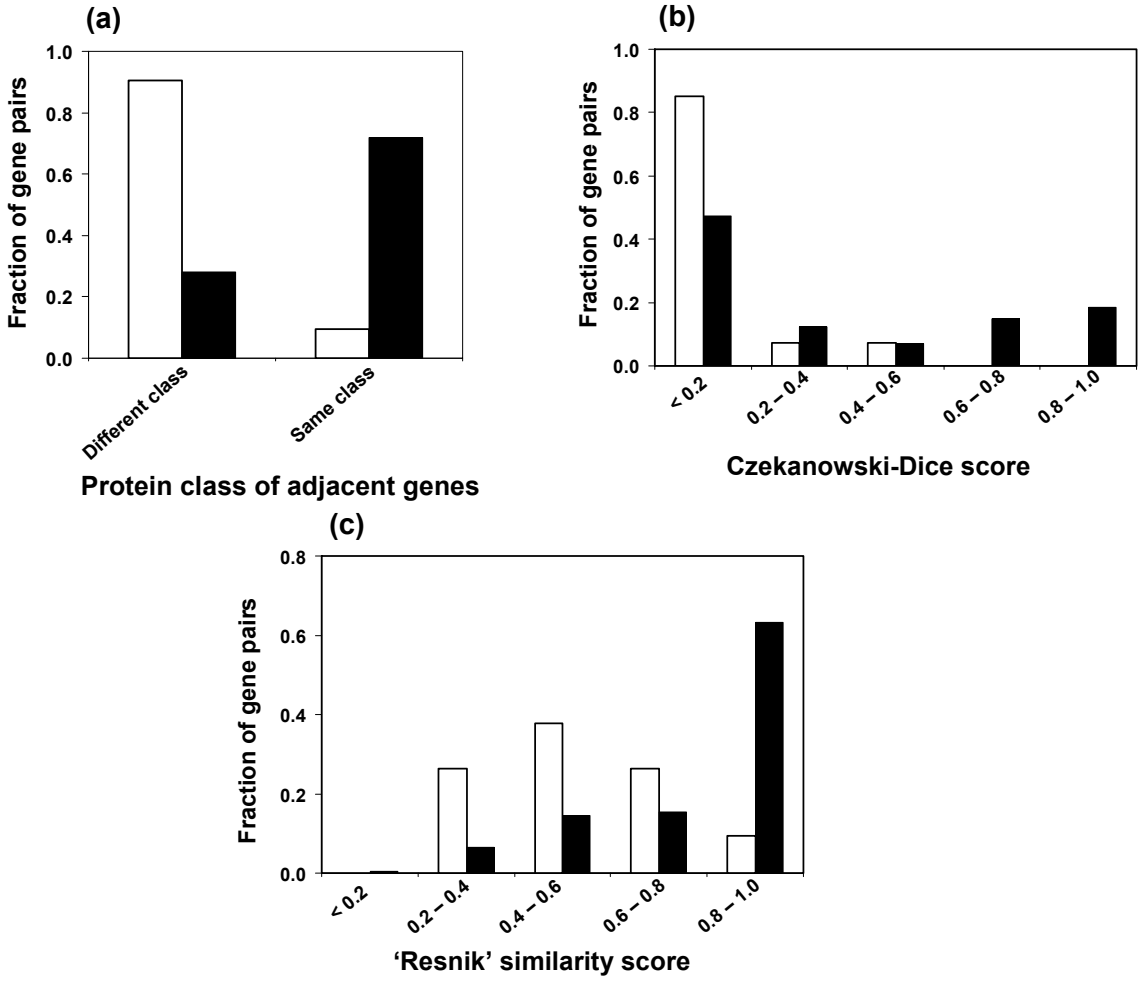

Figure 5.2. Comparison of functional similarity between adjacent genes in KOPs and NOPs. (a) Similarity (same or different) of the protein classes of adjacent genes, (b) Czekanowski-Dice score between adjacent gene based on the commonality of their GO terms, (c) Information theoretic similarity score (proposed by Resnik[273]) between adjacent genes based on their Biological Process GO terms. (■) KOPs and (□) NOPs.

Based on Gene Ontology (GO) classification, Czekanowski-Dice score was used to estimate the functional similarity between adjacent genes. In 269 of the 426 KOPs, each of the two adjacent genes is associated with at least one GO term. The Czekanowski-Dice score is greater than 0.6 for 90 (33%) of these 269 KOPs. However, the score is less than 0.6 for all the

NOPs. Further, the score is less than 0.2 for 85% of NOPs in contrast to 47% of KOPs (Figure 5.2b).

Lastly, since GO is a directed acyclic graph, we also examined functional similarity measures based on the information theory that account for the hierarchical structure of GO. Briefly, the hierarchy of a GO term is embedded in its degree of information. Every GO term is associated with an information content (*IC*) that is inversely proportional to the number of times that term or its direct or indirect child term appears in a functional database of genes. Thus, GO terms with broad functionality (e.g., GO:0009987, cellular process) at the top of the hierarchy have low *IC* (*IC* = 2.7 for GO:0009987), whereas GO terms lower in the hierarchy (e.g. GO:0006313, DNA-mediated transposition) that depict specialized functions, have higher *IC* (*IC* = 11.6 for GO:0006313). The similarity between two GO terms is the *IC* of their lowest common ancestor (also known as the minimum subsumer), which has the highest *IC* amongst all the ancestors of the two GO terms, as proposed by Resnik[273]. A gene however, is often associated with multiple GO terms. Thus, while comparing a pair of adjacent genes, the minimum subsumer of every combination of their GO terms was estimated. The maximum *IC* amongst these subsumers was assigned as the similarity between the two genes. Figure 5.2c shows the histogram of the 'Resnik' similarity scores between adjacent genes in KOPs and NOPs. Adjacent genes in 36% of NOPs have similarity score greater than 0.6. In contrast, 79% of KOPs have similarity score greater than 0.6, highlighting that GO-based functional information can be used for operon predictions. Modifications of the Resnik method proposed by Lin[274], and Jiang and Conrath[275] also displayed a similar discriminatory behavior between KOPs and NOPs. Similarity scores were estimated for GO annotations based on biological process and molecular function, which are two of the three organizing principles of Gene Ontology.

5.4.1.3 Conservation of gene order

Genes in the same operon are often conserved across multiple genomes. This feature has been previously used for operon prediction in several prokaryotes[186]. For every pair of adjacent genes in *S. coelicolor,* the number of bacterial genomes in which their orthologs are present in the same order was therefore used as a feature for operon prediction.

5.4.1.4 Other features

As described in Chapter 4, intergenic distance and the probability of the presence of a transcription terminator in the intergenic region of a pair of adjacent genes were also used as features for operon prediction.

5.4.2 BINARY CLASSIFICATION RESULTS

Binary support vector machine (SVM) classifiers were constructed for differentiating KOPs and NOPs using every feature individually as well as a combination of all the features.

5.4.2.1 K-fold cross-validation results

A 10-fold cross-validation scheme and ROC graphs were used to compare the performance of classifiers constructed from different features (Figure 5.3). Functional similarity measures based on Gene Ontology (Czekanowski-Dice metric and information theoretic metrics) are weak predictors of operons. At a false positive rate (FPR) of 25% they can accurately classify merely 45% of KOPs. Classifier based on protein classification scheme performs marginally better than the ones based on GO. At a FPR of 20%, it can classify nearly 58% of KOPs. Similarly, the classifier based on conservation of gene order can correctly identify 50% of KOPs at 25% FPR. The classifier based on intergenic distance outperforms all the classifiers based on functional similarity and gene order conservation. Interestingly, gene expression similarity based on transcriptome data is the single most predictive feature for operons with a recall of 82% at 25% FPR. The combination of all these features results in a vivid improvement in model predictability. At a FPR of 10%, the SVM model based all the features can accurately classify 92% of KOPs.

Table 5.2 shows a comparison of the AUC of all the classifiers. Transcriptome-derived classifier IV is the single most predictive feature with an AUC of 0.87, which is significantly better than the intergenic distance-based classifier III (p-value = 2.8×10^{-2}, paired t-test). The radial SVM model based on all the features, which has an AUC of 0.97, outperforms all the classifiers based on single features, including classifier IV (p-value = 1.6×10^{-4}, paired t-test).

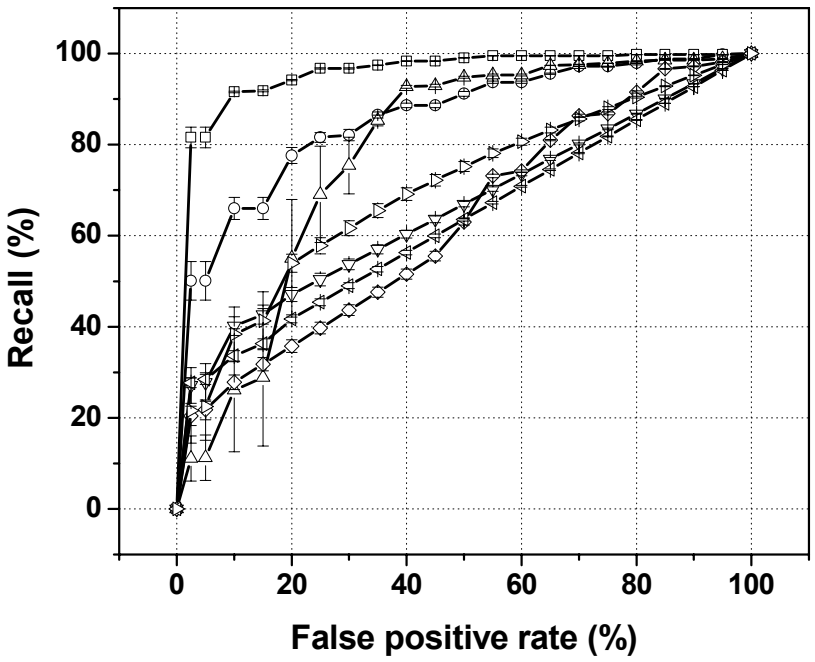

Figure 5.3. Comparison of different SVM classifiers by 10-fold cross-validation and ROC graphs. False positive rate is the percentage of NOPs misclassified as operon pairs. Recall is the percentage of KOPs correctly classified as operon pairs. (◁) Czekanowski-Dice functional similarity metric; (◊) Information theoretic functional similarity metric; (∇) Conservation of gene order; (▷) Functional similarity based on protein classification scheme; (Δ) Intergenic distance; (o) Gene expression similarity; (□) All features. Error bars indicate ± 1 standard deviation (*n* = 10).

Table 5.2. Comparison of the AUC of different classifiers. The null hypothesis was tested by comparing the AUC of 10 ROC graphs for each classifier by one-tailed paired *t*-test

No.	Feature(s)	Average AUC	*p*-value	Null hypothesis
I	Functional similarity			
	a. Czekanowski-Dice metric	0.65		
	b. Information theoretic metric	0.65		
	c. Protein classification scheme-based	0.72	2.3×10^{-2}	$AUC_{Ic} - AUC_{Ia} > 0$
II	Conservation of gene order	0.68	1.1×10^{-1}	$AUC_{II} - AUC_{Ia} > 0$
III	Intergenic distance	0.80	1.5×10^{-3}	$AUC_{III} - AUC_{Ia} > 0$
IV	Transcriptome	0.87	2.8×10^{-2}	$AUC_{IV} - AUC_{III} > 0$
V	All features	0.97	1.6×10^{-4}	$AUC_{V} - AUC_{IV} > 0$

5.4.3 A WHOLE-GENOME OPERON MAP

5.4.3.1 Pair-wise prediction of adjacent genes

The SVM model based on all features was used to predict the operon status of all 4965 same-strand pairs in the *S. coelicolor* genome. The model assigns a score (*s*) to every gene pair. A total of 2479 of the 4965 same-strand pairs have a positive score indicating that adjacent genes in those pairs were predicted to be co-transcribed. Gene expression similarity as well as intergenic distance correlates strongly the score of gene pairs. As the score increases, the gene expression correlation between adjacent genes in a pair increases (Pearson correlation = 0.64). Similarly, as the score increases, the intergenic distance between adjacent genes decreases (Pearson correlation = -0.53) indicating that adjacent genes in high scoring pairs are separated by a short intergenic distance.

5.4.3.2 Prediction of transcription units

The adjacent gene pairs with positive score were grouped into cistrons. A total of 5346 transcription units were predicted. Among them, 3957 are monocistronic, and the remaining 1389 operons have two or more genes. This distribution of cistron sizes is shown in Figure 5.4. Thirteen operons have 10 or more genes including the two largest 21-gene operons, *SCO0381-SCO0401* and *SCO4701-SCO4721*.

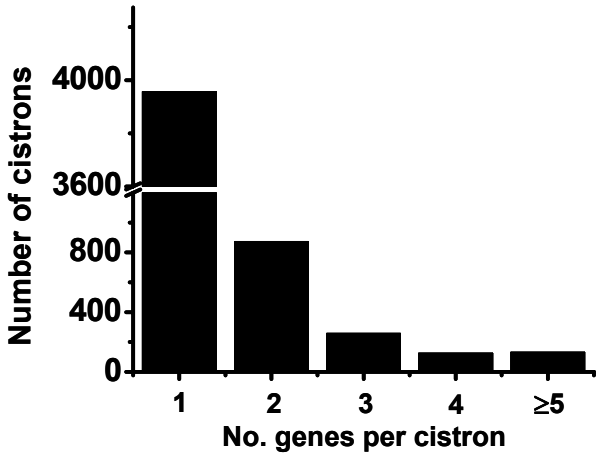

Figure 5.4. Distribution of cistron sizes

5.4.4 IDENTIFYING TRANSCRIPTIONAL INTERACTIONS IN *S. COELICOLOR*

We used ARACNe (algorithm for accurate reconstruction of cellular network)[284] to infer transcriptional interactions of the form 'A regulates B', where A encodes a transcription factor. The methodology of ARACNe is explained in the Methods section. Briefly, ARACNe compares the expression of every combination of two genes to identify the pairs with statistically significant and high mutual information. The premise of ARACNe is that genes with a regulatory linkage will exhibit a high degree of expression dependency or correlation. However, a high expression similarity between two genes does not necessarily translate into a direct interaction. For example, if the protein encoded by gene A activates gene B, and protein B triggers gene C, then gene A and gene C may have similar expression patterns. ARACNe employs the concept of data processing inequality (DPI) to eliminate potentially indirect interactions thereby reducing the number of false positive interactions.

We implemented ARACNe for *S. coelicolor* using the entire transcriptome dataset comprising 524 cell samples. Since adjacent genes are, in many cases, co-regulated as polycistrons, we utilized our predicted operon map to account for this additional level of regulation. Thus, we identify the interactions of the form A regulates B, where both A and B are cistrons. Cistron A contains at least one gene encoding a transcription factor or a putative DNA-binding protein.

5.4.4.1 Data preprocessing

As discussed in the previous chapter, a critical requirement for estimating expression similarity between any two genes is that the genes should exhibit sufficient expression dynamics. Lack of temporal dynamics or a high level of noise in expression profiles can increase the likelihood of predicting a false interaction. In order to reduce false positive discoveries, the transcriptome data was preprocessed to eliminate genes that do not exhibit a threshold level of dynamics in their expression patterns (see Methods). Among the 7825 genes in *S. coelicolor* genome, the selection criterion was satisfied by 6225 genes. Thereafter, the *z*-normalized expression values for adjacent genes in the same predicted cistron were averaged to obtain the mean expression pattern of every cistron. A total of 4399 cistrons were used for further analysis. Among the 6225 selected genes, 747 genes encode putative transcription factors (such as two-component systems, serine-threonine kinases, sigma factors, transcriptional regulators belonging to known families) and putative DNA-binding proteins. These putative regulators were mapped to

692 cistrons (more than one regulatory gene belongs to the same cistron in some cases). ARACNe was implemented on the expression patterns of 4399 cistrons, of which 692 regulatory cistrons contain at least one gene encoding a putative regulator.

5.4.4.2 Predicted transcriptional network

The mean expression pattern of every regulatory cistron was compared to the expression pattern of each of the 4399 cistrons using ARACNe to identify 7170 interactions between 3527 cistrons. The overall structure of the entire network is shown below (Figure 5.5). Each node in the network represents one cistron and every edge between the two nodes represents a potential regulatory interaction between a cistron encoding at least one putative transcription factor and another cistron (which may or may not encode proteins with regulatory functions). Note that every interaction is statistically significant with a p-value $< 1.0 \times 10^{-9}$.

We examined the number of interactions (called degree of connectivity, k) associated with every regulatory cistron or node. The global connectivity properties of the network are shown in Figure 5.6. The network displays a power-law relationship given by $p = 46.6 \times k^{-2.71}$ where, p is the probability that a regulatory node has k interactions. This is indicative of a scale-free network structure. A small fraction of the regulatory nodes (called 'hubs') are highly connected and they account for a large number of interactions. For example, the top four hubs, each of which interacts with at least 50 cistrons, account for 3% of all the predicted linkages (Table 5.3). Interestingly, three of these four hubs encode two-component systems that regulate gene expression by sensing environmental cues[30]. The average interaction strength (i.e. the average mutual information between a regulatory node and its direct targets) for all the interactions in the entire network is 0.137. It is noteworthy that for three of the four hubs, the average interaction strength is significantly greater than 0.137 (p-values $< 1.0 \times 10^{-4}$, Kolmogorov-Smirnov test) (Table 5.3). This further indicates substantial gene expression similarity between the regulatory hubs and the cistrons that they are predicted to regulate.

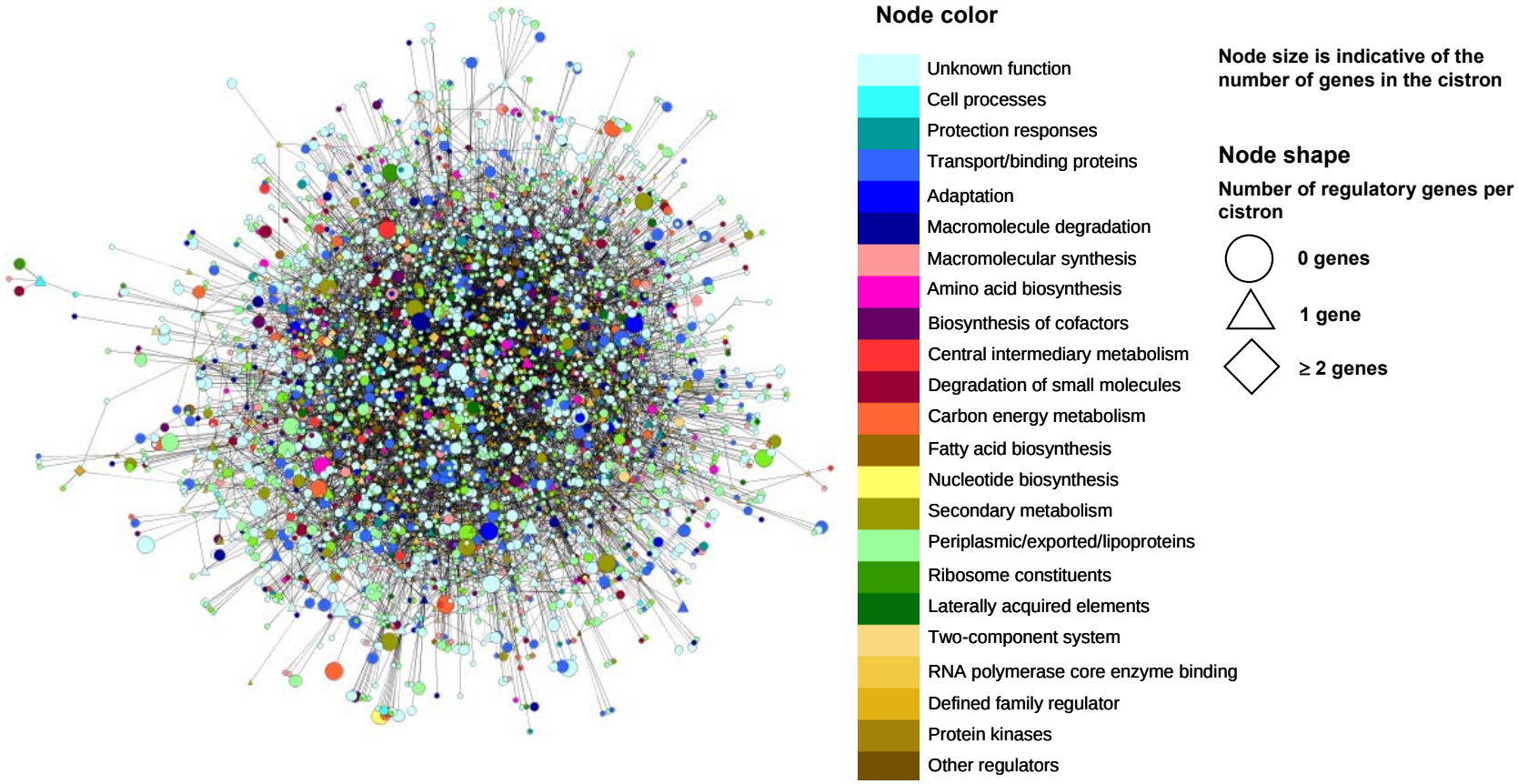

Figure 5.5. Predicted transcriptional regulatory network for *S. coelicolor*. Each node corresponds to a cistron and every edge represents a regulatory interaction between two nodes. The entire network comprises 3527 nodes and 7170 edges. Every node is colored by the primary function that is most common among the genes in that cistron. Node shape is indicative of the number of genes in the cistron that encode a putative transcription factor.

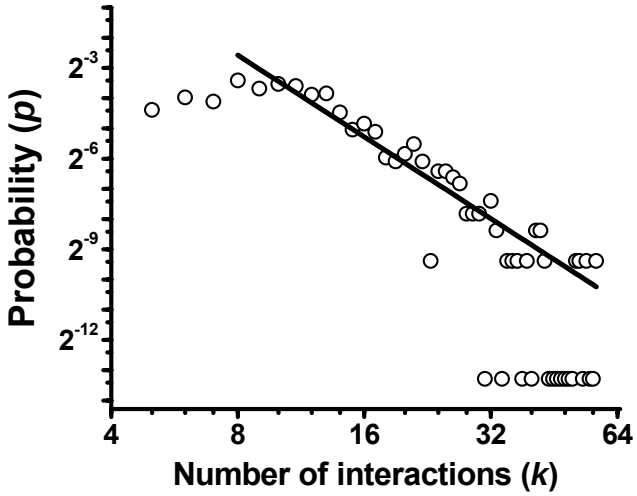

Figure 5.6. Global connectivity properties of predicted *S. coelicolor* transcriptional network. A power-law behavior is observable between the probability that a regulatory node is connected to *k* nodes and its degree of connectivity, *k*.

Table 5.3. Most inter-connected hubs in the predicted *S. coelicolor* transcriptional network

No.	Hub regulatory gene(s)	Gene function	No. predicted interactions	Average interaction strength (MI)[§]
1	*SCO0588*	Putative sensor kinase	54	0.212
2	*SCO3063*	Putative two-component system response regulator	51	0.137
3	*SCO3986*	Putative GntR-family transcriptional regulator	57	0.154
4	*SCO5454-5455*	Putative two-component system sensor kinase and response regulator	52	0.174

§ MI – Mutual information between the cistron encoding the regulator and its interacting partner

5.4.4.3 Identification of biologically relevant network modules

Each regulatory node and the cistrons directly interacting with the node comprise a network module. The entire network was dissected into 692 network modules. Seven modules, which

contain less than five genes, were discarded. We further examined whether the genes in the same network module are functionally-related. Among the 7825 genes in *S. coelicolor* genome, 4889 genes were categorized into 137 functional classes based on a protein classification scheme. Functional annotation based on Gene Ontology (GO) classification was also used. A total of 4560 genes were annotated with one or more GO terms. A total of 173 GO terms based on the three organizing principles, biological process, cellular component, and molecular function, were used. Fisher's exact test was used to identify the networks where a significant fraction of genes are associated with a particular functional class or GO term. At an enrichment *p*-value threshold of 1.0×10^{-4}, 146 unique modules are enriched with 161 functional classes (a few modules are enriched for more than one functional class). Similarly, 115 modules are enriched for 200 GO terms with an enrichment *p*-value less than 1.0×10^{-4}. Strikingly, even at a high threshold of 1.0×10^{-8}, 65 modules are significantly enriched with 67 functional classes and 26 modules are enriched with 53 GO terms.

5.4.4.3.1 Secondary metabolite-related network modules

Twenty-three modules are enriched with functions related to secondary metabolite and polyketide synthesis at a stringent *p*-value threshold of 1.0×10^{-8}. Three of the top ten modules are significantly enriched for genes related to the synthesis of the three known antibiotics in *S. coelicolor*. The modules for actinorhodin (Figure 5.7a), undecylprodigiosin (Figure 5.7b), and calcium-dependent antibiotic (CDA) (Figure 5.7c) are shown. The genes involved in the regulation and synthesis of an unknown type I polyketide (PK) are also distinctly enriched in one of the modules (Figure 5.7d). The central hub for each module is the cistron containing the gene encoding the pathway specific activator of the corresponding antibiotic gene cluster. These genes are *SCO5085* (*actII-ORF4*)[40-43] for actinorhodin, *SCO5877* (*redD*)[44, 46] for undecylprodigiosin, *SCO3217* (*cdaR*)[47] for CDA, and *SCO6280* (*cpkO*)[70] for type I PK. Many of the previously reported regulatory interactions were also identified in our analysis, as shown in Figure 5.7. For example, *scbA* (*SCO6266*) and *cpkO* (*SCO6280*) are regulated by the gene product of *scbR* (*SCO6265*)[70, 258]. The consistency between the predicted and experimentally known interactions provides further evidence that the analysis can successfully predict true positive interactions. Further, our approach identifies several novel interactions that are statistically significant. For instance, a potential interaction between *cpkO* and *SCO7192* (Figure 5.7d) was predicted by

ARACNe. With a mutual information of 0.25, the transcript profiles of these genes are highly correlated across the microarray dataset (Figure 5.8). *SCO7192* encodes a putative sigma factor with unknown function. It would be highly interesting to examine whether *SCO7192* plays a role in regulating type I PK through an interaction with its pathway activator *cpkO*. The number of false positive interactions identified by this approach is also likely to be non-negligible. For example, the generated network predicts a direct interaction between *scbR* (*SCO6265*) and *SCO6273-SCO6275* (*cpkABC*). However, the two nodes are known to interact through the intermediate node *SCO6280-SCO6281* (*cpkOH*)[70]. Nonetheless, network predictions based on a combination of temporal transcriptome data and functional characteristics can provide significant insights about the complex processes that regulate secondary metabolism.

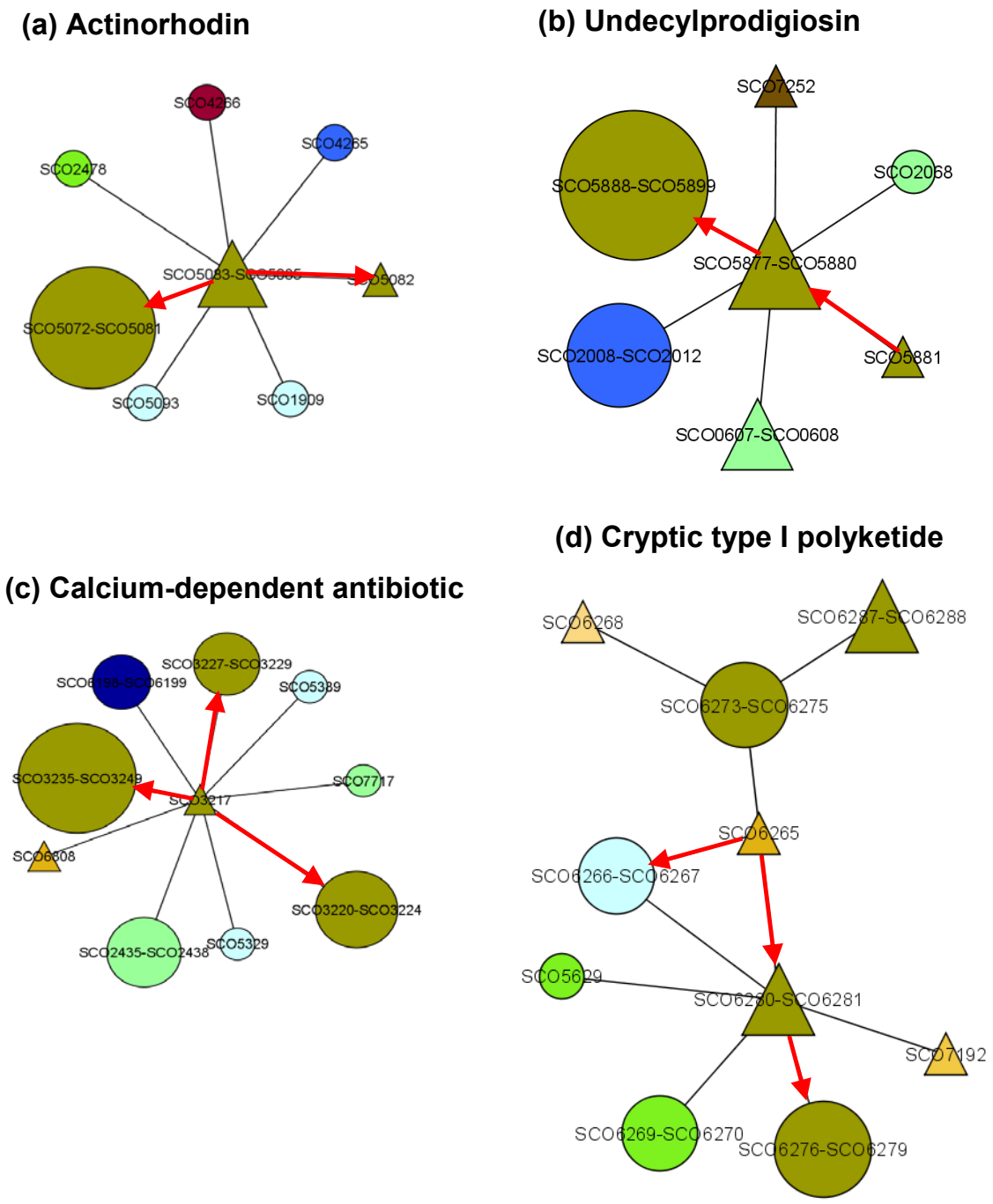

Figure 5.7. Network modules related to secondary metabolite synthesis. Each module is significantly enriched for a secondary metabolite shown. Previously reported experimentally verified interactions are highlighted by red arrows. The node color, size, and shape represent cistron function, number of genes, and number of regulatory genes, respectively, as described in Figure 5.5. (a) Actinorhodin (Enrichment *p*-value = 2.1×10^{-12} for polyketide synthesis); (b) Undecylprodigiosin (*p*-value < 10^{-12} for polyketide synthesis); (c) Calcium-dependent antibiotic (*p*-value = 8.6×10^{-12} for secondary metabolism); (d) Cryptic type I polyketide (*p*-value < 10^{-12} for secondary metabolism)

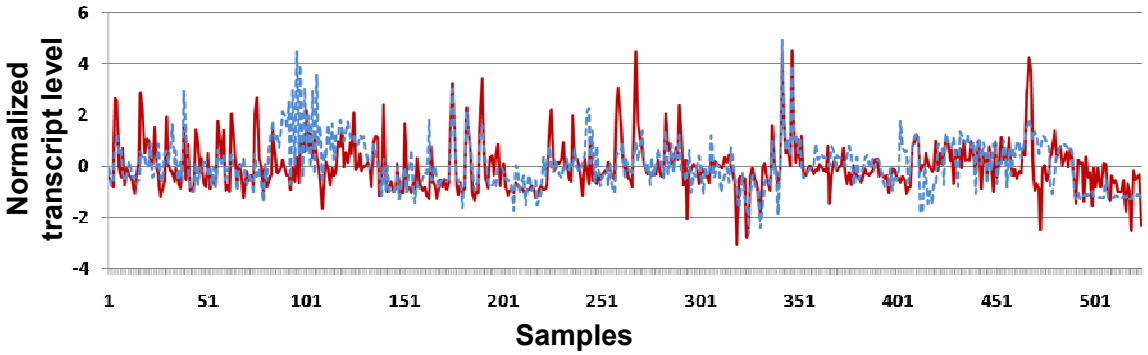

Figure 5.8. Identifying putative transcriptional interactions. The temporal profiles of *cpkO* (*SCO6280*) (red) and *SCO7192* (blue) across the entire microarray dataset.

5.4.4.3.2 Other biologically relevant modules

We also examined the predicted transcriptional network to identify modules that are enriched with genes involved in other cellular processes. One of the significant modules comprises many genes that participate in metabolism of nitrogen compounds (Figure 5.9a). The module includes genes encoding a nitrogen regulatory protein (*SCO5584/glnK*), putative ammonium transporter (*SCO5583/amtB*), glutamine synthetase (*SCO2210/glnII*), glutamine-hydrolyzing NAD^+ synthase (*SCO2238/nadE*), and nitrite reductase (*SCO2486-SCO2488/nirBCD*). The functional homogeneity of the module genes and the statistically significant similarity between their transcript profiles is suggestive of regulatory connections between elements of the module. Three other top scoring modules (based on enrichment *p*-value) are related to protein synthesis (ribosomal proteins) (Figure 5.9b), protein folding (Figure 5.9c), and electron transport (Figure 5.9d). Another three-cistron module is involved in phosphate metabolism (Figure 5.9e). The module comprises a two-component system PhoR-PhoP which senses depletion of phosphate in the growth environment. Upon activation, the phosphorylated response regulator PhoP activates the expression of the *pho* regulon[85]. The experimentally verified *pho* regulon includes the genes encoding the two-component system (*SCO4229-SCO4230/phoRP*), the phosphate ABC transporter (*SCO4139-SCO4142/pstSCAB*), and the phosphate transport system regulator (*SCO4228*/*phoU*). In this study, *pstSCAB* was successfully identified as a direct target of PhoP based on transcriptome and functional characteristics (Figure 5.9e).

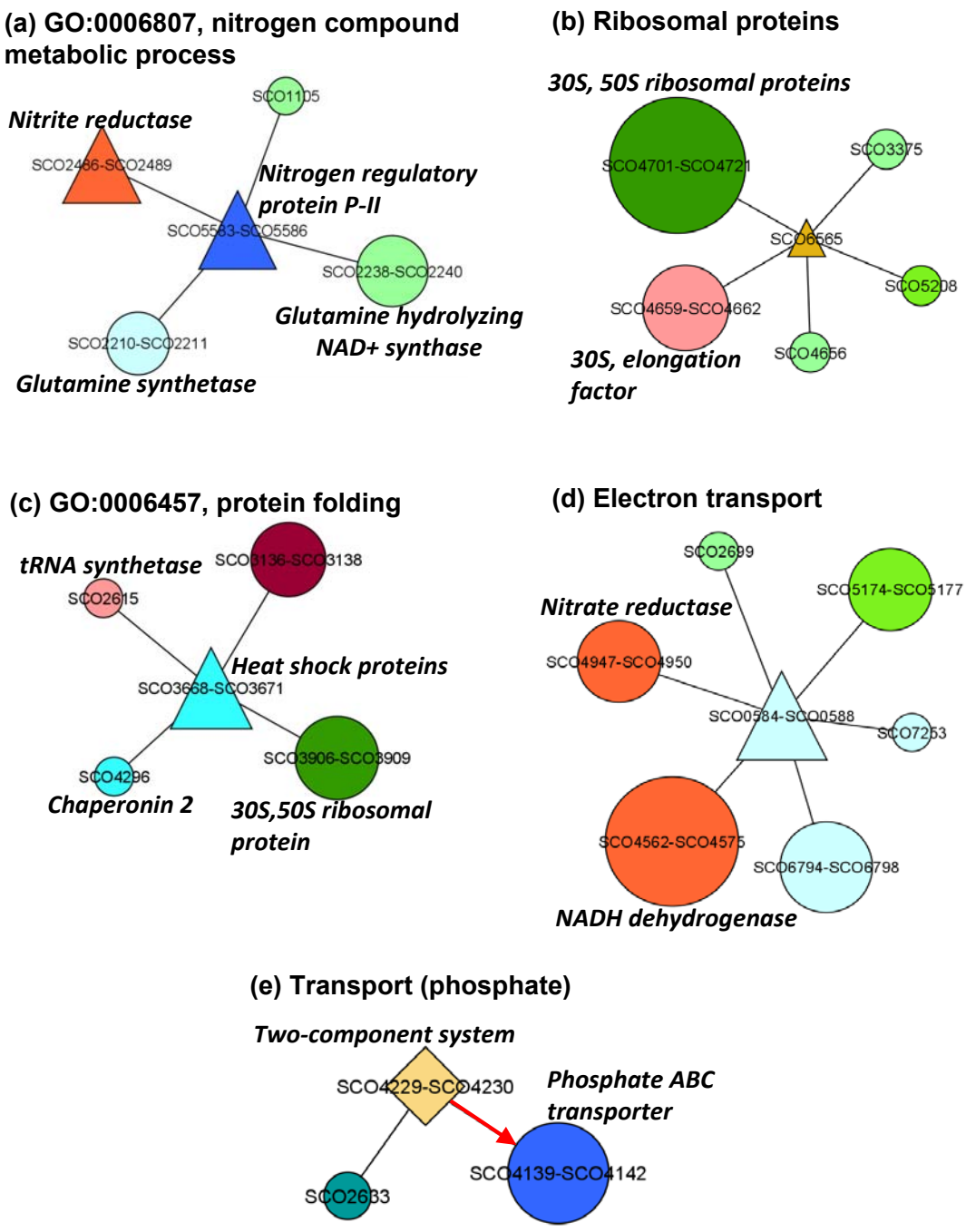

Figure 5.9. Other biologically relevant modules. Each module is significantly enriched for a functional class or a GO term shown. Previously reported experimentally verified interactions are highlighted by red arrows. The node color, size, and shape represent cistron function, number of genes, and number of regulatory genes, respectively, as described in Figure 5.5. Modules enriched for (a) GO:0006807 (p-value = 5.4×10^{-7}); (b) Ribosomal proteins (p-value < 10^{-12}); (c) GO:0006457 (p-value = 8.0×10^{-8}); (d) Electron transport (p-value = 1.1×10^{-12}); (e) Transport (p-value = 5.7×10^{-4})

5.5 DISCUSSION

Transcriptional control of gene expression is a fundamental process in prokaryotes. However, the gene regulatory circuits in *S. coelicolor* that manifest complex phenotypes such as morphological differentiation and biosynthesis of secondary metabolites are not well understood. Temporal gene expression profiles procured under diverse genetics and environmental perturbations can provide valuable information for deciphering the mechanisms of interplay between the regulatory genes and their ensuing effects on secondary metabolism. In addition to these regulatory cascades, groups of chromosomally contiguous genes are coordinately regulated in bacteria as a single transcription unit. In the previous chapter, a whole-genome operon map was constructed and a substantial fraction of the predictions were verified. Here, we refine the operon map using a more extensive feature set.

5.5.1 REFINING THE WHOLE-GENOME OPERON MAP

5.5.1.1 Features for operon prediction

5.5.1.1.1 Temporal transcript profiles

Transcriptome data derived from diverse experimental conditions can significantly improve operon predictions, as described previously (Chapter 4). All the publicly available temporal transcriptome data for *S. coelicolor* in Gene Expression Omnibus (GEO), ArrayExpress, and Stanford Microarray Database (SMD) at the time of this study was also included (Table 5.1). Based on this larger database of gene expression profiles comprising 524 cell samples, 50% of known operon pairs (KOPs) could be predicted with negligible false positive rate (FPR) (Figure 5.3) in contrast to an 8% FPR observed for the previous transcriptome dataset (Figure 4.4). Also, with an AUC (area under ROC graph) of 0.87 the transcriptome dataset used in this study outperforms the previous transcriptome-derived model (AUC = 0.81). Further, transcriptome similarity is the best feature for operon prediction amongst all the features used in this analysis (Table 5.2).

5.5.1.1.2 Limited predictability of features derived from functional similarity and conserved gene groupings

Genes in the same operon frequently participate in the same biological process and their orthologs are often conserved in many other prokaryotic organisms. Therefore, we extended our

feature set to include characteristics based on functional similarity, and the number of genomes in which a pair of adjacent and orthologous genes are conserved. With AUCs ranging between 0.65 and 0.72, the SVM models derived from these individual features offer small improvements compared to a random classifier (AUC = 0.5). This is partly due to the lack of sufficient information for estimating these feature values for a significant portion of training set. For example, GO-based functional similarity could not be estimated for adjacent genes in 167 (39%) KOPs due to lack of GO classification for the genes. Similarly, only 157 (39%) KOPs have chromosomally contiguous orthologs in other bacterial genomes. It was estimated that only 30-50% of gene pairs *E. coli* genome can be accurately predicted using this comparative feature[186].

Nonetheless, the integration of all these features results in a noticeable improvement in operon predictions. Ninety-two percent of KOPs can be accurately classified at a FPR of 10%. Further, this integrative model with an AUC of 0.97 represents a substantial improvement of our previous model (AUC = 0.91) (Chapter 4) that was predominantly based on intergenic distance and a smaller set of transcriptome data.

5.5.1.2 Comparison with previous predictions

The pair-wise predictions of the integrative SVM model for all the same-strand pairs were compared with our previous model predictions. Among the 4965 pairs, the predictions for 4448 pairs remain the same (i.e., a positive score or a negative score predicted by both models), indicating a good agreement between the two models. The gene pairs for which the predictions of the two models disagree include 112 gene pairs with $s_{new} > 0.5$ and $s_{prev} < -0.5$. Adjacent genes in these pairs were predicted to be co-transcribed by the new SVM model but not the previous model. Among these 112 pairs, the order of adjacent genes in 46 pairs is conserved in at least 10 other bacterial genomes indicating a high likelihood for co-transcription. The genes in additional 13 pairs have high expression correlation $r > 0.5$ (based on the complete microarray dataset). Among the remaining 53 gene pairs, the adjacent genes in 48 pairs have a high degree of functional similarity (i.e. they share the same functional class or they are largely associated with the same GO terms). Thus, almost all the 112 gene pairs display distinctly predictive characteristics in at least one of the features, further demonstrating that the current model can effectively integrate predictive features from heterogeneous sources. Further, only 7 gene pairs have $s_{new} < -0.5$ and $s_{prev} > 0.5$. These pairs were predicted to be co-transcribed by the previous

model only. This implies that the current and the previous models have substantial agreement in their 'negative calls' (i.e. predicting that adjacent genes in a pair are *not* co-transcribed).

5.5.2 REVERSE ENGINEERING TRANSCRIPTIONAL NETWORK OF *S. COELICOLOR*

The gene products of various cistrons typically function as a part of an interconnected network in the context of a biological process such as regulation of secondary metabolism. Onset of secondary metabolism, which is coordinated with morphological differentiation, is a probable omen for the impending nutrient starvation (e.g. phosphate limitation[84]). These events depict a major transition in the life-cycle of *S. coelicolor*. The regulatory circuits that control and regulate the onset of these critical events are also likely to be highly dynamic, interconnected as well as synchronized to ensure timely execution of these events. However, deciphering these networks is an arduous task for several reasons. First, the regulatory programs are large and are likely to engage several tens to hundreds of genes. Moreover, additional genes products that affect one or more antibiotics and other secondary metabolites are discovered frequently (e.g., DasR[82], SarA[82]). Second, the genes can be regulated at different levels of regulatory hierarchy, such as transcriptional, post-transcriptional, translational, and post-translational, as observed previously in several studies (discussed in Chapter 2). Third, extensive cross-talk between different secondary metabolic pathways has been observed[200]. Fourth, a statistical dependency between the expression profiles of two genes can result either from a direct or an indirect interaction. Identifying and eliminating the indirect interactions is a non-trivial task.

5.5.2.1 Scale-free topology of transcriptional network identified by ARACNe

ARACNe algorithm employed in this study identified pair-wise dependencies between the temporal expression profiles of two cistrons. Thereafter, the data processing inequality was used to eliminate several potentially indirect dependencies. Moreover, the use of cistron as the basic unit of regulation, rather than individual genes, is in accordance with the notion of a transcription unit in prokaryotes. Another advantage of this approach is that unlike clustering algorithms (such as *k*-means, self-organizing maps) where cistrons or genes are assigned to mutually exclusive groups, a cistron can participate in multiple subnetworks, thus allowing a gene to engage in different biological functions.

Based on the temporal transcriptome data from 524 samples across more than 25 different strains of *S. coelicolor*, ARACNe identified a transcriptional network with scale-free connectivity distribution. This network topology suggests that a small fraction of regulatory nodes are highly

connected with greater than 40 interacting partners. Scale-free architecture has been observed in networks derived from transcriptional interactions[284], metabolic reactions[286], protein-protein interactions[287], as well as Gene Wiki[288] (a recent open-source initiative for annotating gene functions). The probability p that a regulatory node has k interactions follows a power-law behavior. The predicted scale-free transcriptional network of *S. coelicolor* displays a power-law connectivity distribution $p \sim k^{-\gamma}$. The degree exponent γ is indicative of the network diameter (i.e., the shortest path between any two nodes). Most biological networks have a small diameter with degree exponent $2 < \gamma < 3$ (Barabasi and Oltvai[256]). For *S. coelicolor* transcriptional network γ is equal to 2.71, further highlighting that the network is ultra-small. Jeong *et al.*[286] have argued that the scale-free biological networks with small diameter are resilient to genetic and environmental perturbations thus offering potential growth and survival advantage to the organism.

5.5.2.2 Limitations of network inference based on transcriptome data

An implicit assumption of most reverse engineering approaches based on microarray data is that all the network components are fully observed. However, this assumption does not hold in the presence of additional layers of gene regulation. Although on a global level, mRNA abundance correlates with the protein levels of the corresponding genes, discrepancies between mRNA and protein profiles have been noted for several genes in *S. coelicolor*[91]. Moreover, due to post-translational modifications (e.g. phosphorylation of two-component systems), the active protein levels cannot be reliably estimated from transcript levels. These uncertainties introduce hidden variables which are not observed in transcriptomic studies. Due to this limitation of partial observability, it may be impossible to identify all the direct interactions and eliminate those interactions that arise due to indirect statistical dependencies[289]. It is conceivable that many of these regulatory predictions can be substantiated and improved by combining gene expression data with other genomic data sources such as functional annotation, *cis*-regulatory motifs, associations discovered by text-mining biomedical literature, and protein-protein interactions.

5.5.2.3 Combination of transcriptome-derived network inferences with functional annotation

In this study, causal interactions predicted between transcription factors and their target genes were further assessed by functional enrichment analysis to identify network modules with coherency of biological function. The entire network was dissected into 692 modules, where each

module comprises a regulatory node and all its predicted direct targets. Among the 692 network modules, 188 modules are significantly enriched for at least one protein functional class or GO term (Enrichment *p*-value < 1.0×10^{-4}). Further, 72 of these modules meet a highly stringent statistical criterion of functional enrichment (*p*-value < 1.0×10^{-8}) suggestive of their involvement in specific biological functions. Secondary metabolism and polyketide synthesis are the enriched functions in 23 of these 72 modules.

A vast majority of 7170 interactions predicted in this study are novel and not yet experimentally verified. Techniques such as chromatin immunoprecipitation (ChIP) and ChIP-on-Chip can be used to validate many of these causal interactions. However currently, genome-scale experimental data for protein-DNA interactions in *S. coelicolor* is unavailable. Prioritization of these predicted inferences will undoubtedly assist any future attempts to further analyze or verify these interactions. Computational tools for *cis*-regulatory motif discovery can be used in addition to functional enrichment analysis, to select promising network motifs and interactions as candidates for experimental studies. A recent report on operon prediction in *S. coelicolor* using computationally identified transcription factor binding sites suggests that valuable information can be drawn out of these additional sources[290].

5.6 CONCLUDING REMARKS

Streptomyces coelicolor, a soil-inhabiting prokaryote, with nearly 8000 genes, exhibits vastly dynamic patterns of gene expression. The dynamic life-style is reflective of its adaptability in an ever-changing soil environment to undergo differentiation and synthesize secondary metabolites. Genome-scale temporal transcriptome data obtained under diverse genetic and environmental perturbations can provide valuable cues for comprehending the regulatory mechanisms underlying these biological phenomena. Here, we implement a systematic approach for mining large volumes of transcriptome data to predict the transcription regulatory network of *S. coelicolor*. The network comprises more than 7000 direct associations between putative transcription factors and more than 3500 predicted cistrons in *S. coelicolor*. The network displays a scale-free architecture with a small-world property observed in several biological networks in bacteria as well as higher organisms. A substantial percentage of these interactions comprise network modules with coherency of biological function. Further attempts to integrate diverse genomic dataset will seek to improve the sensitivity and specificity of these network predictions.

Such integrative efforts substantiated with experimental validation present a highly promising systems approach for elucidating the regulatory determinants of secondary metabolism.

CHAPTER 6 MINING TRANSCRIPTOME DATA FOR FUNCTION-TRAIT RELATIONSHIP OF HYPER PRODUCTIVITY OF RECOMBINANT ANTIBODY

6.1 SUMMARY

In the past decade we have witnessed a drastic increase in the productivity of mammalian cell culture-based processes. High-producing cell lines that synthesize and secrete these therapeutics have contributed largely to the advances in process development. To elucidate the high productivity trait, the transcriptomes of eleven NS0 cell lines with a wide range of productivity were compared. Gene selection and pattern classification approaches were combined to learn support vector machines (SVM)-based predictive models to discriminate between high producers and low producers. Based on transcriptome-derived classification of high and low producers, gene set testing (GST) analysis was used to identify physiological functions that are altered in high producers. Three complementary tools for gene set testing – gene set enrichment analysis (GSEA), gene set analysis (GSA), and MAPPFinder, were used to identify groups of functionally coherent genes that are up- or down-regulated in high producers. Major functional classes identified include those involved in protein processing and transport, such as protein modification, vesicle trafficking, and protein turnover. A significant proportion of genes involved in mitochondrial ribosomal function, cell cycle regulation, cytoskeleton-related elements are also differentially altered in high producers. The observed correlation of these functional classes with productivity suggests that simultaneous modulation of several physiological functions is a potential route to high productivity.

6.2 INTRODUCTION

Many antibodies, including over a dozen that have been successfully introduced to clinical applications in the past decade are important therapeutics for cancer, arthritis and other diseases. Many of those antibodies are produced in quantities exceeding thousands of kilograms annually;

the increased demand has prompted increasing efforts in cell line and process development. Advances have led to a drastic increase in productivity in the past few years. By and large, the pursuit for higher productivity has been conducted by systematic yet empirical screening. The desire to better understand the complex trait of hyproductivity has led to a number of studies comparing the transcriptome[107, 291] or proteome[107, 112, 292, 293] of cell lines with varying productivities. With their large-scale surveying power, transcriptome and proteome analyses certainly hold promise for discerning the genotypic characteristics of hyper productivity.

In a previous study, the transcriptome profiles of seven high and four low recombinant IgG-producing NS0 cell lines were analyzed to identify genes which are significantly different between the two groups[107]. As in other studies, the high and low producers were classified heuristically based on final titer of IgG in culture. It was not clear whether all the producers share common features in their transcriptome. Nevertheless, through differential expression and functional analysis, it was highlighted that several genes related to protein synthesis and cell cycle were differentially expressed between high and low producers.

In this study, a different but complementary approach was undertaken for analyzing transcriptome data of high and low producers. Instead of analyzing the cell lines by heuristically defined high and low producer groups, a pattern recognition technique was used to classify the cell lines into distinct groups according to their transcriptome. We employed support vector machines (SVM) to develop supervised models to classify transcriptomes into 'high' or 'low' producer categories. SVMs have found increasing applications in transcriptome analysis due to their superior performance on genome-scale datasets compared to other discriminatory approaches[238].

Biological interpretation of genes that are differentially expressed (often referred to as 'expression signature') between two or more phenotypes is facilitated by grouping them into few functional classes. However, the list of differentially expressed genes depends on the stringency of the statistical threshold used. The difficulty in identifying genes which are truly differentially expressed is further compounded by the observation that for mammalian cells in culture, changes in productivity levels or even some culture conditions are accompanied by only modest alterations in the expression levels of individual genes[106-108, 293]. This is in contrast to gene expression changes seen in microbial populations or in stem cells in early stages of differentiation.

Gene set testing (GST) tools were used to assess gene expression alteration at functional class level rather than at individual gene level (for a recent review on GST, see[294]). In this approach, functionally-related genes are combined *a priori* into gene sets and transcriptome data is evaluated in terms of these gene sets instead of individual genes. The correlation of a gene set to a phenotype is evaluated by comparing the observed number of genes in a gene set, which have altered expression level with the expected number under a null hypothesis. Large percentage of genes in functions related to protein processing and secretion, such as Golgi apparatus, the cytoskeletal network, and protein degradation were altered between high and low producers. Differential transcript changes were also observed in cell cycle-related genes. The relatively modest changes at individual gene level between the two groups suggest that the expression changes are not localized but a broad range of functional modulation is likely to accompany the process of high producer selection during cell line development.

6.3 MATERIALS AND METHODS

6.3.1 CELLS AND SAMPLE PREPARATION

The eleven GS-NS0 cell lines and their cultural conditions and have been described previously[107]. The average productivity of the seven high producers is approximately five times the average productivity of the four low producers. In addition, biological replicates of four high producers (H1-H4) and two low producers (L1-L2) were performed under the same culture and sample preparations conditions, as reported previously.

6.3.2 MICROARRAY HYBRIDIZATION

GeneChip® Mouse genome 430A 2.0 (MOE430A 2.0) (Affymetrix, Santa Clara, CA) was used for assaying the transcriptome of the six biological replicate cultures. MOE430A 2.0 contains 22,690 probes representing approximately 14,000 well-characterized mouse genes. Biotinylated cRNA was prepared as per the protocol described in the Affymetrix Technical Manual. Fifteen micrograms of biotinylated cRNA was used for hybridization. The arrays were scanned at University of Minnesota Biomedical Image Processing Laboratory.

6.3.3 MICROARRAY DATA PROCESSING

The raw intensity data from each array was normalized using Affymetrix Microarray Suite (MAS) version 5.0, which includes background correction, perfect match (PM) adjustment, and calculation of expression summary from 11 probe pairs using one-step Tukey's Biweight method for estimation of robust mean. The probe intensities from each array were scaled to an average of 500. Further, a quantile normalization procedure was employed at probe level to ensure that probe intensities from different arrays have the same distribution. Using a one-sided Wilcoxon signed rank test, the MAS 5.0 algorithm also determines a 'detection' *p*-value for every probe. A *p*-value < 0.04 was used as the criterion to call a transcript 'present'. Transcripts with absolute intensity, averaged across all the samples, less than 60 were discarded before further analysis.

6.3.4 DIFFERENTIAL EXPRESSION ANALYSIS

Significance analysis of microarrays (SAM) version 3.0 was used to identify genes that are differentially expressed between the high producers and low producers[114]. SAM combines a *d*-statistic with repeated sample permutations to determine the percentage of genes that are identified as differentially expressed by chance, i.e., false discovery rate (FDR). A threshold of 10% FDR was used in this study. SAM outputs a *q*-value for every probe, which is an estimate of the FDR incurred when that probe, and all the probes with a lower *q*-value are called significantly differentially expressed. In this study, all the probes with *q*-value $\leq 10\%$ were considered as differentially expressed.

6.3.5 GENE SELECTION AND SUPPORT VECTOR MACHINES (SVM) CLASSIFICATION

A differential expression-based recursive gene selection approach was combined support vector machines (SVM)[202] to construct and evaluate models for binary classification of high and low producers. A linear kernel was used to estimate the similarity between two producers. A leave-one-out (*loo*) cross-validation scheme was employed. In each *loo* resample, ten producers were used as training set. Based on the transcriptome data of ten producers, every probe was ranked based on its degree of differential expression between the two classes using SAM. The probes were sorted based on increasing order of their differential expression established by their *q*-values. The top-ranking 500 probes were used to train a linear SVM classifier. The classifier was then used to assign a score to the 11^{th} test producer. In the next two iterations of gene selection, top-ranking 100 and 50 probes were used for SVM classification. In the next *loo*

resample, another combination of ten producers was chosen for training. Again, the probes were ranked based on differential expression metric, *q*-value, and the top-ranking 500, 100, and 50 probes were used to train SVM models. The models were used to calculate the score for the left-out producer. The *loo* resampling was repeated until every producer was used as a test object in one resample.

The scores of the test producers were subsequently normalized on a scale of 0 – 100, where 0 and 100 correspond to the producers with the lowest and highest scores, respectively. The scores of the producers were compared with their known class to evaluate the accuracy of the SVM classifier. A low score is indicative of a low producer and a high score indicates that the producer was correctly identified as a high producer.

6.3.6 FUNCTIONAL ANALYSIS

Gene set testing (GST) was performed on 242 gene sets to identify those that correlate with the phenotype distinction between the high producer and the low producer groups. A set of genes involved in the same biological function is defined as a gene set. Three different GST tools – MAPPFinder[295], gene set enrichment analysis (GSEA)[118, 119], and gene set analysis (GSA)[296], were used. For MAPPFinder, which is built into the software package Gene Map Annotator and Pathway Profiler (GenMAPP)[297], the criteria of *q*-value ≤ 10% and two different fold change thresholds (1.2 and 1.4) were used to identify differentially expressed genes as input. For each method, the null distribution was estimated by 1000 permutations. The enrichment of every gene set is characterized by a *p*-value. In this study, gene sets with a *p*-value ≤ 0.06 in at least two of the three GST methods were identified as significantly enriched. GSEA was used as a module in GenePattern[298] and GSA was available in SAM version 3.0[114].

6.4 RESULTS

In a recent study the transcriptome and proteome profiles of eleven GS-NS0 cell lines with different productivities were compared. The eleven cell lines were broadly classified into two groups as seven high and four low producers[107]. Several differentially expressed transcripts were involved in protein synthesis and cell cycle related pathways suggesting a direct or indirect correlation between those biological functions and high productivity. Here, we extend the

analysis by combining a pattern recognition approach with functional investigation measures to systematically explore the physiological traits that impart high productivity.

6.4.1 CLASSIFICATION OF PRODUCERS WITH DIFFERENT PRODUCTIVITY

The transcriptome data of eleven recombinant NS0 cell lines with different IgG productivities was examined using SVM to discriminate these producers into high producer or low producer groups. The methodology is explained in Materials and Methods. Briefly, all the individual transcripts are ranked according to their ability of distinguish the two groups. The top-ranking 500, 100, and 50 probes (n = 500, 100, 50) are then selected for SVM classification. The entire procedure was embedded in a leave-one-out (*loo*) resampling scheme for cross-validation.

The scores for all the producers, obtained from the *loo* procedure, are shown in Table 6.1. Four high producers (H1-H4) have scores above 80 for all the values of n. The score of each producer, averaged over different values of n, is greater than 85. In contrast, the average scores for the four low producers (L1-L4) are below 50. Thus, based on the transcriptome data, four of the seven high producers can be discriminated from the low producers. The average scores of the remaining three high producers (H5-H7) range between 0 and 35, suggesting that the distinction of their transcriptome from the low producers is not apparent.

Table 6.1. Scores for different producers based on SVM classification

Producer	Score for top-ranking n probes			Average score
	n = 500	n = 100	n = 50	
H1	82.5	84.3	93.8	86.9
H2	94.7	100.0	100.0	98.2
H3	93.1	95.8	85.7	91.6
H4	100.0	95.2	96.7	97.3
H5	0.0	0.0	0.0	0.0
H6	24.0	35.3	44.6	34.7
H7	11.6	8.7	19.4	13.2
L1	53.4	54.5	35.4	47.8
L2	27.4	25.9	42.4	31.9
L3	20.7	18.4	33.1	24.1
L4	8.3	19.2	17.3	14.9

6.4.2 FUNCTIONAL ANALYSIS

To discern possible physiological functions that confer the high productivity trait, we set out to identify functionally related genes that are altered between the two groups. Since the four high producers H1-H4 and the four low producers L1-L4 were placed into two groups based on transcriptome classification, they were used for further analysis. The three producers, H5-H7, whose transcriptome profile could not be classified distinctively as high producers, were excluded from the functional analysis.

Further, biological replicates of six of the remaining producers (H1-H4 and L1-L2) were performed and included in functional analysis to increase the confidence levels. M-A plots of various binary combinations of samples (arrays) were examined to ensure consistent normalization from replicates. No intensity-dependent bias in fold change was observed. Furthermore, inclusion of data from biological replicates did not change the membership of the four high and four low producers based on SVM classifiers.

Pathway analysis was performed using three different GST tools – MAPPFinder, GSEA, and GSA. A total of 242 MicroArray Pathway Profiles (MAPPs) for the mouse genome were obtained from GenMAPP website (http://www.wikipathways.org/index.php/Portal:GenMAPP). Each MAPP, a collection of genes involved in the same biological function or pathway, was considered as a gene set. The functional classes are categorized according to the organizing principles of Gene Ontology (GO) – cellular component, biological process, and molecular function.

Each of the three tools uses a different methodology to test the null hypothesis for every gene set. GSEA and GSA rank all the transcripts in the dataset using a class-correlation metric. Signal-to-noise ratio was used for GSEA, whereas GSA uses *t*-statistic as the class-correlation metric. Thus, for example, the genes which are upregulated in high producers are ranked high, whereas those that are downregulated in the high producers are ranked at the bottom of the rank-ordered list. The ranking scheme thus does not explicitly require a threshold for differential expression. Genes in a particular gene set or functional class are mapped on this ranked list. GSEA uses a Kolmogorov-Smirnov-like statistic for every gene set to test the enrichment of genes at the extremes of the ranked list[118], whereas GSA employs a maxmean statistic [296]. In contrast, MAPPFinder uses a list of differentially expressed genes as input. The overrepresentation of differentially expressed transcripts in a gene set is tested using the

hypergeometric distribution. A *z*-score computes the difference between the fraction of genes differentially up or downregulated in a gene set and the overall fraction expected in the population [295]. A gene set with a higher-than-expected proportion of differentially expressed genes has a high *z*-score and hence a higher likelihood of a correlation to productivity phenotype. The results of MAPPFinder are, however, dependent on the user-defined threshold for differential expression. To that end, we employed significance analysis of microarrays (SAM) [114] to identify differentially expressed genes. The transcriptome analysis on cultured mammalian cell in the past few years has generally demonstrated that the degree of differential expression observed is not of very large magnitude. Whether the hyper productivity trait is the manifestation of vast number of genes, each altering at a relatively minute level, or large but localized expression changes in a small number of genes, is still an open question. We thus employed two different criteria for differential expression of individual transcripts: (i) *q*-value ≤ 10% (10% FDR) and at least a 1.2 fold change, (ii) *q*-value ≤ 10% and a fold change of at least 1.4.

A distinction between GSEA, GSA and MAPPFinder is thus the reliance of MAPPFinder on a user-defined differential expression criterion. A potential drawback of this user-dependence is that the fraction of genes in a gene set that are called differentially expressed changes depending on the criterion, which in turn can affect the results of such a 'discrete' method [120]. Furthermore, change in the activity of a biological pathway can be effected by a modest change in a large number of genes involved in the pathway, many of which may not satisfy a stringent differential expression criterion. Modest changes can be identified more readily by quantifying the *shift* in distribution of a differential expression metric for a set of functionally-related genes, compared to the overall distribution for all genes[119, 120]. Gene set enrichment analysis (GSEA) proposes such a 'continuous' methodology whereby all the genes in a dataset are ranked according to a class-correlation metric, and enrichment of a gene set is based on the non-random positioning of its members in the ranked list[118, 119]. Gene set analysis (GSA) method proposed potential improvements to GSEA algorithm. GSA uses a different enrichment statistic and a modified procedure for estimating the null distribution[296].

The significance of each functional class or a gene set is characterized by an enrichment *p*-value. A low *p*-value for a gene set suggests that a significant fraction of transcripts in that set have altered expression levels between the two phenotypic groups. Since the three GST tools use different methodologies, the gene sets identified can also differ. The gene sets identified as

significant (p-values ≤ 0.06) by each method were cross-compared. Eight sets which were identified in at least two of the three methods are listed in Table 6.2. A positively enriched gene set is one in which many genes in that functional class are upregulated in the high producers. Similarly, in a negatively enriched gene set, genes downregulated in high producers are overrepresented.

The biological process of cell cycle (GO:0007049) was the only functional class identified as significantly enriched by GSEA and MAPPFinder (p-value ≤ 0.06) and marginally enriched by GSA (p-value = 0.062) (Table 6.2). Among the genes involved in cell cycle progression that are represented on the MOE430A array, 28% were differentially expressed, which includes 24 upregulated and 9 downregulated. Other functional classes that constitute different molecular functions such as isomerase activity (GO:0016853), structural constituent of ribosome (GO:0003735), GTPase regulator activity (GO:0030695), and ligase activity (GO:0016874) were also correlated to the phenotypic difference between the high and low producer groups. Golgi apparatus (GO:0005794), cytoskeleton (GO:0005856), and chromatin (GO:0000785) are the ontological classes under cellular component that were identified as altered between the high and low producer groups. In the ensuing sections, several of these functional classes are elaborated with emphasis on differentially expressed genes in each class

6.4.3 GENES ENRICHED IN HIGH AND LOW PRODUCER CLASSES

GSEA identifies a subset of genes (called leading-edge subset) in each gene set that are key contributors to the enrichment of the functional class. These genes reside at the top or bottom region of the rank-ordered list. A significant proportion of these leading edge genes also meet the criteria used to identify differentially expressed genes by SAM. In six of the eight gene sets, at least 60% of the genes in the leading-edge subset also satisfy the differential expression criteria (q-value $\leq 10\%$ and fold change ≥ 1.2). For every gene set, the percentage of gene probes that are represented on the microarray is also shown in Table 6.2. At least 19% of genes present on the array are differentially expressed in each functional class.

Table 6.2. Functional gene sets identified by different gene set testing methods

Gene set (Functional class)	No. genes in gene set	% present on MOE430A	% of genes D.E.[a]	Enrichment p-value for				Direction of alteration
				GSEA	GSA	MAPPFinder		
						DE. criteria		
						1[a]	2[b]	
Cellular component								
Golgi apparatus	336	69	19	0.150	0.053	0.035	0.008	Up
Cytoskeleton	189	68	20	0.027	0.042	0.231	0.119	Down
Chromatin	118	59	26	0.002	0.034	0.162	0.005	Down
Biological process								
Cell cycle	166	72	28	0.039	0.062	0.037	0.027	Up
Molecular function								
Isomerase activity	126	73	19	0.016	0.113	0.016	0.285	Up
Structural constituent of ribosome	164	77	19	0.069	0.059	0.024	0.331	Up
GTPase regulatory activity	174	66	20	0.045	0.106	0.014	0.028	Down
Ligase activity	160	67	27	0.078	0.057	0.021	0.173	Down

(a) D.E. criteria 1: q-value ≤ 10% AND fold change ≥ 1.2; (b) D.E. criteria1: q-value ≤ 10% AND fold change ≥ 1.4

These enriched genes and their extent of differential expression for different functional classes are listed in Tables 6.3-6.8 (listed at the end of this chapter). For each gene, the average hybridization intensity from the four high producers is also listed. To gauge the range of signal intensities from every array, the 25th, 50th, and 75th percentiles correspond to signal intensities of 49, 164, and 595, respectively. Following is a discussion on the possible role of these functional classes in conferring the hyper productivity trait. The focus is on the gene class rather than individual genes. A brief annotation on the functions of key genes is included in these tables.

6.4.3.1 Golgi apparatus

Golgi apparatus (GO:0005794) was identified as significantly upregulated in two of the three GST tools. Two hundred thirty-three genes present on MOE430A array are annotated as belonging to Golgi apparatus gene set according to the GenMAPP database. Genes in this functional class encode proteins that are involved in post-translational modification and protein trafficking and secretion. Among the 37 transcripts differentially expressed in this gene set by at least 1.4 fold (Table 6.3), as many as 27 are upregulated. All the 27 upregulated genes are in the leading-edge subset identified by GSEA further suggesting that upregulation of these transcripts in high producers plays a role in enhancing productivity. Some of these differentially expressed genes localized in Golgi can be clustered into categories of similar functions, notably vesicle transport and glycosylation. These categories are described below.

6.4.3.1.1 Vesicle transport

Ten genes, nine of which are upregulated, are related to vesicle transport. Protein processing in Golgi involves transport of cargos via membrane vesicles from one Golgi compartment to another as well as from Golgi to other cellular destinations. The correct delivery of membrane vesicles to their receiving targets is mediated by two complementary sets of transmembrane proteins: vesicle SNARE (*v*-SNARE) proteins and target membrane-specific SNARE (*t*-SNARE) proteins. *v*-SNARE on the membrane vesicle and *t*-SNARE on the target membrane interact to form a trans-SNARE complex that facilitates fusion of the vesicle to the target membrane. At least 30 different SNARE proteins are present in mammalian cells. Two of the nine upregulated transcripts encode the SNARE proteins, Gosr1 and Gosr2. The assembly of COPI-coated vesicles is initiated by activation of a small G protein, Arf1 on the Golgi membrane, followed by Arf1-mediated recruitment of the preassembled heptameric COPI coat complex.

GTPase-activating protein Arfgap1 stimulates GTP hydrolysis of Arf1 thereby contributing to COPI vesicle budding [299]. The transcript of Arfgap1 is upregulated by 1.5 fold in high producers. Interestingly, although the Arf1 transcript, with *q*-value of 25%, did not meet our statistical criterion, its expression level in each of the four high producers is ~1.2 fold higher than the average of low producers. Notably, the average signal intensity of Arf1 is 5629, which corresponds to the 98th percentile of intensities on every microarray. This suggests that Arf1 is one of the most highly abundant transcripts in the cell, and a 20% upregulation, if true, may alter cellular capacity for vesicle transport.

6.4.3.1.2 Protein glycosylation

Eight differentially expressed transcripts listed in Table 6.3 encode glycosyltransferase enzymes. These membrane-bound enzymes reside mainly in different Golgi compartments (*cis*, *medial*, *trans*) and catalyze the transfer of various sugar moieties to the newly formed protein transported from ER. The signal intensity for Man2a1 transcript is 2340 (93rd percentile) suggesting that the transcript is abundant and its upregulation by 70% in high producers is quite significant. However, not all the glycosyltransferase enzymes probed were upregulated. Genes encoding four glycosylation enzymes are downregulated by two-fold or greater in high producers (Table 6.3). There appears to be a change in the ratio of α-2,6-, α-2,8-, and α-2,3-sialyltransferase, and an enhanced mannosidase transcript level in high-producing cells. Alpha-mannosidase I and II are responsible for the trimming of high mannose glycan on glycoproteins before glycan extension. These are the first steps of glycan processing in Golgi, whereas sialations are the final steps of glycan synthesis for glycoproteins. It is also interesting to note that in high-producing cells the transcript of α-1,6-fucosyltransferase is two-fold lower. The product of fucosyltransferase, fucosylated *N*-glycan, has been reported to confer immunoglobulin G with lower antibody-dependent cellular cytotoxcity (ADCC) compared to the unfucosylated product[300]. It will be interesting to examine whether the glycoform of the high antibody producing cells indeed has altered glycans.

6.4.3.2 Cytoskeleton

Twenty percent of the genes in this functional category are differentially expressed. Members of the cytoskeleton class encode proteins that associate with one or more filamentous elements that form the cellular scaffold essential for maintaining cell shape, exerting and distributing mechanical force and various other functions such as exocytosis, endocytosis,

mitosis, and cell motility. They also play an essential role in intracellular protein transport, especially vesicle transport (reviewed in[301, 302]). Among the 23 genes that are differentially expressed in this functional class by at least 1.4 fold, eight genes that are downregulated (Table 6.4). The most significant is Hook homolog 2 (Hook2), which was not detected (detection *p*-value > 0.04) in any high producer in contrast to an average intensity of 646 in low producers. Hook proteins attach to microtubules at their N-terminal domains and the C-terminal domain associates with different organelles. Myo9b, which serves as a molecular motor to drive intracellular cargo on actin filaments, is also downregulated by 1.4 fold at transcript level. With the diverse functions that cytoskeletal elements are involved in, it is difficult to point the exact cellular functions that may have been altered. But combined with the altered functional classes identified, it seems logical to speculate that the change in cytoskeletal elements may be related to vesicle trafficking and protein secretion. Interestingly, a recent 2D-gel based proteomic investigation of four IgG_4-producing NS0 cell lines by linear regression analysis of functionally-related proteins also indicated a correlation between cytoskeleton-related proteins and antibody productivity[112].

6.4.3.3 Chromatin

The third cellular component ontological class – chromatin, comprises of proteins involved in packaging DNA to form the condensed chromosome. The transcript levels of a large percentage (26%) of genes in this gene set were altered between high and low producer groups.

Seventy genes from this functional class are present on the MOE430A array as annotated by the GenMAPP database, of which 16 are differentially expressed by at least 1.4 fold (Table 6.5). The most notable gene in this class encodes tripartite motif 28 (Trim28). Genes encoding proteins involved in transcriptional silencing, such as DNA methyltransferase 3A (Dnmt3a) and methyl-CpG binding protein 1 (Mbd1), were also downregulated in high producers by more than 1.5 fold. Four other differentially expressed genes in this class encode histone proteins, which form the building blocks of chromatin. Modifications of these proteins alter DNA accessibility thereby regulating cellular processes such as transcription and replication.

6.4.3.4 Cell cycle progression

Among the 119 genes involved in cell cycle progression that are present on MOE430A array, 29 were identified as differentially expressed with a *q*-value and fold change threshold of 10% and 1.4, respectively (Table 6.6). Among these, 21 are upregulated and 8 are downregulated. Thirteen of the upregulated genes encode products involved in mitotic (M) phase of cell cycle

progression. One of the highly expressed and upregulated genes encodes a component of anaphase-promoting complex/cyclosome (APC/C). APC/C associates with cell division cycle 20 (Cdc20) and other E1 ubiquitin-activating and E2 ubiquitin-conjugating enzymes during various stages of mitosis. APC/C regulates progression through M phase by 26S proteasome-mediated degradation of cyclin A and cyclin B, that arrest cell cycle. Ras-association domain family protein 1 (Rassf1), whose transcript is also upregulated in high producers by 2.2 fold, binds to Cdc20 during prometaphase to inhibit APC/C activity[303].

Cell growth and death is the outcome of a concerted effect of several exogenous and endogenous factors. Low growth rate and high cell viability during the production phase is desirable in a high-producing cell line. Previous attempts to tap cell cycle regulation have focused on inducing expression of factors such as cyclin-dependent kinase inhibitors that can arrest cell growth (reviewed in Seth *et al.*[2]). The upregulation of negative effectors of cell proliferation, Rassf1 and Mad2l1, which arrest G2/M phase progression, and Gmnn (Geminin), which inhibits transition from G2 to S phase, may reflect subtle changes in growth regulation in high-producing cells.

While these results suggest that transcript level alteration in several cell cycle-related genes is correlated with productivity, it is likely that several other unaccounted gene products also modulate cell proliferation at various levels of regulatory hierarchy. Gene set testing methods such as GSEA and GSA are motivated by the correlation of a pathway with one phenotype or another, i.e. the genes involved in a common biological function are upregulated in one phenotype with respect to another, or vice versa. Alteration of many physiological processes can be invoked by differential up or downregulation of several genes involved. However, regulation of cell cycle is a homeostatic balance of many positive and negative factors. For regulatory networks involving intricate interaction of positive and negative elements, the state of the functional class may not be easily identified as up- or down-regulated.

6.4.3.5 Ribosomal constituents

Several genes in this class have modest, albeit significant changes in expression level. Among the 14 genes differentially expressed by at least 1.4 fold, ten are upregulated and four are downregulated (Table 6.7). Eight of the ten upregulated transcripts encode mitochondrial ribosomal proteins. These mitochondrial ribosomal proteins are encoded in the nucleus and are responsible for translation of mitochondrial genes. Mitochondrial ribosome consists of a small

28S subunit and a larger 39S subunit. Among the eight upregulated transcripts, six encode 39S subunit proteins and two encode 28S subunit proteins. Although mitochondria harbor between 5-15% of eukaryotic proteome, few of the proteins are synthesized in mitochondria[304]. The mitochondrial genome has 13 protein-coding genes, which predominantly encode enzymes involved in the oxidative phosphorylation pathway for ATP synthesis[305]. These genes are not represented on the MOE430A array. Hence, the effect of the upregulation of a large number of mitochondrial ribosomal proteins on protein synthesis cannot be confirmed.

6.4.3.6 Ligase activity

The set of genes encoding enzymes involved in ligation was also enriched (Table 6.8). Among the 20 differentially expressed, eight are upregulated and twelve are downregulated. All the 12 downregulated genes were also identified in GSEA as members of the leading-edge subset. Although the genes in this functional category are involved in the molecular function of ligation, they are not all involved in the same biological function. However, among the 20 differentially expressed genes, eight are involved in ubiquitin ligation. The ubiquitin-proteasome system plays a major role in cellular protein turnover. The ubiquitin-mediated proteasomal degradation pathway is comprised of several steps during which the ubiquitin moiety is activated by attachment to ubiquiting-activating enzyme (E1) and transferred to ubiquitin-conjugating enzyme (E2). A third component, ubiquitin-ligase (E3) acts in association with E2 enzyme and binds to specific protein degrading signals on the target protein and a polyubiquitin chain is attached to the protein. This acts as a recognition signal for the 26S proteasome for protein degradation. Among the eight differentially expressed genes in this gene set, five encode E2 enzymes, and three correspond to E3 enzymes.

Three differentially expressed genes in this functional category encode tRNA synthetases, which are enzymes that read the trinucleotide sequence on the corresponding tRNAs and ligate the appropriate amino acid. The three differentially expressed transcripts encode tRNA synthetases corresponding to phenylalanine, glycine, and alanine. An additional five transcripts coding for tRNA synthases of leucine, glutamate, glutamine, and histidine are differentially expressed with fold change between 1.2 and 1.4 fold. Additionally, four differentially expressed genes with ligase activity encode acyl-CoA synthetases, two of which are upregulated. Members of this family of enzyme convert long chain fatty acids into fatty acyl-CoA esters. These enzymes are required for synthesis of cellular lipids and also fatty acid degradation.

6.5 DISCUSSION

6.5.1 IDENTIFICATION OF MOLECULAR SIGNATURE FOR PRODUCTIVITY TRAIT

There is a profound interest in understanding the foundation of biological variability in the productivity of recombinant mammalian cells. It is customary to classify producing cells into high or low productivity groups. However, productivity is not an exactly defined trait. At least three elements may affect the final titer of recombinant protein profoundly: (i) the specific recombinant protein secretion rate, (ii) the growth rate and the growth extent, and (iii) the duration of sustained viability upon reaching maximum cell concentration. A super producer may have acquired some elements of all those positive characteristics, whereas a moderate producer may have only some of those positive characteristics. Even within each functional characteristic, there exist multiple routes to achieve the same superior features. For example, an elevated energy metabolism can potentially be accomplished through enrichment of mitochondria in each cell, enhanced expression of selective genes in mitochondrial metabolism, etc. Furthermore, the range of productivity is likely to be a continuum, rather than an arbitrary cut-off of high and low-producing classes.

Given the heterogeneous nature of cells in the same productivity class and the relatively arbitrary nature of conventional classification of high and low producers, the transcriptomes of cells in the same class may not all share common features. Employing a differential expression-based gene selection and SVM classification approach, discriminatory features in the transcriptome of eleven cell lines of known productivity were selected and used to classify the cell lines into two groups. Predictions based on cross-validation suggest that four (H1-H4) of the seven high producers can be discriminated from the low producers. The observed exclusion of the three cell lines from the classified 'high' category reflects the differences in the transcriptome fingerprint of the various high producers. It is conceivable that those three cell lines had taken 'alternative routes' to achieve a high productivity that is reflected at transcriptome level. It is also possible that due to the small number of producers used to develop classifiers, a broad range of features of the high-producing population could not be attained resulting in misclassification of the three cell lines. It is worth noting that the average productivity of the four differentiable high producers is 50% higher than the average productivity of the remaining three high producers and 5.3 fold greater than the average productivity of the four low producers. Having a small sample

size is almost an inherent problem of comparative transcriptome analysis of high-low producers. All efforts in cell line development aim to generate high-producing clones; few low-producing lines are kept, let alone characterized. Even the number of high-producing cell lines selected for characterization is usually small. This is a drastic contrast to studies of different disease phenotypes, which often have a very large patient base with tens to hundreds of samples for identification of molecular signatures[142, 306]. Nonetheless, even though a model constructed with few producers may not offer high predictive ability, it can provide useful insights to the physiology that underlies the productivity trait.

6.5.2 SIGNIFICANCE TESTING FOR FUNCTIONAL ANALYSIS

In search of genes conferring hyper productivity through transcriptome data analysis, one invariably resorts to statistical testing by setting criteria of fold differential expression, false discovery rate, etc. More stringent criteria imply a lower risk of false positive calls. In seeking pivotal genes responsible for a complex trait, a question inevitably arises whether the trait is caused by colossal alterations in a small number of master genes or by minute variations globally distributed in many functional classes. Small changes in transcript levels can be physiologically significant, especially when many genes involved in the same pathway or functional class change simultaneously. We thus employed multiple methods for functional class analysis. Two of the three methods used do not rely on setting an explicit criterion to preselect differentially expressed genes. Rather a differential expression metric is used to rank all the genes in the dataset, and the distribution of ranks of a set of functionally-related genes is used to characterize alteration of a functional class.

The functional class analysis identified several classes in cellular component, biological process, and molecular function ontology as significantly changed, as described earlier. However, many differentially expressed genes that do not fall into one of the identified functional classes may also play a significant role. Genes which are differentially expressed but do not fall into one of the significant classes were further examined. It was noted that many are involved in protein trafficking (Table 6.9). Together with the leading edge genes in identified functional classes, an overall picture of changes observed can be broadly depicted in Figure 6.1a. Functionally, the leading edge genes in the three cellular component classes (Golgi apparatus, cytoskeleton and chromatin) are involved in protein synthesis, processing and transport. Spatially, the enriched genes in the classes related to biological processes and molecular functions are distributed in

cytosol, mitochondria, and endoplasmic reticulum. The three dimensions of functional class analysis (cellular component, biological process, and molecular function) are closely interconnected, as expected. Node 1 comprises members of transcriptional regulation that are present in chromatin functional class (Table 6.5). Genes involved in protein synthesis, namely the genes encoding ribosomal proteins and tRNA synthetases in cytoplasm (node 2) and mitochondrial ribosomal proteins (node 3) are differentially expressed between high and low producers (Table 6.7, 6.8). Nodes 4-8 depict functions related to protein processing and secretion. Nodes 4, 5, and 6 represent various components of protein transport mediated by COPII, COPI, and clathrin-coated vesicles, respectively. Node 7 describes constituents of Golgi apparatus including glycosyltransferase enzymes (Table 6.3). The cytoskeletal network (node 8) comprises actin and microtubule filaments and molecular motors that drive vesicular cargo from one organelle to another (Table 6.4). Lastly, node 9 depicts the machinery for protein degradation, which includes several members of ubiquitin-mediated protein degradation pathway (Table 6.8).

Nodes 4, 5, and 6, which describe vesicle-mediated transport in early secretion pathway, are shown in expanded form in Figure 6.1b. The four essential steps in this trafficking include cargo selection followed by vesicle assembly and budding, vesicle transport to the target organelle by cytoskeletal motors, tethering of vesicles to the target membrane, and the final step involves fusion of the vesicular and target membranes. COPII vesicles are required for ER to early-Golgi transport. Transcripts of several genes involved in COPII vesicle transport are differentially expressed (Table 6.9a). Sar1a, a small GTPase, which is activated by nucleotide exchange on the ER membrane, is crucial for vesicle assembly. Rab1 and Rab2 are monomeric GTPases that regulate tethering and fusion of COPII vesicles to early Golgi or ER-Golgi intermediate compartment (ERGIC). Vdp and Trappc5 are among several Rab effectors that interact with Rab1 to ensure specificity of vesicle tethering to target membrane. Both Vdp and Trappc5 are upregulated in high producers. According to a recent report, Vdp is also upregulated in sodium butyrate-treated mouse hybridoma cells (MAK) and CHO cells by 1.3 and 1.5 fold, respectively[108]. Interestingly, Vdp is also among the genes identified as Xbp1 targets during B cell differentiation[128]. The final step of vesicle fusion is facilitated by ER-to-Golgi SNARE proteins, Gosr1 (Table 6.3) and Sec22b. Upregulation of these protein trafficking components in high producers is consistent with an earlier report suggesting that ER to *cis*-Golgi transport is the rate-limiting step in mammalian as well as insect cells[307]. Several components of COPI vesicle-mediated intra-Golgi transport and retrograde transport from Golgi to ER (node 5) are also

differentially expressed at the transcript level (Table 6.9b). Of particular note are the genes encoding N-Wasp and Arhgap21 that coordinate actin assembly on COPI vesicles in an Arf1-dependent manner. Others members including components of vesicle tethering complex COG (Cog4 and Cog8), Arf1-activating protein (Arfgap1), and a SNARE protein (Gosr2) that facilitate COPI-mediated vesicle transport are also upregulated in high producers at the transcript level (Table 6.3). The transcript of Napg, which encodes the γ-SNAP protein involved in disassembly of SNARE complex for vesicle membrane recycling, is also differentially expressed (Table 6.3).

6.6 CONCLUDING REMARKS

In this study, pathway-level analysis was performed on a set of high and low producers to identify physiological functions that are differentially altered between the two groups. Gene set testing (GST) tools were employed to discern statistical significance at a functional level, rather than at individual gene level. Analysis based on GST indicates that several functional classes, including protein processing constituents in the Golgi apparatus, cytoskeleton-related, and cell cycle-related functions were altered. Taken together our data and previous reports on transcriptome and proteome analysis of high and low producers seem to suggest that a number of functional classes are involved in enhanced productivity, including protein processing, vesicle trafficking and cell growth regulation. The study of hyper productivity is likely to benefit from combining high-low producer comparisons with comparative investigations on culture conditions that increase productivity. As such results begin to accumulate in the near future we can expect our understanding of this complex trait to expand and our ability to direct cells towards hyper productivity to greatly advance.

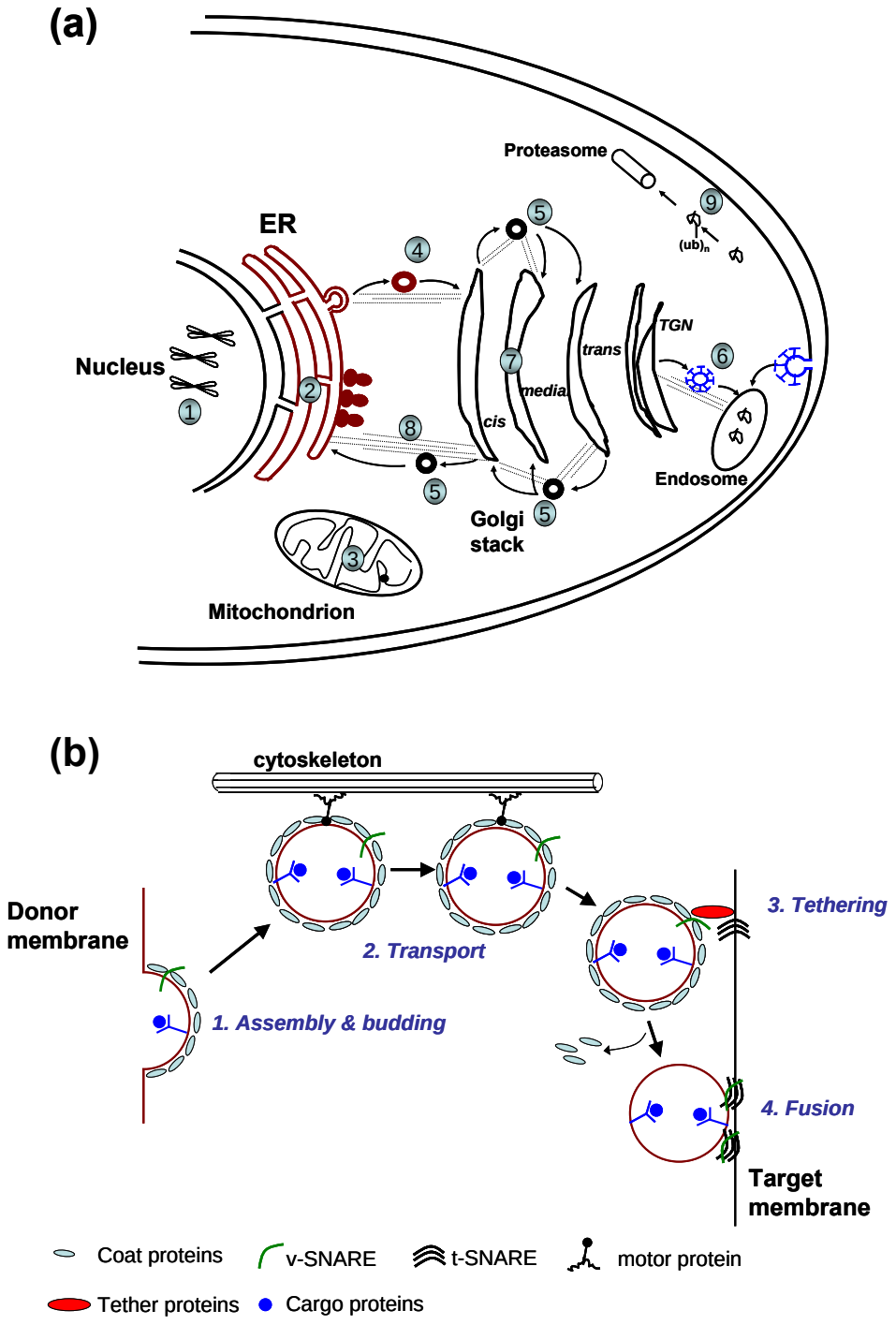

Figure 6.1. Genes differentially expressed between high and low-producing NS0 cell lines grouped according to intracellular function. (a) Each node depicts an intracellular process with large number of differentially expressed genes as identified by GST. (b) Schematic of the steps involved in vesicle-mediated transport (nodes 4, 5, and 6) (Modified from Cai *et al.*[308])

Table 6.3. List of differentially expressed genes in the functional class 'Golgi apparatus'

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	Vesicle transport					
Cog4	Component of oligomeric Golgi complex 4	(+) 2.3	0.48	311.7	Yes	A component of octameric COG complex, that plays a crucial role in tethering transport vesicles from late Golgi or early endosomes to *cis*-Golgi[309]
Ap1g1	Adaptor protein complex AP-1, gamma 1 subunit	(+) 1.9	0.34	666.9	Yes	γ-subunit of adaptor-related protein complex 1, which is important for formation of clathrin-coated pit on vesicles of *trans*-Golgi network (TGN).
Ap3s2	Adaptor-related protein complex 3, sigma 2 subunit	(+) 1.7	1.68	721.9	Yes	σ_2 subunit of clathrin-related adaptor complex 3 (AP3). AP3 complex likely to be associated with TGN and peripheral endosome-like structures involved in protein sorting[310]
Gosr1	Golgi SNAP receptor complex member 1	(+) 1.6	3.96	280.5	Yes	A v-SNARE that is a key component of Golgi 20S SNARE complex involved in vesicle transport from ER to *cis*- and *medial*-Golgi[311]
Copb2	Coatomer protein complex, subunit beta 2 (beta prime)	(+) 1.5	6.62	1149.5	Yes	A subunit of the coatomer protein complex that forms the coat of non-clathrin coated vesicles
Arfgap1	ADP-ribosylation factor GTPase activating protein 1	(+) 1.5	1.68	431.8	Yes	GTPase activating protein, which is involved in COPI vesicle-mediated protein transport between Golgi cisternae and retrograde transport from *cis*-Golgi to ER.
Cog8	Component of oligomeric Golgi complex 8	(+) 1.4	6.62	707.9	Yes	Component of COG complex that plays a crucial role in determining the protein transport capacity of Golgi
Gosr2	Golgi SNAP receptor complex member 2	(+) 1.4	3.10	396.0	Yes	A Golgi SNARE (GS27), which participates in protein transport from *medial*-Golgi to *trans*-Golgi and TGN[312]
Arfrp1	ADP-ribosylation factor related protein 1	(+) 1.4	1.68	243.4	Yes	
Napg	N-ethylmaleimide sensitive fusion protein attachment protein gamma	(-) 1.4	6.62	460.7	No	γ-isoform of N-ethylmaleimide-sensitive factor (NSF) attachment protein. Involved in disassembly of T-SNARE/V-SNARE/SNAP25 complex after vesicle fusion to facilitate their recycling[313].

	Protein glycosylation					
St3gal2	ST3 beta-galactoside alpha-2,3-sialyltransferase 2	(+) 2.0	0.00	313.1	Yes	Enzyme that can transfer sialylic acid residue to glycoproteins that have Galβ1,3GalNAc as the terminal disaccharide[314].
Man2a1	Mannosidase 2, alpha 1	(+) 1.7	5.47	2339.5	Yes	α-mannosidase II enzyme found primarily in *medial* Golgi. Catalyzes the cleavage of α(1,3) and α(1,6)-mannose residues from the high mannose glycan resulting in the formation of a core glycan structure - $glcNAc_2Man_3$, that is common to all *N*-glycans.
Man1a2	Mannosidase, alpha, class 1A, member 2	(+) 1.5	6.62	686.7	Yes	α-mannosidase IA, which cleaves α(1,2)-mannose residues from the high mannose oligosaccharide
Man1a	Mannosidase 1, alpha	(+) 1.5	3.09	696.3	Yes	
St6gal1	Beta galactoside alpha 2,6 sialyltransferase 1	(-) 2.4	5.47	204.2	No	
St8sia4	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	(-) 2.3	1.68	324.4	No	
Fut8	Fucosyltransferase 8	(-) 2.1	2.15	1198.5	No	Enzyme that can transfer fucose from GDP-fucose to first galactose of *N*-glycan through α-1,6-linkage
Galnt11	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 11	(-) 2.0	0.00	227.5	No	
	Others					
Sgpp1	Sphingosine-1-phosphate phosphatase 1	(+) 1.9	5.47	148.5	Yes	
Slc35a5	Solute carrier family 35, member A5	(+) 1.8	3.09	583.4	Yes	
Emid1	EMI domain containing 1	(+) 1.7	0.48	1147.8	Yes	
Nsg2	Neuron specific gene family member 2	(+) 1.7	8.25	1178.8	Yes	
Mbtps1	Membrane-bound transcription factor peptidase, site 1	(+) 1.7	0.48	1174.9	Yes	Involved in proteolytic activation of activating transcription factor 6 (Atf6) - a key transducer or ER stress. Mbtps1 also important for achieving cholesterol homeostasis by activation of two sterol regulatory element binding proteins (SREPB1 and SREPB2)[315]

Chst12	Carbohydrate sulfotransferase 12	(+) 1.6	5.47	336.7	Yes
Tmed2	Transmembrane emp24 domain trafficking protein 2	(+) 1.6	0.71	1694.2	Yes
Adam10	A disintegrin and metallopeptidase domain 10	(+) 1.5	2.15	1098.7	Yes
Golph3	Golgi phosphoprotein 3	(+) 1.5	5.47	1935.5	Yes
Gopc	Golgi associated PDZ and coiled-coil motif containing	(+) 1.5	6.62	238.0	Yes
Bicd2	Bicaudal D homolog 2	(+) 1.4	1.68	277.1	Yes
Golga4	Golgi autoantigen, golgin subfamily a, 4	(+) 1.4	6.62	291.0	Yes
Chrnb1	Cholinergic receptor, nicotinic, beta polypeptide 1	(+) 1.4	6.62	200.0	Yes
Gabarapl2	Gamma-aminobutyric acid (GABA-A) receptor-associated protein-like 2	(+) 1.4	2.15	2037.4	Yes
Aph1a	Anterior pharynx defective 1a homolog (*C. elegans*)	(-) 1.7	0.90	363.8	No
Lman2	Lectin, mannose-binding 2	(-) 1.6	1.24	798.3	No
Psen2	Presenilin 2 (PS2)	(-) 1.6	6.62	467.5	No
Clcn3	Chloride channel 3	(-) 1.5	8.25	1251.2	No
Rnpep	Arginyl aminopeptidase (aminopeptidase B)	(-) 1.4	3.10	387.3	No

* Fold change (F.C.), (+) Upregulated in high producer (H), (-) downregulated in high producer; $^{\varepsilon}$ *q*-value calculated for each probe using SAM; $^{\xi}$ Average intensity of high producers on Affymetrix array (MOE430A); § Core member of the functional gene set identified as the leading-edge subset in GSEA (applicable for Tables 6.3 – 6.9)

Table 6.4. List of differentially expressed genes involved in cytoskeleton function

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	Actin-binding					
Tpm2	Tropomyosin 2, beta	(+) 2.1	3.09	307.3	No	Members of an actin-binding protein family that stabilize actin filaments and regulate access to other actin-binding proteins
Tpm1	Tropomyosin 1, alpha	(+) 1.8	0.00	159.3	No	
Tmsb10	Thymosin, beta 10	(+) 1.8	3.97	457.0	No	
Coro1c	Coronin, actin binding protein 1C	(+) 1.7	1.68	207.6	No	
Shrm	Shroom	(+) 1.6	1.68	95.2	No	
Tmod3	Tropomodulin 3	(+) 1.6	3.96	866.1	No	
Rdx	Radixin	(+) 1.5	6.62	348.2	No	
Sntb2	Syntrophin, basic 2	(+) 1.4	0.71	238.8	No	
Hip1r	Huntingtin interacting protein 1 related	(+) 1.4	8.25	414.4	No	
Dst	Dystonin	(-) 1.9	3.09	270.0	Yes	
Myo9b	Myosin IXb	(-) 1.4	5.47	244.3	Yes	Encodes an actin-based motor
	Others					
Elmo1	Engulfment and cell motility 1	(+) 2.1	0.34	328.8	No	
Pxn	Paxillin	(+) 1.5	1.68	174.1	No	
Nf2	Neurofibromatosis 2	(+) 1.5	3.09	311.4	No	
Bicd2	Bicaudal D homolog 2	(+) 1.4	1.68	277.1	No	
Ctnnb1	Catenin (cadherin associated protein), beta 1	(+) 1.4	3.97	1167.3	No	An adherens junction protein that mediates cell-cell communication. Also interacts with TCF/LEF family of transcription factors to activate cyclin D1 transcription for G1/S phase transition[316]
Sirt2	Sirtuin 2	(+) 1.4	3.96	595.8	No	A tubulin deacetylase that regulates exit from the mitotic phase of the cell cycle [317]
Hook2	Hook homolog 2 (Drosophila)	(-) 17.4	0.00	37.2	Yes	A Hook protein which attaches to microtubules at its N-terminal domain and the C-terminal domain associates with different organelles
Jak2	Janus kinase 2	(-) 1.9	0.00	245.0	Yes	
Arpc5l	Actin related protein 2/3 complex, subunit 5	(-) 1.9	0.00	682.3	Yes	
Add3	Adducin 3 (gamma)	(-) 1.7	0.00	582.8	Yes	

Sspn	Sarcospan	(-) 1.5	8.25	161.4	Yes
Ptpn21	Protein tyrosine phosphatase, non-receptor type 21	(-) 1.4	3.96	186.4	Yes

Table 6.5. List of differentially expressed genes in the gene set 'Chromatin'

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	Histones					
Hist3h2a	Histone 3, H2a	(+) 1.7	5.47	180.6	No	
H2afy3	H2A histone family, member Y3	(-) 2.4	6.62	139.7	Yes	
H2afy	H2A histone family, member Y	(-) 2.0	6.62	1690.8	Yes	
H2afz	H2A histone family, member Z	(-) 1.5	0.28	7553.1	Yes	
	Transcriptional regulation					
Trim28	Tripartite motif protein 28	(-) 3.1	0.00	2322.4	Yes	Member of TRIM family of transcription factors that negatively regulate transcription from RNA polymerase II promoters
Dnmt3a	DNA methyltransferase 3A	(-) 2.6	2.15	114.4	Yes	Involved in *de novo* methylation of DNA
Mbd1	Methyl-CpG binding domain protein 1	(-) 1.6	0.40	1373.2	Yes	Member of methyl-CpG-binding domain proteins that interact with histone deacetylases to form transcriptional repressor complexes.
	Others					
Hmgb1	High mobility group box 1-like	(+) 2.1	3.97	1809.8	No	
Cbx2	Chromobox homolog 2 (Drosophila Pc class)	(+) 1.7	6.62	213.2	No	
Baz1b	Bromodomain adjacent to zinc finger domain, 1B	(+) 1.7	2.15	1040.1	No	
Cbx8	Chromobox homolog 8 (Drosophila Pc class)	(+) 1.5	0.71	276.7	No	
4930548 G07Rik	RIKEN cDNA 4930548G07 gene	(+) 1.5	8.25	127.2	No	
Smarcc 1	SWI/SNF related, actin dependent regulator of chromatin, subfamily C, member 1	(+) 1.4	8.25	869.0	No	
Suv39h 2	Suppressor of variegation 3-9 homolog 2 (Drosophila)	(+) 1.4	8.25	363.3	No	
Asf1b	ASF1 anti-silencing function 1 homolog B *(S. cerevisiae)*	(-) 2.0	0.00	591.4	Yes	
Cbx1	Chromobox homolog 1 (Drosophila HP1 beta)	(-) 1.6	8.25	154.2	Yes	

Table 6.6. List of differentially expressed genes involved in cell cycle progression

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	G2/M phase transition					
Rassf1	Ras association (RalGDS/AF-6) domain family 1	(+) 2.2	0.71	703.8	Yes	Binds to Cdc20 to inhibit the activity of anaphase-promoting complex/cyclosome (APC/C) - a large multisubunit E3 ubiquitin ligase, during prometaphase of mitosis [303]. Also inhibits cellular progression from G1 to S phase by post-translational inhibition of cyclin D1[318]
Anapc5	Anaphase-promoting complex subunit 5	(+) 1.8	0.90	2043.3	Yes	Component of the APC/C complex. In active form, from prometaphase to telophase, APC/C promotes degradation of proteins, such as cyclin A and cyclin B, that arrest cell cycle
Mad2l1	MAD2 (mitotic arrest deficient, homolog)-like 1 (yeast)	(+)1.7	2.15	2156.0	Yes	A component of the mitotic spindle checkpoint, that binds to Cdc20 and APC/C during metaphase to inhibit APC/C ligase activity[319, 320]
Chfr	Checkpoint with forkhead and ring finger domains	(+) 1.7	2.15	445.4	Yes	Functions as a checkpoint for entry into mitotic phase of cell cycle
Stag1	Stromal antigen 1	(+) 1.6	3.09	554.6	Yes	
Smc4l1	SMC4 structural maintenance of chromosomes 4-like 1	(+) 1.6	1.24	1489.7	Yes	
Rad21	RAD21 homolog *(S. pombe)*	(+) 1.6	2.15	1512.8	Yes	
Rbl1	Retinoblastoma-like 1 (p107)	(+) 1.5	1.24	216.9	Yes	
Chek1	Checkpoint kinase 1 homolog *(S. pombe)*	(+) 1.5	5.47	871.0	Yes	
Nek2	NIMA (never in mitosis gene a)-related expressed kinase 2	(+) 1.5	8.25	543.0	Yes	
Cdc23	CDC23 (cell division cycle 23, yeast, homolog)	(+) 1.5	8.25	435.8	No	

Sirt2	Sirtuin 2 *(S. cerevisiae)*	(+) 1.4	3.96	595.8	Yes	
Sept11	Septin 11	(+) 1.4	1.67	362.6	Yes	
Spag5	Sperm associated antigen 5	(-) 1.5	3.09	621.5	No	Involved in regulation of mitotic spindle apparatus
	Others					
Ccnd2	Cyclin D2	(+) 1.8	0.00	1054.9	Yes	Interacts with cyclin dependent kinases Cdk4 and Cdk6 for G1/S phase transition
Ccnc	Cyclin C	(+) 1.7	8.25	130.3	Yes	
Mapk6	Mitogen-activated protein kinase 6	(+) 1.7	3.09	753.2	Yes	
Gmnn	Geminin	(+) 1.7	0.48	4444.5	Yes	Accumulates during S and G2 phases and inhibits DNA replication by interacting with Cdt1, a replication initiation factor[321]
Calm2	Calmodulin 2	(+) 1.6	5.47	3099.9	Yes	
Ccng2	Cyclin G2	(+) 1.6	3.96	698.3	Yes	
Calm3	Calmodulin 3	(+) 1.4	5.47	685.3	Yes	
Cdc7	Cell division cycle 7 *(S. cerevisiae)*	(+) 1.4	0.90	500.1	Yes	
Siah1a	Seven in absentia 1A	(-) 1.8	0.40	152.2	No	Encodes a E3 ubiquitin ligase involved in protein degradation
Lzts2	Leucine zipper, putative tumor suppressor 2	(-) 1.6	6.62	269.5	No	
Ahr	Aryl-hydrocarbon receptor	(-) 1.5	6.62	71.3	No	
Pdcd4	Programmed cell death 4	(-) 1.5	3.09	671.5	No	
Cspg6	Chondroitin sulfate proteoglycan 6	(-) 1.4	3.96	1343.9	No	
Txnl4	Thioredoxin-like 4	(-) 1.4	8.25	160.9	No	
Chaf1b	Chromatin assembly factor 1, subunit B (p60)	(-) 1.4	6.62	272.9	No	

Table 6.7. List of differentially expressed gene in the functional class 'Structural constituent of ribosome'

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	Mitochondrial ribosomal proteins					
Mrpl1	Mitochondrial ribosomal protein L1	(+) 1.6	3.96	547.9	Yes	These genes encode components of the large 39S subunit of mitochondrial ribosomes
Mrpl37	Mitochondrial ribosomal protein L37	(+) 1.6	6.62	1531.3	Yes	
Mrpl52	Mitochondrial ribosomal protein L52	(+) 1.5	3.09	841.2	Yes	
Mrpl19	Mitochondrial ribosomal protein L19	(+) 1.5	5.47	885.2	Yes	
Mrpl38	Mitochondrial ribosomal protein L38	(+) 1.4	3.09	113.8	Yes	
Mrpl27	Mitochondrial ribosomal protein L27	(+) 1.4	1.24	2263.0	Yes	
Mrpl44	Mitochondrial ribosomal protein L44	(-) 1.5	0.40	1497.2	No	
Mrpl43	Mitochondrial ribosomal protein L43	(-) 1.4	2.15	1343.7	No	
Mrps23	Mitochondrial ribosomal protein S23	(+) 1.6	5.47	952.2	Yes	These genes encode components of the small 28S subunit of mitochondrial ribosomes
Mrps5	Mitochondrial ribosomal protein S5	(+) 1.5	5.47	522.8	Yes	
	Others					
Nola2	Nucleolar protein family A, member 2	(+) 1.5	3.09	5299.2	Yes	
Rps10	Ribosomal protein S10	(+) 1.4	8.25	114.4	Yes	
Rpl5	Ribosomal protein L5	(-) 1.8	0.00	253.0	No	
Mrp63	Mitochondrial ribosomal protein 63	(-) 1.4	6.62	132.2	No	

Table 6.8. List of differentially expressed genes involve in the functional class 'Ligase activity'

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)[ε]	Avg. Int.[ξ]	Core subset[§]	Annotation
	Ubiquitin ligation					
Ube2s	Ubiquitin-conjugating enzyme E2S	(+) 1.4	8.25	6047.6	No	These genes encode members of ubiquitin-conjugating enzyme (E2) family
Ube2d2	Ubiquitin-conjugating enzyme E2D 2	(-) 1.7	0.00	546.2	Yes	
Ube2j1	Ubiquitin-conjugating enzyme E2, J1	(-) 1.6	3.96	362.8	Yes	
Ube2f	Ubiquitin-conjugating enzyme E2F (putative)	(-) 1.5	0.90	578.7	Yes	
Ube2t	Ubiquitin-conjugating enzyme E2T (putative)	(-) 1.4	8.25	1121.6	Yes	
Wwp2	WW domain containing E3 ubiquitin protein ligase 2	(+) 1.4	1.68	422.6	No	These genes encode members of E3 ubiquitin ligase family
Siah1a	Seven in absentia 1A	(-) 1.8	0.40	152.2	Yes	
Wwp1	WW domain containing E3 ubiquitin protein ligase 1	(-) 1.5	3.09	234.9	Yes	
	tRNA synthetases					
Aars	Alanyl-tRNA synthetase	(+) 1.5	0.34	1915.5	No	
Farsla	Phenylalanine-tRNA synthetase-like, alpha subunit	(-) 1.7	1.24	265.8	Yes	
Gars	Glycyl-tRNA synthetase	(-) 1.4	6.62	1827.8	Yes	
	Acyl-CoA synthetases					
Acsl6	Acyl-CoA synthetase long-chain family member 6	(+) 1.7	8.25	101.3	No	
Acsl1	Acyl-CoA synthetase long-chain family member 1	(+) 1.6	3.96	189.2	No	
Acss1	Acyl-CoA synthetase short-chain family member 1	(-) 1.5	3.09	156.0	Yes	
Acsl5	Acyl-CoA synthetase long-chain family member 5	(-) 1.4	3.09	2753.8	Yes	
	Others					
Brap	BRCA1 associated protein	(+) 1.8	2.15	146.3	No	
Lig3	Ligase III, DNA, ATP-dependent	(+) 1.8	1.67	386.7	No	
Chfr	Checkpoint with forkhead and ring finger domains	(+) 1.7	2.15	445.4	No	
Gclm	Flutamate-cysteine ligase , modifier subunit	(-) 1.6	2.15	1454.4	Yes	
Rnf14	Ring finger protein 14	(-) 1.4	3.09	667.5	Yes	

Table 6.9. Differentially expressed genes involved in early secretion pathway at nodes 4 and 5

(a) ***Node 4: COPII vesicle-mediated ER-to-Golgi transport***

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)$^{\varepsilon}$	Avg. Int.$^{\xi}$	Annotation
Sec31	SEC31-like 1 (*S. cerevisiae*)	(+) 1.5	3.96	1138.0	Component of the Sec13-Sec31 heterotetramer which is the outer structural layer of COPII coat
Trappc5	Trafficking protein particle complex 5	(+) 1.3	3.09	167.8	A subunit of heptameric TRAPPI tethering complex that activates Rab1 through GTP exchange[322]
Rab1	RAB1, member RAS oncogene family	(+) 1.3	6.62	5875.2	A member of the Rab family of monomeric GTPases. The GTP-bound form of Rab1 regulates tethering and fusion of COPII vesicles to the target membrane[323]
Sar1a	SAR1 gene homolog A (*S. cerevi*siae)	(+) 1.3	10.44	2604.1	A small GTPase that is recruited to the ER membrane, which in turn recruits several components of COPII coat for vesicle formation and cargo selection
Vdp	Vesicle docking protein p115	(+) 1.2	10.44	1229.8	A coiled-coil tethering protein that is recruited to COPII coat in a Rab1-dependent manner and interacts with a subset of v-SNAREs to promote vesicle tethering and fusion[323]
Sec22b	SEC22 vesicle trafficking protein homolog B *(S. cerevisiae)*	(-) 1.6	0.40	353.0	*Sec22b* encodes a t-SNARE protein for ER-Golgi transport
Rab2	RAB2, member RAS oncogene family	(-) 1.2	6.62	1442.0	Another member of Rab family that is essential for vesicle-mediated transport from ER to Golgi or pre-Golgi compartments[324]

(b) ***Node 5: COPI vesicle-mediated intra-Golgi transport and retrograde transport from Golgi to ER***

Gene Symbol	Gene Title	F.C.*	*q*-val. (%)	Avg. Int.$^{\xi}$	Annotation
Arf3	ADP-ribosylation factor 3	(+) 2.0	0.48	211.2	A member of the Class I family of Arf proteins that regulate assembly of several coat proteins including COPI and clathrin coats
Wasl	Wiskott-Aldrich syndrome-like (human) (N-WASP)	(+) 1.5	0.90	334.1	N-WASP regulates actin assembly on COPI vesicles by stimulating the Arp2/3 (actin-related protein 2/3) complex[325]
Arf2	ADP-ribosylation factor 2	(+) 1.4	10.44	167.5	Another member of class I family of Arf proteins
Arhgap21	Rho GTPase activating protein 21	(+) 1.3	6.62	825.7	Serves as a GTPase-activating protein for the Rho-family GTPase, Cdc42, which in turn activates N-WASP for actin polymerization

CHAPTER 7 MINING CELL CULTURE PROCESS DATA TO UNVEIL HIGH PRODUCTIVITY CHARACTERISTICS

7.1 SUMMARY

Many recombinant proteins today are manufactured in modern production plants with sophisticated process control strategies, which significantly impact the growth and productivity characteristics of mammalian cells. Historical archives of these processes present numerous avenues to discover the subtle and hidden associations between process attributes and cellular physiology that determine the performance and robustness of bioprocesses. Here, we present a systematic effort to integrate vast volumes of on-line as well as off-line temporal process data from production 'trains' to develop predictive models for the final titer of the recombinant protein product. Support vector regression models based on data from 30 productions runs demonstrate that process information from inoculum bioreactors and the early stages of production-scale reactors provides valuable cues for process outcome.

7.2 INTRODUCTION

The worldwide demand for therapeutic proteins has increased drastically in the last two decades. The ever-increasing demand for these biologics has been met, in part, by engineering mammalian cells for increased specific productivity (reviewed elsewhere[2]). Since the capital costs of a new production facility can be prohibitively high, improving the production capacity of existing facilities and sustaining the robustness of the cell culture bioprocess is extremely critical. Robustness of a mammalian cell culture process is determined by product quantity as well as product quality, which among other factors is governed by post-translational modifications such as glycosylation. During the past decade, advances in process monitoring and control strategies have resulted in enormous improvements in process performance[326]. Fed-batch processes have become a popular choice for commercial manufacturing of recombinant proteins (reviewed in Wlaschin and Hu[327]). By employing strategies for controlled nutrient feeding, the product titer has

increased by almost a 100-fold in the past two decades. However, despite the process engineering attempts, run-to-run variability in process outcome is not uncommon.

Most manufacturing facilities commissioned in the past few years are highly automated. Several tens to hundreds of process parameters are routinely acquired and archived electronically, not only at the production scale, but also during cell expansion in the inoculum 'train'. Process data from a single production run can result in millions of data points. The challenges and opportunities associated with mining these process datasets to identify high productivity traits were highlighted in Chapter 3. Despite previous attempts, scrutinizing vast volumes of production-scale process data and on-line implementation of such schemes remains an arduous challenge. Recent advances in machine learning techniques present unprecedented opportunities for investigating the under-utilized resources of process archives to decipher the distinguishing features of a high productivity process. In this study, we analyzed vast volumes of off-line and on-line cell culture process data from acquired during production runs at Genentech's recombinant protein manufacturing facility at Vacaville, CA. Process data from 30 runs were scrutinized to investigate the causes of fluctuations in process outcome and to identify the distinguishing characteristics of high productivity processes. We propose an adaptable data mining framework to integrate off-line and on-line temporal process data to construct support vector regression models that can predict process outcome and identify critical parameters that shed insights on process productivity.

7.3 METHODS

7.3.1 DATA PREPROCESSING

The smaller of two manufacturing buildings at Genentech's Vacaville facility comprises cell culture bioreactors at scales ranging from 20L to 12000L. The recombinant mammalian cells for production are expanded from the cell bank to 20L scale and then step-wise to the manufacturing scale at 12000L. At each of these scales, cells are expanded for approximately 75 hours in each bioreactor scale. The cells are cultured at 12000L scale for approximately 11 days. In this study, process data of 30 runs, from 400L, 2000L, and 12000L scale bioreactors was analyzed.

The antibody titer at the end of a process run was measured and normalized for analysis. The titers of the 30 runs distribute over a range (Figure 7.1).

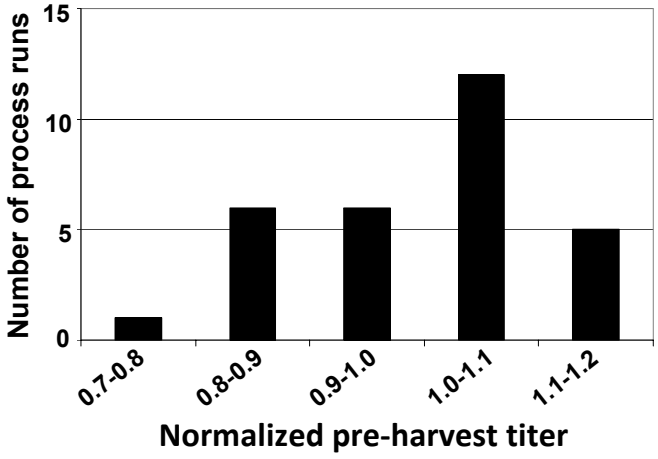

Figure 7.1. Histogram of the normalized pre-harvest titer of thirty production runs

7.3.1.1 On-line parameters

The manufacturing facility is equipped with fully automated control and data logging systems whereby acquired process data are recorded and archived on-line electronically. The number of parameters acquired on-line at the three different scales of 400L, 2000L, and 12000L are 130, 145, and 158, respectively. The on-line parameters include control parameters and control action parameters. The former category includes parameters such as dissolved oxygen (DO), pH, and vessel temperature that are controlled at specific levels (e.g., Temperature control at 37^{o}C), whereas the latter category includes parameters such as controller responses, the sparge rates of air and oxygen to control DO, and the rates of base addition and carbon dioxide sparge to control pH. Other important parameters such as, vessel volume and overlay gas flow rates, are also acquired on-line. The volumetric oxygen uptake rate (OUR) is estimated approximately every four hours, whereas all other on-line parameters are acquired more frequently (seconds or minutes) over the entire duration of the run that lasts several days. In addition to these parameters whose values are continuous, there are 'discrete' parameters such as the state of different valves, which is often binary (OFF/ON state). These valves control different ports for addition of inoculum, media, base, antifoam, and gas sparging among others. As many as 40, 48, and 55

parameters related to states of different valves are recorded on-line at 400L, 2000L, and 12000L, respectively.

On-line acquired data were preprocessed using a moving window average method. A time window of 100 minutes was selected. At every time point, a parameter value was approximated as the average of all the measurements for the parameter within the time window. For instance, the processed value at time *t* is the average of measurements at time *t, t+1*, *t+2*,…, *t+99* minutes. The raw and the preprocessed temporal profiles of CO_2 sparge rate, and DO controller output the 12000L scale of one run are shown in Figure 7.2a as examples. The preprocessed profile delineates the temporal patterns of the parameters without the disturbances at the local timescales. On-line parameters at all the three bioreactor scales were preprocessed by the moving window average method.

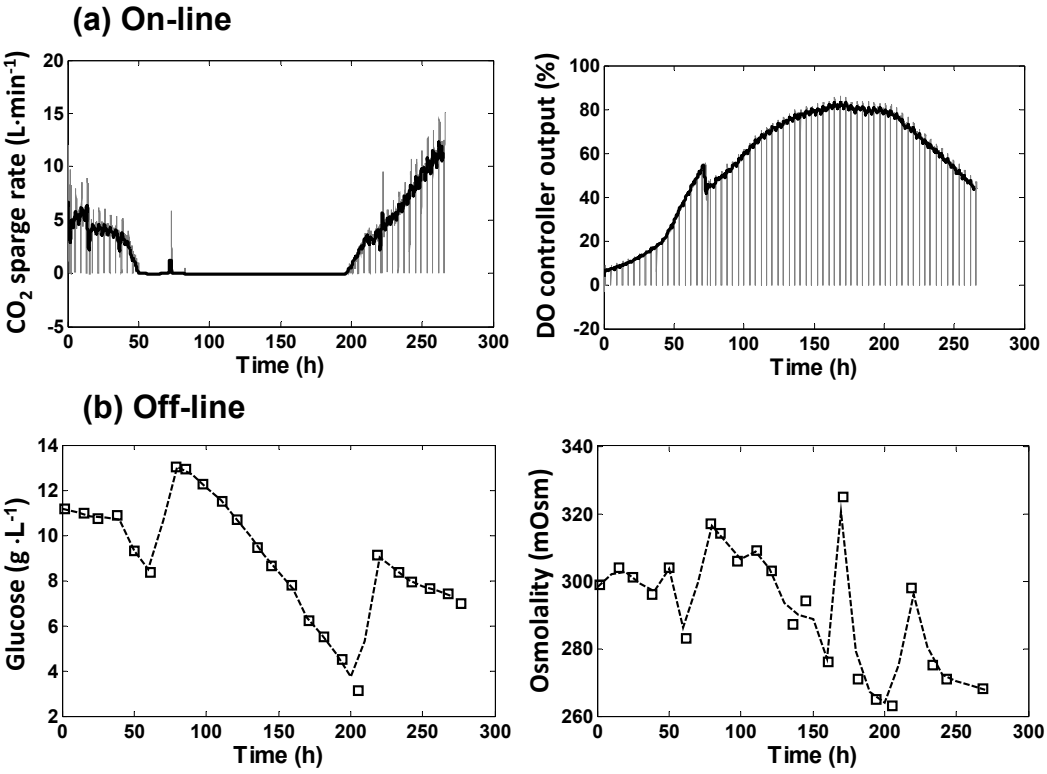

Figure 7.2. Preprocessing of cell culture process data. (interpolated data. (a) On-line parameters were preprocessed by a moving window average method (see Methods). The panels display the temporal profiles of carbon dioxide sparge rate (left panel) and DO controller output (right panel) from a 12000L fed-batch culture. (——) measured, (——) preprocessed. (b) Off-line parameters were preprocessed by a linear interpolation scheme. The panels show the temporal profiles of glucose concentration (left panel) and medium osmolality (right panel) from a 12000L fed-batch culture. (□) measured, (---) preprocessed.

7.3.1.2 Off-line parameters

A number of parameters related to nutrient consumption and metabolite production are measured off-line by periodic withdrawal of samples from the bioreactor (Table 7.1).

Table 7.1. Summary of process parameters at different bioreactor scales

Off-line parameter	On-line parameter
Physical and state parameters	***Controlled parameters***
Dissolved carbon dioxide	Dissolved oxygen (primary probe)
Dissolved oxygen	Dissolved oxygen (secondary probe)
Vessel temperature	Accepted dissolved oxygen
pH (off-line)	pH (on-line)
Chemical parameters	***Control action parameters***
Lactic acid concentration	Dissolved oxygen (DO) controller output
Glucose concentration	Air sparge rate
Sodium ion concentration	Air sparge set point
Ammonium ion concentration	Total air sparged
Physiological parameters	Oxygen sparge rate
Viable cell density	Total oxygen sparged
Viability	pH controller output
Packed cell volume	Total base added
Integral of packed cell volume¶	CO_2 sparge rate
	Total CO_2 sparged
	Total gas sparged
	Others
	Oxygen uptake rate
	Overlay flowrate
	Exhaust valve pressure¶
	Backpressure¶

¶ Estimated only at 12000L scale

The parameters include physical and state parameters, chemical parameters, and physiological parameters. A total of 12, 12, and 13 parameters were measured periodically at the bioreactor scales of 400L, 2000L, and 12000L, respectively. Due to the differences in sampling frequencies of the off-line parameters, all the off-line measurements were preprocessed by linear interpolation. Figure 7.2b shows the temporal profiles of glucose concentration and medium osmolality at the 12000L scale for one run. The interpolated patterns adequately represent the dynamics of the measured parameters. Each parameter was 'sampled' at a uniform interval of 10 hours from the interpolated dataset.

7.3.2 ESTIMATION OF SIMILARITY BETWEEN PARAMETER PROFILES

A crucial aspect of our approach is to compare the 'likeness' of any two runs based on the process parameter profiles. In Chapter 3, we proposed a two-step method to compute the similarity between two runs (say run 1 and run 2). In the first step, individual parameter profiles (e.g. osmolality, CO_2 sparge rate, etc) from run 1 are compared with the corresponding parameter profiles in run 2 and a similarity score is computed for each parameter. In the second step, the individual parameter-wise similarity scores are integrated to estimate the overall similarity between the two runs.

The temporal profile of a process parameter $(\mathbf{p})$ was compared between any two runs (denoted by *i* and *j*) by Euclidean distance metric $\left(d_{ij}^{p}\right)$.

$$\boldsymbol{d}_{ij}^{\boldsymbol{p}} = \left\|\mathbf{p_i} - \mathbf{p_j}\right\| = \sqrt{\sum_{k=1}^{l}\left(\boldsymbol{p}_{ik} - \boldsymbol{p}_{jk}\right)^2}$$

where, $\boldsymbol{p}_{ik}$ corresponds to the measured value of the parameter at time point *k* in run *i*.

For *n* different runs, the parameter profiles $\left(\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_n}\right)$ were compared in a pairwise manner. The resulting Euclidean distances were scaled between 0 and 5, where 0 corresponds to the highest similarity, and 5 corresponds to the lowest similarity between two profiles. Euclidean distance metric $\left(d_{ij}^{p}\right)$ was translated into a similarity metric $\left(s_{ij}^{p}\right)$ by the exponential transformation.

$$\boldsymbol{s}_{\boldsymbol{ij}}^{\boldsymbol{p}} = \exp\left(-\boldsymbol{d}_{\boldsymbol{ij}}^{\boldsymbol{p}}\right)$$

The similarity metric ranges between 0 (dissimilar profiles) and 1 (identical profiles). All the pairwise estimates of the similarity of a parameter profile across different runs comprised a similarity matrix for that parameter. The similarity matrix is symmetric, and positive semidefinite (i.e. all the eigenvalues of the matrix are non-negative), thus satisfying the Mercer's theorem. The similarity matrix is, therefore, a valid Mercer kernel.

The likeness between two runs (*i* and *j*) was computed by a weighted linear combination of the similarity between individual parameter profiles. For example, for three parameters $(\mathbf{p}, \mathbf{q}, \mathbf{r})$ measured in runs *i* and *j*, the overall similarity is estimated as:

$$s_{ij} = w_p s_{ij}^p + w_q s_{ij}^q + w_r s_{ij}^r$$

where, w_p, w_q, w_r are the weighting factors for parameters $\mathbf{p}, \mathbf{q},$ and $\mathbf{r}$, respectively.

7.3.3 ESTIMATION OF PARAMETER WEIGHT

A weight was assigned to every parameter by comparing the similarity of that parameter profile between any two runs with the difference in the outcome of the two runs. Final product titer was used as a measure of process outcome. For every parameter, all possible combinations of two runs were compared. The difference in their final titers was correlated to the similarity between the temporal profiles of the parameter using Spearman's rank correlation coefficient (*rho*). The weights for individual parameters were obtained by scaling *rho* such that the sum of all the weights is equal to one.

7.3.4 SUPERVISED MACHINE LEARNING

7.3.4.1 Support vector regression (SVR)

A set of *n* training runs can be denoted as $(\mathbf{x}_i, y_i) \ \forall \ i = 1, 2, ..., n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the space of all input parameters, and y_i is the outcome of i^{th} run. A support vector regression (SVR) models seeks to identify a regression function $f(\mathbf{x}) = \mathbf{w} \bullet \mathbf{x} + b$ that minimizes the difference (i.e. the error) between the true process outcome (y_i) and the model-predicted outcome $(f(\mathbf{x}_i))$. A *v*-SVR algorithm was employed to estimate the regression function[328]. For each run $(\mathbf{x}_i, y_i)$ an error $|y_i - f(\mathbf{x}_i)|$ of up to ε is considered acceptable. Differences

exceeding ε are penalized by a slack variable $(\xi_i$ or $\xi_i')$ and the *a priori* chosen cost function (*C*). The parameters $(\mathbf{w}, b)$ of the regression function are obtained by solving a constrained optimization problem:

$$\min_{\mathbf{w},b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n}(\xi_i + \xi_i')\right)\right\}$$

subject to the following inequality constraints $(\forall i = 1, 2, ..., n)$

$$(\mathbf{w} \bullet \mathbf{x}_i + b) - y_i \le \varepsilon + \xi_i$$

$$y_i - (\mathbf{w} \bullet \mathbf{x}_i + b) \le \varepsilon + \xi_i'$$

$$\xi_i, \xi_i' \ge 0 \; ; \; \varepsilon \ge 0$$

The *v*-SVR algorithm seeks to minimize the error ε. The parameter *v* is a non-negative constant that determines the balance between the complexity of the model and the extent of the error ε. LIBSVM[329], an implementation of *v*-SVR in C was used for training and validation of SVR models. The default value of $v = 0.5$ was used. Models were constructed for three different cost parameters, $C = 0.1$, 0.5, and 1.0.

7.3.4.2 Model training and evaluation

A ten-fold cross-validation approach was used assess the generalizability of the *v*-SVR approach (Figure 7.3). The dataset was randomly divided into ten groups; nine of these groups were used as the training set to learn the SVR model, and the tenth group was used to test the prediction accuracy of the model by using it to predict its titer. This process was repeated ten times, each time using a different group for testing. Note that for each one of the ten constructed models, only the runs making up the training set for that model were used to compute the parameter weights for the profile similarities.

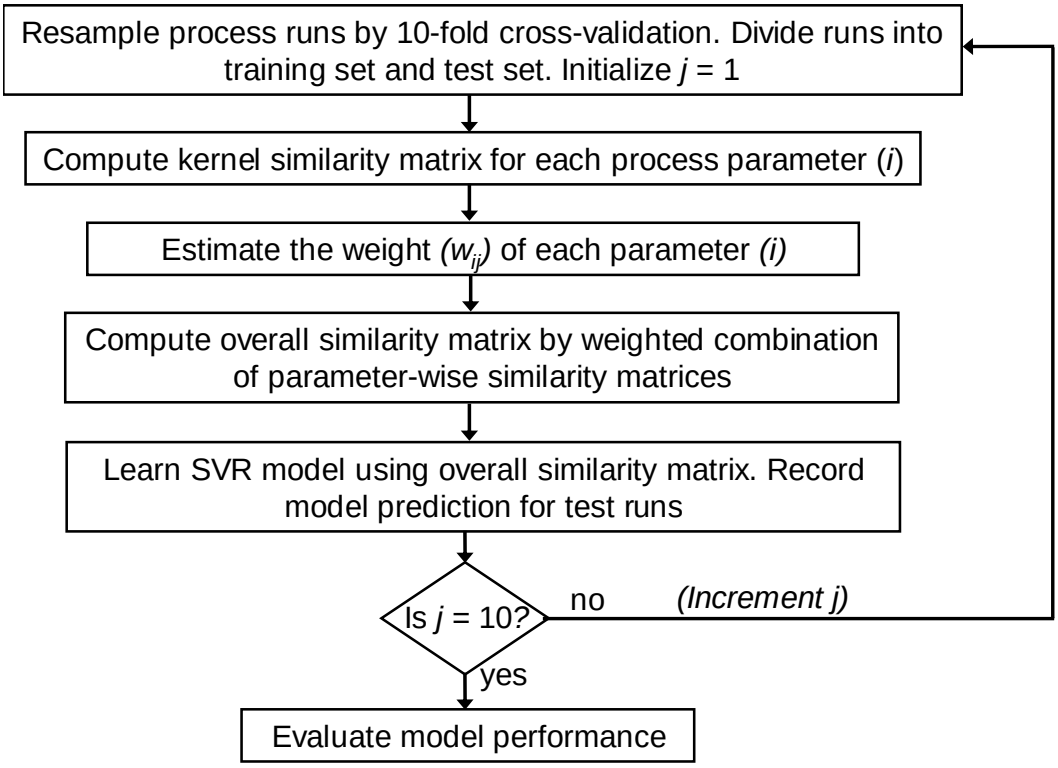

Figure 7.3. Flow diagram of the proposed methodology for knowledge discovery in manufacturing cell culture process datasets

Model performance was assessed by comparing the model-predicted titers and the actual titers of the test runs using Pearson's correlation coefficient and root mean square error (RMSE). RMSE is a measure of the average error in predicting the final titer of a run:

$$\boldsymbol{RMSE} = \sqrt{\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{f}\left(\mathbf{x}_\mathbf{i}\right)\right)^2}$$

where, $\boldsymbol{y}_i$ and $\boldsymbol{f}\left(\mathbf{x}_\mathbf{i}\right)$ are the actual and the model-predicted titers of run *i,* respectively

7.4 RESULTS

In this study, we implement a data mining strategy to preprocess, compare, and integrate process data across several runs in order to identify important process parameters at various stages of process runs and utilize that information to develop models to predict a critical process outcome – final product titer.

7.4.1 SELECTION AND PREPROCESSING OF BIOPROCESS DATA

All the parameters acquired on-line were examined across several runs to identify a subset of parameters based on their physiological importance, temporal dynamics, and also the data types. The selected parameters at different bioreactor scales are summarized in Table 7.1. A total of 18 on-line parameters were selected at 12000L scale, and 16 parameters were selected at each of the two smaller scales (400L and 2000L). In order to reduce noise at local time scales and also dampen the effect of the local discontinuities observed due to periodic interventions, all the on-line parameters were preprocessed using a moving window average method. Off-line parameters were also preprocessed by linear interpolation (see Methods).

The resulting preprocessed data comprises a total of 87 temporal parameters: 31 (18 on-line, 13 off-line) at 12000L, and 28 (16 on-line, 12 off-line) each at 2000L and 400L scales. The preprocessed parameter profiles can be compared to identify the similarities and differences in their temporal patterns across different runs.

7.4.2 KERNEL TRANSFORMATION AND COMPARISON OF PROCESS RUNS

7.4.2.1 Parameter profile-based comparison of process runs

In computing the similarity between two runs, individual parameter profiles (e.g. osmolality, CO_2 sparge rate, etc) were first compared, and a similarity score was computed for each parameter. In the second step, similarity scores of different parameters were integrated to estimate the overall similarity between the two runs.

A Euclidean distance metric was used in the first step for parameter-wise comparison. Euclidean distance was converted to a similarity score by an exponential kernel transformation (see Methods). Thus, for example, the osmolality profile from the 12000L scale of every process run was compared with the osmolality profiles at the 12000L scale of all the other 29 runs in a pair-wise manner. The results comprise a 30×30 kernel similarity matrix for osmolality. Similarly, kernel similarity matrices were generated for all the 87 temporal parameters at the three different bioreactor scales.

The second step involves aggregation of the parameter-wise similarity scores for an overall estimate of the likeness of two process runs. Here, we propose a scheme for weighted combination of the parameter-wise similarity scores such that critical process parameters have a greater contribution to the overall similarity between two runs.

7.4.3 PRODUCTIVITY-BASED APPROACH FOR PARAMETER WEIGHTING

A simple, yet proficient approach was used to weight the process parameters. The trend associated with changes in the temporal profile of a process parameter across runs with different outcomes was used to weight the parameter. This can be illustrated with an example of four process runs (runs 1, 2, 3, and 4) with normalized final product titers of 0.8, 0.9, 1.0, and 1.1, respectively. The temporal profile of a parameter in run 1 is pair-wise compared with the parameter profiles in the other three runs. Consider a case where all the three similarity scores are 0.9 (on a 0-1 scale). In this case the parameter has a comparable profile across all the four runs. This indicates that the parameter does not provide much information for deciphering the differences in the outcome of the four runs. In contrast, consider a case where the three similarity scores (between runs 1-2, 1-3, and 1-4) are 0.9, 0.6, and 0.4, respectively. Here, the *increasing* titer-difference between runs 1-2, 1-3, and 1-4 correlates with *decreasing* similarity between the parameter profiles. This opposing trend can be quantified by the Spearman's rank correlation (*rho*), a non-linear measure for assessing the correlation between two variables. Thus, for example, in the latter case, the *rho* between titer-differences (0.1, 0.2, and 0.3) and similarity scores (0.9, 0.6, and 0.3) for the parameter is -1.0 indicating that the parameter profile can discriminate between runs with different titers.

For every process parameter, the trend between its profile similarity score (between every pair of runs) and the titer-difference (between the two runs) was assessed across all the 30 runs by the *rho* metric. Figure 7.4 shows the similarity matrices for six process parameters acquired at 12000L scale. The 30 runs are arranged in an increasing order of final titer, such that run 1 has the lowest titer and run 30 has the highest titer. Every element of a parameter similarity matrix represents the similarity score of the temporal profiles of that parameter between two runs. Note that the diagonal element, where a parameter profile is compared to itself is always 1 (i.e. a profile is self-identical). As one moves away from the diagonal, the titer-difference between the runs increases. This correlates strongly with decreasing profile similarity for lactic acid, total base added, and oxygen sparge rate, but not for dissolved oxygen (Figure 7.4). The correlations (*rho*) for the parameters are shown in parenthesis. *Rho* metric can be used to assign a degree of importance to every parameter. Weighting factors are proportional to *rho* such that parameters with high negative *rho* have greater weights.

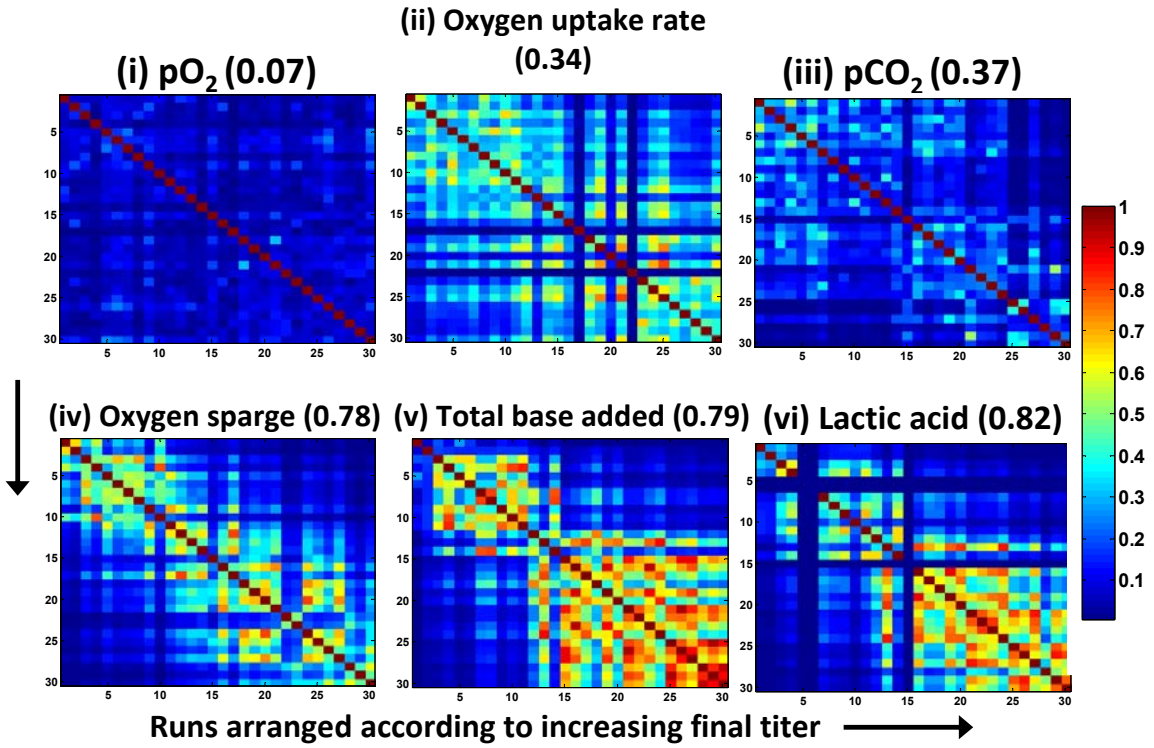

Figure 7.4. A differential weighting scheme for process parameters. The kernel similarity matrices for six process parameters acquired at 12000L scale are shown. The parameters are (i) Dissolved oxygen (pO_2), (ii) Oxygen uptake rate (OUR), (iii) Dissolved carbon dioxide (pCO_2), (iv) Oxygen sparge rate, (v) Total base added, (vi) Lactic acid. Each element (*i,j*) of a parameter kernel matrix represents the similarity (s_{ij}) between the temporal profiles of that parameter between two runs (run *i* and run *j*). The similarity score ranges from 0 (blue) for dissimilar profiles to 1 (red) for identical profiles. Note that the matrix is symmetric, i.e. $s_{ij} = s_{ji}$, and diagonal values are 1, i.e. $s_{ii} = 1$. The runs are arranged in increasing order of normalized final titer – run 1 has the minimum titer of 0.76 and the run 30 has the maximum titer of 1.17. A correlation between decreasing parameter profile similarity and increasing titer-difference between runs is observable as a red-to-blue gradient. The correlation is strong for lactic acid, total base added, and oxygen sparge rate. A weak correlation exists for pCO_2 and OUR, whereas pO_2 is uncorrelated. The number in parenthesis is the absolute value of Spearman's rank correlation coefficient for each of the six parameters

7.4.4 INTEGRATION OF ALL PROCESS PARAMETERS

The overall similarity between any two runs was estimated as a weighted combination of all parameter-wise similarity scores. The overall similarity was estimated for every combination of two process runs to obtain a 'fused' kernel matrix. Using the comparative information in this kernel matrix, we constructed models to predict the final outcome of different runs.

7.4.5 PREDICTIVE DATA MINING USING SUPPORT VECTOR REGRESSION

7.4.5.1 Process Datasets

A supervised machine learning method, support vector regression (SVR), was used to construct models for predicting process outcome, the final product titer. Several SVR models were formulated to investigate the progression of the production trains by gradually incrementing the dimensionality of the process dataset (Figure 5a, top panel). Thus, the first dataset comprises on-line and off-line parameters from 400L inoculum bioreactors only. The fused kernel matrix comprises the similarities of 16 on-line and 12 off-line parameters acquired at 400L scale. Based on this kernel matrix, SVR constructed a model to predict the final titer of the run. Note that the final titer is estimated approximately two weeks after the cells are transferred from 400L to 2000L, which in turn is used to inoculate the fed-batch cultures in 12000L bioreactors. Thus, SVR models based on 400L process data employs data very early in the cell culture process to predict the final process outcome.

The second dataset combines process data from both the inoculum bioreactors, 400L and 200L to predict the final titer using SVR. Similarly, the subsequent four datasets (dataset 3, 4, 5, and 6) increase the dimensionality of the previous dataset by incorporating process data up to day 3, day 5, day 7, and day 9 of the 12000L bioreactors. Lastly, the seventh dataset integrates all the process data from the three scales to formulate SVR models for titer prediction. This cumulative organization of process data allows a comparison of the predictability of the final titer by the SVR models at various stages of the production train. In addition, the stage-wise comparison is advantageous for identifying critical process parameters at the various stages of production.

7.4.5.2 SVR models for predicting process outcome

As described in Methods, a 10-fold cross-validation scheme was used to evaluate and compare the SVR models constructed for all the seven datasets described above. Random predictors were also used to assess the significance of the SVR models. Based on 10000 simulations of randomized titer prediction, the Pearson's correlation (between actual and predicted titer) is expectedly zero, and the root mean square error (RMSE) is 0.162. In contrast, the SVR model based solely on process data from 400L inoculum bioreactors (dataset 1) has significantly high predictability with Pearson's correlation (r) and RMSE (ε) of 0.34 and 0.099, respectively (Figure 7.5). This indicates a noticeable improvement compared to random prediction. As early as two weeks before cell harvest and downstream protein purification, the

final titer can be predicted with nearly 40% accuracy. Incorporation of process data from 2000L incolum reactors (dataset 2) results in a marginal decrease in predictability of the SVR model with the evaluation metrics, r and ε, as 0.29 and 0.100, respectively (Figure 7.5).

A marked improvement in model predictability is observed when process data from the first three days of 12000L production scale bioreactors is included (dataset 3). The evaluation metrics, r and ε increase to 0.61 and 0.088, respectively. This trend of increasing predictability is conspicuous for datasets 4-7 where process data from additional days of 12000L bioreactors is added sequentially. Thus, by the 9th day post inoculation of 12000L bioreactors (dataset 6), the final titer can be predicted with very a high accuracy (r = 0.92, ε = 0.055) (Figure 7.5).

The contribution of every process parameter in these SVR models is determined by a differential weighting scheme (described earlier). However, regardless of the degree of significance, all the process parameters contribute to model formulation, resulting in a high data dimensionality. The detrimental effects of this 'curse of high dimensionality' on data mining methods are well-known[330]. To alleviate this effect, the differential weighting scheme was also used to reduce data dimensionality be pre-selecting a subset of top-weighting process parameters. SVR models were thereafter formulated using the parameter subset only. For each of the seven process datasets, the top 2, 5, and 10 process parameters were selected by the weighting scheme and SVR models were constructed by combining the kernel similarity matrices of the selected parameter only. Caution was exercised to avoid 'selection bias' by performing parameter selection only on the training runs (without the inclusion of test runs)[183]. For six of the seven datasets, SVR models based on parameter subsets result in a reduction in RMSE indicating that outcome predictability is enhanced (Figure 7.5b). For example, in dataset 4, which includes process data from inoculum bioreactors and up to the first 5 days of 12000L bioreactors, the RMSE reduces by 17% from 0.082 (SVR model with all parameters) to 0.068 for the SVR model with only the top 2 parameters. Also, the Pearson's correlation between the actual titer and the model-predicted titer increases from 0.70 to 0.76. This implies that by day 5 in the 12000L bioreactor, the final titer can be predicted with good accuracy by the SVR model. Lastly, the top 10-parameter SVR model for dataset 6 has a very good performance (r = 0.94, ε = 0.043) evincing that the final titer can be predicted very accurately by day 9 in the 12000L bioreactors.

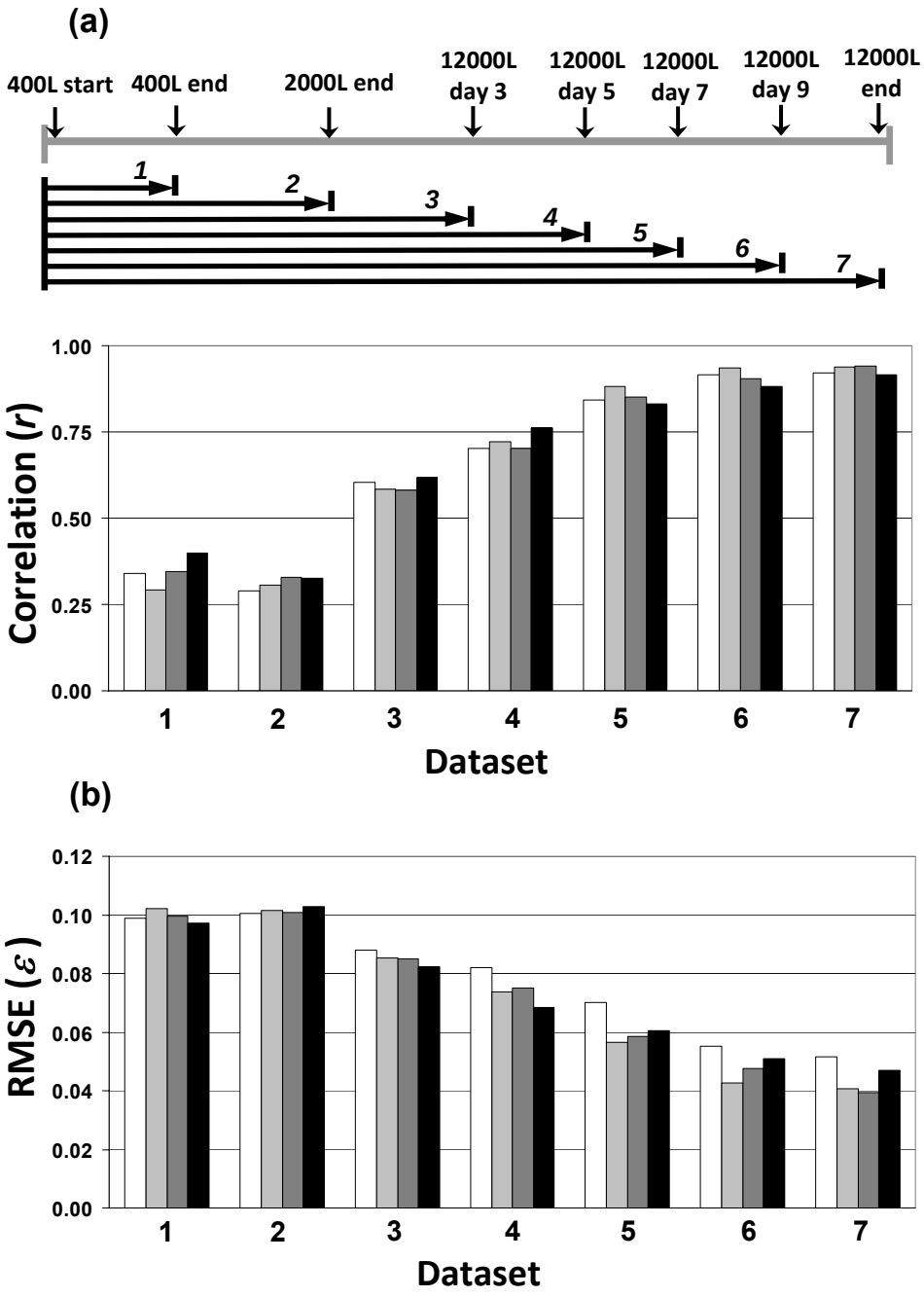

Figure 7.5. Evaluation of SVR model performance. (a) Top panel: The process dataset from the inoculum bioreactors (400L and 2000L) and the production bioreactor (12000L) was divided into seven datasets in order to analyze process data in a stage-wise cumulative manner. The time scale of each dataset is shown in the process timeline. For each dataset, a 10-fold cross-validation procedure was used to assess SVR model performance. Bottom panel: Pearson's correlation (*r*) between SVR model-predicted titer and actual titer of test runs, (b) Root mean square error (*ε*) for SVR models. SVR model based on: (□) all parameters, (▫) top 10 parameters, (▪) top 5 parameters, (■) top 2 parameters.

7.4.6 STAGE-SPECIFIC IDENTIFICATION OF CRITICAL PROCESS PARAMETERS

7.4.6.1 Weight-based assessment of process parameters

The Spearman's rank correlation (*rho*) is used to assess the weight of every parameter in distinguishing between high and low titer runs. Figure 7.6 shows the *rho* metric for every process parameter in 12000L-scale. The *rho* metric was estimated for every parameter at each of the five stages (day 3, 5, 7, 9, and all days). Recall that a parameter with negative *rho* (and therefore a higher weight) correlates with deviations in process outcome. To facilitate comparison, the negative of *rho* (e.g., +0.6 instead of -0.6) is shown in Figure 7.6. A subset of parameters comprising DO controller output, oxygen sparge rate, osmolality, and glucose concentration exhibit a strong increase in *rho* at day 9 (dataset 6) compared to previous days (day 3, 5, 7). DO controller output and osmolality profiles for the top 5 runs with highest titer and the top 5 runs with lowest titers are shown in Figure 7.7a. It is evident that during the first 160 hours, the profiles for these two parameters are similar between the high and low titer runs. A conspicuous difference between the two classes (high and low runs) emerges after 160 hours. For the runs with low-titer, there is a discernible increase in medium osmolality and a decrease in DO controller output revealing a decrease in viability of the runs with low-titer. This is also evident by the high weight for viability parameter at day 9 (Figure 7.6). Thus, the model is successful in discerning the temporal effects of the process parameters.

7.4.6.2 Association between early stage process parameters and process outcome

Particularly informative are the parameters that are indicative of process class in the early stages of a process run as they can potentially be used as an early warning system to detect atypical batches thereby providing the process specialist an opportunity to pursue remedial action. At the early stages, (day 3 and day 5), the parameters including concentrations of lactic acid and sodium ion, total base added, and pH controller output have high weighting factors (Figure 7.6) indicating that the profiles of these parameters differ between high-titer and low-titer runs as early as day 3. The profiles for lactic acid and total base added during the first three days are shown in Figure 7.7b. A greater accumulation of lactic acid is observable in runs with low-titer, which correlates with the greater amount of base added as a control response to maintain a constant pH. The early differences at 12000L scale between high and low-titer runs prompted us to investigate

the process data from inoculum train to identify cues for the departures in process outcome. The average Spearman's correlation (*rho*) for process parameters acquired at 400L and 200L scale are shown in Figure 7.8. Unlike the observations at 12000L scale, 22 out of 28 parameters at each scale have *rho* less than 0.05 indicating that these parameters are not correlated to deviations in final titer. A noticeable exception is the lactic acid profile at 400L scale (Figure 7.9). The concentration of lactic acid is appreciably higher in the runs with low final titer. This is also evident from the greater concentration of sodium in the low-titer runs after 50 hours due to base addition to neutralize the decrease in pH caused by lactic acid accumulation (Figure 7.9). These observations are striking in that they suggest that the history of the inoculum plays an important role in determining the final process outcome.

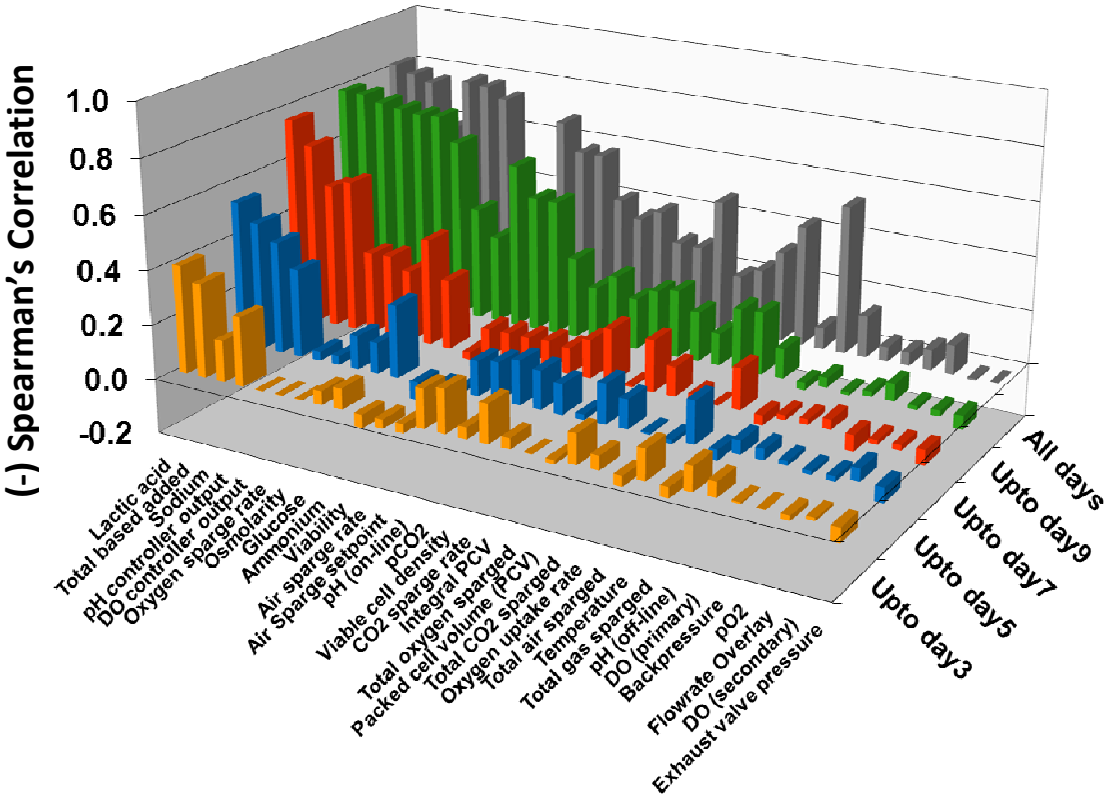

Figure 7.6. Relative importance of process parameters at different stages of the production phase (12000L scale). The Spearman's correlation coefficient (*rho*) for all the 12000L process parameters for data up to: day 3, day 5, day 7, day 9, and all days are shown. Due to the opposite nature of the trend, the negative of *rho* is shown (i.e., *rho* of -0.6 is displayed as +0.6)

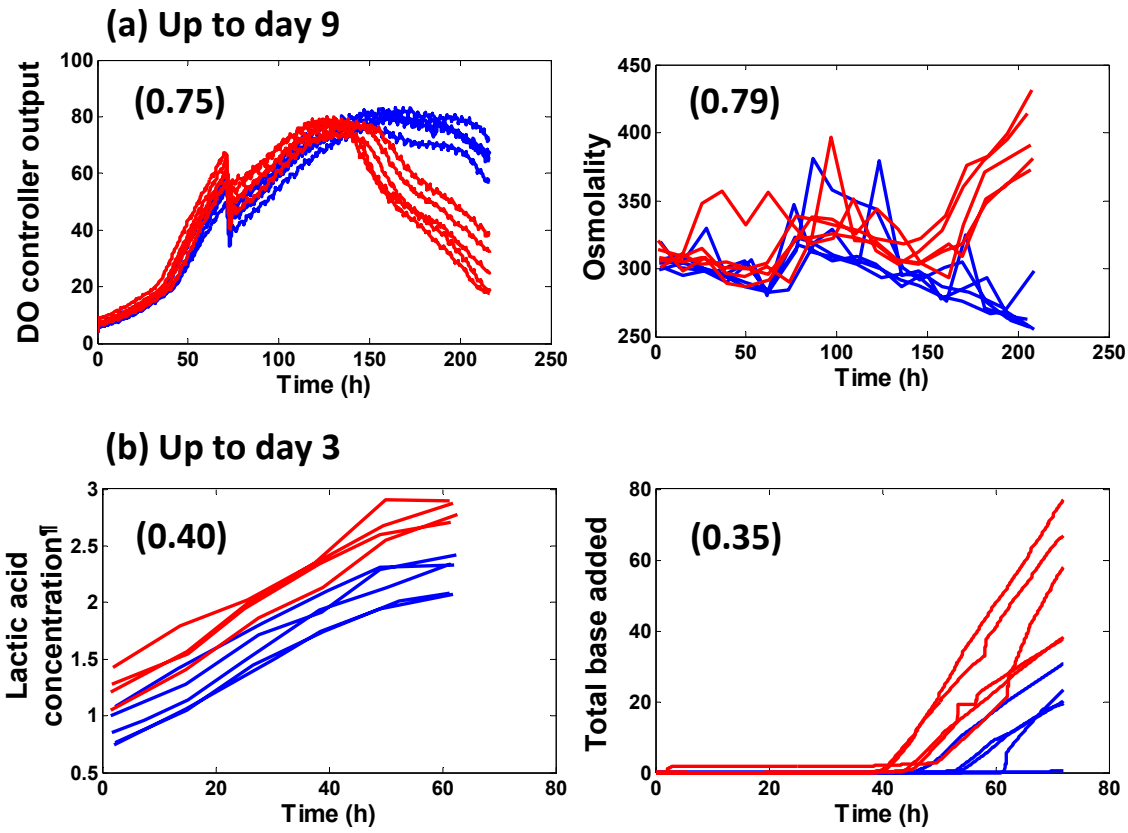

Figure 7.7. Selected critical process parameters at different stages of the production phase. (a) Process data up to day 9. The profiles for DO controller output (left panel) and medium osmolality (right panel) are shown. (b) Process data up to day 3. The profiles of lactic acid concentration (left panel) and total base added (right panel) are shown. The *rho* for each parameter is shown in parenthesis. Top 5 runs with low final titer (in red), top 5 runs with high final titer (blue). ¶Lactic acid concentration was measured in four of the five low titer runs.

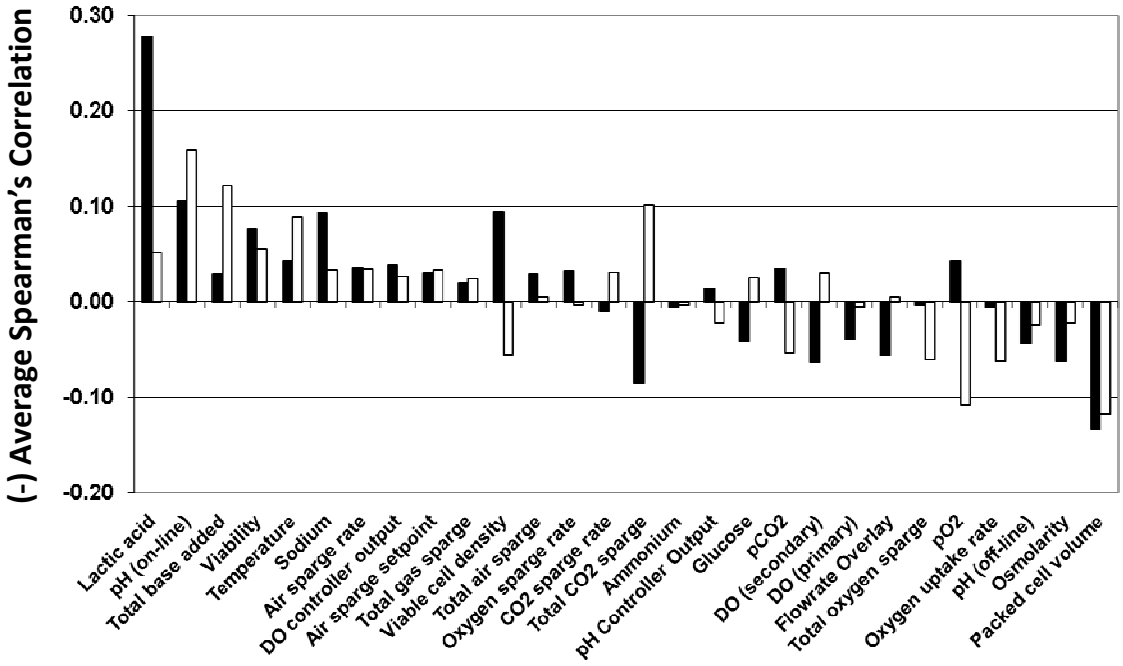

Figure 7.8. Relative importance of process parameters acquired at 400L and 2000L scale inoculum bioreactors. The Spearman's correlation coefficients (*rho*) for all the process parameters are shown. Due to the opposite nature of the trend, the negative of *rho* is shown. (■) 400L, (□) 2000L.

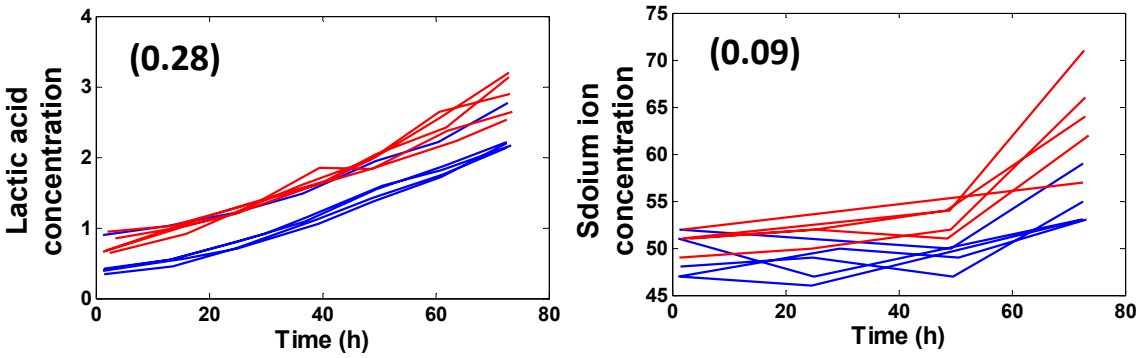

Figure 7.9. Critical process parameters measured at 400L scale. The profiles of lactic acid concentration (left panel) and sodium ion concentration (right panel) are shown. The *rho* for each parameter is shown in parenthesis. Top 5 runs with low final titer (in red), top 5 runs with high final titer (blue).

7.5 DISCUSSION

7.5.1 AN ADAPTABLE FRAMEWORK FOR MINING PROCESS CELL CULTURE DATA

In this study, a framework was described to systematically interrogate large volumes of manufacturing-scale process data to identify the distinguishing characteristics of a high productivity process. The challenges associated with analyzing complex bioprocess data archives have been highlighted before[150, 331]. Each process run is represented by a large number of parameters measured over long time periods. Further, the task entails comparing process data from a number of different runs. The difficulties associated with comparing multiple temporal parameter between different runs can be alleviated by concatenating all the process parameters from one run as a vector, as described in a previous report[332]. Another approach, described by Kamimura *et al.*[150] employed principal component analysis (PCA) to identify the linear correlations between temporal process parameters as a few principal components (called eigenvectors). The dimensionality of the parameter matrix was then reduced to a vector by considering only the first PC. While this is a satisfactory alternative, the first PC typically fails to capture more than 50% of the dynamics of the data leading to loss of information.

The approach illustrated here does not obligate such measures for dimensionality reduction. The temporal profile of each parameter is compared across all runs to generate a similarity matrix or a kernel matrix for every parameter. Also, the approach does not place any restrictions on the number or the type of parameters. Any number of parameters can be combined without compromising their integrity or their temporal dynamics. Secondly, the approach can aptly address the heterogeneity of different parameters. Off-line and on-line parameters with vastly different sampling frequencies and scales can be combined conveniently. Furthermore, different similarity metrics and kernel transformations can be used for comparing different parameters. The notion of constructing models by integrating diverse data types has been used in several bioinformatics applications, such as predicting protein-protein interactions[333] and predicting gene functions[15].

7.5.2 AN EFFICIENT WEIGHTING STRATEGY FOR INTEGRATING HETEROGENEOUS PROCESS PARAMETERS

One strategy to integrate heterogeneous datasets for comparing different runs is to simply add all the parameter-wise similarity scores between any two runs. For example, if two runs are compared based on similarity scores of the profiles of glucose concentration, gas overlay flow-

rate, and vessel temperature, then the overall similarity between the two runs is the summation of the similarity scores of the above-mentioned three parameters. A drawback of this approach is that the contributions all the parameters are treated as being equal when in reality that is not the case. For example, glucose concentration, which is indicative of glucose consumption rate, is likely to have a greater impact on cellular metabolism than the other two parameters. A simple addition scheme can weaken or completely mask the effect of a few important parameters. A differential parameter weighting scheme was therefore implemented. All the process parameters were weighted according to their relative predictability with respect to process outcome using a non-linear metric, Spearman's correlation coefficient. The weighting scheme was used to sort as well as select critical process parameters at different stages of the process runs. Our model cross-validation results suggest that this approach of selecting a subset of relevant parameters results in models with lower prediction errors (Figure 7.5b).

A critical aspect of knowledge discovery for bioprocesses involves identification of process parameters that are important indicators of process outcome. These indicators provide a set of norms for evaluating and comparing the overall performance of different runs in addition to a straightforward examination of the end-point titer. The data mining strategy proposed here is well-equipped for determination of critical process parameters. The Spearman's rank correlation (*rho*) of a parameter, which is the basis for estimating parameter weight, can serve as a guide for assessing the ability of a parameter to distinguish between high and low-titer runs. Further, the sequential organization of process data allows us to determine predictive parameters at different stages of a run.

The concept of weighting different parameters can be further extended to identify the optimal parameter integration scheme that maximizes model predictability. In a bioinformatics study, Lanckriet *et al.*[170] proposed models for predicting the functional class of a protein based on diverse datasets comprising protein sequence information, transcript expression profile, and known protein-protein interactions. A convex optimization framework was proposed to determine the weighted linear combination of parameters that results in the best predictor of protein functional class.

7.5.3 INFLUENCE OF PROCESS PARAMETERS ON OUTCOME

The adverse effects of lactic acid accumulation on viability and recombinant protein productivity of mammalian cells are well-known. The results here show that differences between the lactic acid profiles of high and low-titer runs emerge in very early stages of culture.

Accumulation of higher levels of lactic acid is an impediment to achieving high cell densities, and therefore, high product titers in mammalian cell culture processes. Several studies have attempted to engineer mammalian cells with reduced lactic acid production by employing strategies to decrease the level of lactate dehydrogenase enzyme[100, 334] or decrease glucose uptake rate by metabolic shift (reduced glucose supply)[335] or glucose transporter engineering[336]. Understanding the cause of increased lactate formation in mammalian cells will certainly aid efforts to increase recombinant protein productivity. It is therefore interesting to note that increased lactate formation in the inoculum train, as early as two weeks before estimation of final titer, can provide an early quantitative indication of process outcome.

SVR models employed in this analysis highlight that process data acquired during cell expansion in the inoculum train and the first three days of the final production phase can predict process outcome with good accuracy (Figure 7.5). This suggests that key events that affect process outcome occur before day 3 in the 12000L bioreactor. Careful examination of the process data, especially the parameters that regulate control actions in these early stages, is necessary to investigate the causes of fluctuations in process outcome. Further, scrutinizing the history of the cells during expansion in the inoculum train can provide insights about how the observed correlations between parameter profiles and process outcome affects cellular physiology. The quantitative effects of the 'quality' of seed on process outcome have been previously highlighted by an unsupervised classification method for an industrial-scale Penicillin production process[332].

The effect of raw materials should also be critically investigated. Particularly interesting is an inspection of how process outcome is affected by variations in the batch (different lots) or the source (different companies) of the raw material. For example, a decision tree-based method identified that the source of a media component, methionine, has a strong correlation with the outcome of the fermentation process[337]. The study also identified that the timing of addition of an unspecified carbohydrate component had an impact on product yield. Due to the complexity of the raw materials and the large number of components involved in cell culture processes, a detailed examination of the effect of raw materials would necessitate further measurements to determine the consumption and production rates of various nutrients and metabolites including amino acids. However, such investigative efforts may be restricted to smaller bioreactor scales. Nonetheless, inspection of the variations in the raw material, particularly complex components such as protein hydrolysates, is warranted.

7.6 CONCLUDING REMARKS

In this study, we develop a predictive data mining tool to analyze historical, production-scale, cell culture process data in an effort to uncover the distinguishing characteristics of high productivity processes. The proposed framework integrates heterogeneous process data comprising hundreds of temporal parameters, measured and archived during production processes, to construct machine learning models for predicting the final titer of the recombinant protein. Based on the models, the final titers can be predicted with reliable accuracies several days prior to process completion. Model results also suggest that the history of the inoculum also plays a role in determining process outcome. The methodology can also be extended to investigate other critical process outputs, such as product quality. With increasing Process Analytical Technology (PAT) initiatives by the U.S. Food and Drug Administration to "design, analyze, and control" manufacturing processes to ensure product quality (http://www.fda.gov/Cder/OPS/PAT.htm), attempts to mine multivariate process data can provide strategies for bioprocess optimization.

CHAPTER 8 SUMMARY AND CONCLUDING REMARKS

The tremendous successes of biotechnologists to deliver ever-increasing quantities of life-saving drugs to patients worldwide are awe-inspiring. The future presents many further opportunities for cell engineering and process intensification to harness the pharmaceutical potential of biological systems.

Today, the fermentation technology for antibiotic production is mature with nearly 250 drugs most of which were approved 20-30 years ago. However, there are renewed concerns about pathogenic microorganisms with multidrug resistance. In an alarming report, Dantas *et al.*[338] showed that hundreds of phylogenetically diverse, soil-dwelling bacteria could not only survive but also grow on antibiotics at clinically relevant dosage. The close phylogenetic relationships between many of these multidrug resistant bacteria and human pathogens bespeaks of the need to understand the physiology of antibiotic-producing Streptomycetes in an effort to discover and engineer novel antibiotics made by these 'soil-warriors'.

This study focused on *Streptomyces coelicolor*, the most genetically characterized species of *Streptomyces* genus. The availability of the whole-genome sequence for *S. coelicolor* has equipped us with functional genomic tools to deconvolute the regulatory aspects of antibiotic production. Here, the temporal transcriptome data from wild-types and several strains constituting a wide gamut of genetic and environmental perturbations was compiled. Since an operon represents the elemental unit of transcriptional regulation in bacteria, the transcriptome data was combined with other relevant attributes to construct a whole-genome operon map for *S. coelicolor* (Chapter 4 and 5). The experimental verification of a significant fraction of the predicted operons further demonstrates the utility of the operon model.

In a subsequent step, the genome-wide transcriptome profiles were analyzed to predict the transcriptional regulation network of *S. coelicolor* (Chapter 5). The network exhibits a scale-free topology previously observed in other living organisms. The network constitutes more than 7,000 direct regulatory edges between genes encoding putative regulators and other cistrons in the genome. Several subgraphs of the network, called modules, are enriched for functions related to primary and secondary metabolism. The network predictions can be further validated by integrating additional data sources such as the presence of a common *cis*-regulatory element in

the promoter segments of several cistrons targeted by one regulator[136]. The enormous growth of omics technologies and the barrage of high-throughput information about the physical as well as genetic associations between genes and proteins have incited much interest in development of tools to mine the exhaustive biomedical literature[339]. Application of such tools to identify literature-reported associations between different genes in *S. coelicolor* and its close relatives also appears promising. Furthermore, the conservation of a regulatory edge in multiple species provides yet another source to validate and identify the conserved regions of an interaction map[340]. Integration of these multiple data sources will allow us to sort and prioritize regulatory interactions, many of which can be experimentally assessed by ChIP or ChIP-on-Chip assays. Lastly, the dynamic activity of a transcriptional network is fine-tuned by cells to adapt to the changes in its environment. Temporal transcriptome data is a valuable resource for discerning the dynamical behavior of network modules in response to various perturbations[341].

The second part of this study explored the functional determinants of high recombinant protein productivity trait in mammalian cells. These protein biologics have provided life-saving therapies to thousands of ailing patients in the past twenty years. Through considerable efforts in cell engineering and process optimization, the volumetric productivities of cell culture processes have risen from few tens of milligrams per liter to nearly five grams per liter in two decades. Further improvements will likely require a better understanding of the physiological attributes that confer growth and productivity characteristics to mammalian cells. Identification of the genetic markers that dictate this complex trait will facilitate rational cell engineering strategies to further improve recombinant protein productivity. Genome-scale technologies provide valuable investigative tools for these discovery efforts.

In this study, high and low recombinant antibody-producing NS0 mouse myeloma cells were compared using transcriptome data (Chapter 6). Instead of identifying individual differentially expressed genes, functional investigation tools were employed to detect sets of functionally related genes that are coordinately associated with the high productivity phenotype. We used a gene set analysis approach to examine molecular phenotypes at functional level to deduce subtle changes in expression patterns that may not be observable at an individual gene level. This is particularly relevant for transcriptome and proteome profiling of high-producing mammalian cells where majority of individual gene expression changes are modest (less than two-fold). Multiple gene set analysis methods were used in this study to identify such subtle but

significant alterations. Biological processes including protein processing and transport, cell cycle regulation, and cytoskeletal functions were correlated with productivity. These findings suggest that high productivity can be achieved by simultaneous modulation of several physiological attributes. A comparison of high-producing cells and productivity enhancing culture conditions will further facilitate generation of biologically relevant hypotheses. Tools that assess the significance of biological pathways by incorporating prior knowledge about the position and importance of various pathway genes can also provide fruitful insights[342]. Further, similar to their applications in studying disease phenotypes such as cancer[122], pattern recognition techniques can be used to identify subpopulations of high-producing cells with coordinated patterns of alteration of multiple pathways.

This study also examined high recombinant protein productivity trait by investigating large amounts of bioprocess data that is customarily archived in modern production plants (Chapter 7). Mining these large databases can unearth hidden patterns that associate cellular physiology with protein productivity. In this study, a data mining framework was introduced to identify predictive correlations between on-line and off-line process parameters and process outcome for several production runs. The analysis identified parameters in inoculum bioreactors and initial stages of production-scale bioreactors as early indicators of outcome. Such data mining models can also provide useful insight about other critical measures of process outcome such as product quality.

In this study, considerable emphasis was placed on developing data mining approaches to analyze and integrate large-scale biological datasets facilitate the interpretation of cellular physiology. Supervised machine learning tools such as support vector machines were used to construct models to predict and verify the operon map of *S. coelicolor* as well as to predict recombinant protein productivity of mammalian cell culture processes. The complexities of biological systems inspire an unabridged, multidimensional approach combining omics datasets and large-scale bioprocess datasets to generate and verify physiologically relevant hypotheses for rational cell engineering and process optimization. As we march forward in this quest, the integration of omics technologies with data mining-based knowledge discovery approaches will play an increasingly important role in our pursuit of a systems level understanding of the physiology of the much desired biological traits for bioprocess enhancement.

REFERENCES

1. Rokem, J.S., Lantz, A.E. & Nielsen, J. Systems biology of antibiotic production by microorganisms. *Nat Prod Rep* **24**, 1262-1287 (2007).
2. Seth, G., Hossler, P., Yee, J.C. & Hu, W.S. Engineering cells for cell culture bioprocessing--physiological fundamentals. *Adv Biochem Eng Biotechnol* **101**, 119-164 (2006).
3. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463-5467 (1977).
4. Mullis, K. et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* **51 Pt 1**, 263-273 (1986).
5. Fleischmann, R.D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).
6. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray [see comments]. *Science* **270**, 467-470 (1995).
7. Lamb, J. The Connectivity Map: a new tool for biomedical research. *Nat Rev Cancer* **7**, 54-60 (2007).
8. Nevins, J.R. & Potti, A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet* **8**, 601-609 (2007).
9. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630-631 (2004).
10. Tan, P.K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684 (2003).
11. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151-1161 (2006).
12. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55-65 (2006).
13. Dupuy, A. & Simon, R.M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* **99**, 147-157 (2007).
14. Joyce, A.R. & Palsson, B.O. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* **7**, 198-210 (2006).
15. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc Natl Acad Sci U S A* **100**, 8348-8353 (2003).
16. Zhu, J. et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet* **40**, 854-861 (2008).
17. Lian, W. in Chemical Engineering and Materials Science, Vol. PhD 192 (University of Minnesota, Minneapolis; 2005).
18. Jayapal, K.P. in Chemical Engineering and Materials Science, Vol. PhD 211 (University of Minnesota, Minneapolis; 2008).
19. Seth, G. in Chemical Engineering and Materials Science, Vol. PhD 147 (University of Minnesota, Minneapolis; 2006).

20. Challis, G.L. & Hopwood, D.A. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. *Proc Natl Acad Sci U S A* **100 Suppl 2**, 14555-14561 (2003).
21. Chater, K.F. & Bibb, M.J. in Products of secondary metabolism, Vol. 7, Edn. second. (eds. H. Kleinkauf & H. von Dohren) 57-105 (VCH, Weinheim, Germany; 1997).
22. Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F. & Hopwood, D.A. Practical Streptomyces Genetics. (John Innes Centre, Norwich; 2000).
23. Hopwood, D.A. Streptomyces genes: from Waksman to Sanger. *J Ind Microbiol Biotechnol* **30**, 468-471 (2003).
24. Bentley, S.D. et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147 (2002).
25. Ikeda, H. et al. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**, 526-531 (2003).
26. Ohnishi, Y. et al. Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350. *J Bacteriol* **190**, 4050-4060 (2008).
27. Paget, M.S., Leibovitz, E. & Buttner, M.J. A putative two-component signal transduction system regulates sigmaE, a sigma factor required for normal cell wall integrity in Streptomyces coelicolor A3(2). *Molecular Microbiology* **33**, 97-107 (1999).
28. Bibb, M.J., Molle, V. & Buttner, M.J. sigma(BldN), an extracytoplasmic function RNA polymerase sigma factor required for aerial mycelium formation in Streptomyces coelicolor A3(2). *J Bacteriol* **182**, 4606-4616 (2000).
29. Viollier, P.H., Weihofen, A., Folcher, M. & Thompson, C.J. Post-transcriptional regulation of the Streptomyces coelicolor stress responsive sigma factor, SigH, involves translational control, proteolytic processing, and an anti-sigma factor homolog. *J Mol Biol* **325**, 637-649 (2003).
30. Hutchings, M.I., Hoskisson, P.A., Chandra, G. & Buttner, M.J. Sensing and responding to diverse extracellular signals? Analysis of the sensor kinases and response regulators of Streptomyces coelicolor A3(2). *Microbiology* **150**, 2795-2806 (2004).
31. Tsao, S., Rudd, B., He, X., Chang, C. & Floss, H. Identification of a red pigment from Streptomyces coelicolor A3(2) as a mixture of prodigiosin derivatives. *J Antibiot (Tokyo)* **38**, 128-131 (1985).
32. Wasserman, H. et al. Biosynthesis of prodigiosin. Incorporation patterns of C-labeled alanine, proline, glycine, and serine elucidated by fourier transform nuclear magnetic resonance. *J Am Chem Soc* **95**, 6874-6875 (1973).
33. Coco, E., Narva, K. & Feitelson, J. New classes of Streptomyces coelicolor A3(2) mutants blocked in undecylprodigiosin (Red) biosynthesis. *Mol Gen Genet* **227**, 28-32 (1991).
34. O'Hagan, D. The polyketide metabolites. (E. Horwood, New York; 1991).
35. Brockmann, H., Pini, H. & Plotho, O. Actinorhodin. *Chem Ber* **83**, 161 (1950).
36. Challis, G. & Chater, K. Incorporation of [U-13C]glycerol defines plausible early steps for the biosynthesis of methylenomycin A in *Streptomyces coelicolor* A3(2). *Chem Commun*, 935-936 (2001).
37. Yamasaki, M., Ikuto, Y., Ohira, A., Chater, K. & Kinashi, H. Limited regions of homology between linear and circular plasmids encoding methylenomycin biosynthesis in two independently isolated streptomycetes. *Microbiology* **149**, 1351-1356 (2003).
38. Chater, K.F. & Bruton, C.J. Resistance, regulatory and production genes for the antibiotic methylenomycin are clustered. *Embo Journal* **4**, 1893-1897 (1985).

39. Huang, J., Lih, C.J., Pan, K.H. & Cohen, S.N. Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in Streptomyces coelicolor using DNA microarrays. *Genes Dev* **15**, 3183-3192 (2001).
40. Rudd, B. & Hopwood, D. Genetics of actinorhodin biosynthesis by *Streptomyces coelicolor* A3(2). *J Gen Microbiol* **114**, 35-43 (1979).
41. Malpartida, F. & Hopwood, D.A. Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature* **309**, 462-464 (1984).
42. Hallam, S., Malpartida, F. & Hopwood, D. Nucleotide sequence, transcription and deduced function of a gene involved in polyketide antibiotic synthesis in *Streptomyces coelicolor. Gene* **74**, 305-320 (1988).
43. Fernandez-Moreno, M.A., Caballero, J.L., Hopwood, D.A. & Malpartida, F. The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the bldA tRNA gene of Streptomyces. *Cell* **66**, 769-780 (1991).
44. Rudd, B. & Hopwood, D. A pigmented mycelial antibiotic in *Streptomyces coelicolor*: control by a chromosomal gene cluster. *J Gen Microbiol* **119**, 333-340 (1980).
45. Malpartida, F., Niemi, J., Navarrete, R. & Hopwood, D.A. Cloning and expression in a heterologous host of the complete set of genes for biosynthesis of the Streptomyces coelicolor antibiotic undecylprodigiosin. *Gene* **93**, 91-99 (1990).
46. Narva, K.E. & Feitelson, J.S. Nucleotide sequence and transcriptional analysis of the redD locus of Streptomyces coelicolor A3(2). *Journal of Bacteriology* **172**, 326-333 (1990).
47. Chong, P.P. et al. Physical identification of a chromosomal locus encoding biosynthetic genes for the lipopeptide calcium-dependent antibiotic (CDA) of Streptomyces coelicolor A3(2). *Microbiology* **144 (Pt 1)**, 193-199 (1998).
48. Takano, E. et al. Transcriptional regulation of the redD transcriptional activator gene accounts for growth-phase-dependent production of the antibiotic undecylprodigiosin in Streptomyces coelicolor A3(2). *Mol Microbiol* **6**, 2797-2804 (1992).
49. Gramajo, H.C., Takano, E. & Bibb, M.J. Stationary-phase production of the antibiotic actinorhodin in Streptomyces coelicolor A3(2) is transcriptionally regulated. *Mol Microbiol* **7**, 837-845 (1993).
50. White, J. & Bibb, M. bldA dependence of undecylprodigiosin production in Streptomyces coelicolor A3(2) involves a pathway-specific regulatory cascade. *Journal of Bacteriology* **179**, 627-633 (1997).
51. Wietzorrek, A. & Bibb, M. A novel family of proteins that regulates antibiotic production in streptomycetes appears to contain an OmpR-like DNA-binding fold [letter]. *Mol Microbiol* **25**, 1181-1184 (1997).
52. Arias, P., Fernandez-Moreno, M.A. & Malpartida, F. Characterization of the pathway-specific positive transcriptional regulator for actinorhodin biosynthesis in Streptomyces coelicolor A3(2) as a DNA-binding protein. *J Bacteriol* **181**, 6958-6968 (1999).
53. Hakenbeck, R. & Stock, J.B. Analysis of two-component signal transduction systems involved in transcriptional regulation. *Methods Enzymol* **273**, 281-300 (1996).
54. Ishizuka, H., Horinouchi, S., Kieser, H.M., Hopwood, D.A. & Beppu, T. A putative two-component regulatory system involved in secondary metabolism in Streptomyces spp. *Journal of Bacteriology* **174**, 7585-7594 (1992).
55. Chang, H.M., Chen, M.Y., Shieh, Y.T., Bibb, M.J. & Chen, C.W. The cutRS signal transduction system of Streptomyces lividans represses the biosynthesis of the polyketide antibiotic actinorhodin. *Molecular Microbiology* **21**, 1075-1085 (1996).

56. Adamidis, T., Riggle, P. & Champness, W. Mutations in a new *Streptomyces coelicolor* locus which globally block antibiotic biosynthesis but not sporulation. *J Bacteriol* **172**, 2962-2969 (1990).
57. Adamidis, T. & Champness, W. Genetic analysis of absB, a Streptomyces coelicolor locus involved in global antibiotic regulation. *J Bacteriol* **174**, 4622-4628 (1992).
58. Brian, P., Riggle, P.J., Santos, R.A. & Champness, W.C. Global negative regulation of Streptomyces coelicolor antibiotic synthesis mediated by an absA-encoded putative signal transduction system. *Journal of Bacteriology* **178**, 3221-3231 (1996).
59. Anderson, T.B., Brian, P. & Champness, W.C. Genetic and transcriptional analysis of absA, an antibiotic gene cluster-linked two-component system that regulates multiple antibiotics in Streptomyces coelicolor. *Mol Microbiol* **39**, 553-566. (2001).
60. McKenzie, N.L. & Nodwell, J.R. Phosphorylated AbsA2 negatively regulates antibiotic production in Streptomyces coelicolor through interactions with pathway-specific regulatory gene promoters. *J Bacteriol* **189**, 5284-5292 (2007).
61. Hong, S.K., Kito, M., Beppu, T. & Horinouchi, S. Phosphorylation of the AfsR product, a global regulatory protein for secondary-metabolite formation in Streptomyces coelicolor A3(2). *Journal of Bacteriology* **173**, 2311-2318 (1991).
62. Matsumoto, A., Hong, S.K., Ishizuka, H., Horinouchi, S. & Beppu, T. Phosphorylation of the AfsR protein involved in secondary metabolism in Streptomyces species by a eukaryotic-type protein kinase. *Gene* **146**, 47-56 (1994).
63. Umeyama, T., Lee, P.C. & Horinouchi, S. Protein serine/threonine kinases in signal transduction for secondary metabolism and morphogenesis in Streptomyces. *Appl Microbiol Biotechnol* **59**, 419-425 (2002).
64. Horinouchi, S. et al. Primary structure of AfsR, a global regulatory protein for secondary metabolite formation in Streptomyces coelicolor A3(2). *Gene* **95**, 49-56 (1990).
65. Vogtli, M., Chang, P.C. & Cohen, S.N. afsR2: a previously undetected gene encoding a 63-amino-acid protein that stimulates antibiotic production in Streptomyces lividans. *Molecular Microbiology* **14**, 643-653 (1994).
66. Lian, W. et al. Genome-wide transcriptome analysis reveals that a pleiotropic antibiotic regulator, AfsS, modulates nutritional stress response in Streptomyces coelicolor A3(2). *BMC Genomics* **9**, 56 (2008).
67. Takano, E. Gamma-butyrolactones: Streptomyces signalling molecules regulating antibiotic production and differentiation. *Curr Opin Microbiol* **9**, 287-294 (2006).
68. Takano, E. et al. Purification and structural determination of SCB1, a gamma-butyrolactone that elicits antibiotic production in Streptomyces coelicolor A3(2). *J Biol Chem* **275**, 11010-11016 (2000).
69. Hsiao, N.H. et al. ScbA from Streptomyces coelicolor A3(2) has homology to fatty acid synthases and is able to synthesize gamma-butyrolactones. *Microbiology* **153**, 1394-1404 (2007).
70. Takano, E. et al. A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of Streptomyces coelicolor. *Mol Microbiol* **56**, 465-479 (2005).
71. Horinouchi, S. A microbial hormone, A-factor, as a master switch for morphological differentiation and secondary metabolism in Streptomyces griseus. *Front Biosci* **7**, d2045-2057 (2002).
72. Onaka, H. et al. Cloning and characterization of the A-factor receptor gene from Streptomyces griseus. *Journal of Bacteriology* **177**, 6083-6092 (1995).

73. Yamada, Y., Sugamura, K., Kondo, K., Yanagimoto, M. & Okada, H. The structure of inducing factors for virginiamycin production in Streptomyces virginiae. *J Antibiot (Tokyo)* **40**, 496-504 (1987).
74. Sato, K., Nihira, T., Sakuda, S., Yanagimoto, M. & Yamada, Y. Isolation and structure of a new butyrolactone from *Streptomyces* sp. FRI-5. *Journal of Fermentation and Bioengineering* **68**, 170-173 (1989).
75. Kinoshita, H. et al. Butyrolactone autoregulator receptor protein (BarA) as a transcriptional regulator in Streptomyces virginiae. *J Bacteriol* **179**, 6986-6993 (1997).
76. Stratigopoulos, G., Gandecha, A.R. & Cundliffe, E. Regulation of tylosin production and morphological differentiation in Streptomyces fradiae by TylP, a deduced gamma-butyrolactone receptor. *Mol Microbiol* **45**, 735-744 (2002).
77. Kitani, S., Yamada, Y. & Nihira, T. Gene replacement analysis of the butyrolactone autoregulator receptor (FarA) reveals that FarA acts as a Novel regulator in secondary metabolism of Streptomyces lavendulae FRI-5. *J Bacteriol* **183**, 4357-4363. (2001).
78. Folcher, M. et al. Pleiotropic functions of a Streptomyces pristinaespiralis autoregulator receptor in development, antibiotic biosynthesis, and expression of a superoxide dismutase. *J Biol Chem* **276**, 44297-44306. (2001).
79. Kim, H.S. et al. Cloning and characterization of a gene encoding the gamma-butyrolactone autoregulator receptor from Streptomyces clavuligerus. *Arch Microbiol* **182**, 44-50 (2004).
80. Champness, W. in Prokaryotic Development. (eds. Y.V. Brun & L.J. Shimkets) 11-31 (ASM Press, Washington D.C.; 2000).
81. Kim, E.S., Hong, H.J., Choi, C.Y. & Cohen, S.N. Modulation of actinorhodin biosynthesis in Streptomyces lividans by glucose repression of afsR2 gene transcription. *J Bacteriol* **183**, 2198-2203. (2001).
82. Rigali, S. et al. Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by Streptomyces. *EMBO Rep* **9**, 670-675 (2008).
83. Hobbs, G., Frazer, C.M., Gardner, D.C.J., Flett, F. & Oliver, S.G. Pigmented antibiotic production by *Streptomyces coelicolor* A3(2): kinetics and the influence of nutrients. *Journal of General Microbiology* **136**, 2291-2296 (1990).
84. Sola-Landa, A., Moura, R.S. & Martin, J.F. The two-component PhoR-PhoP system controls both primary metabolism and secondary metabolite biosynthesis in Streptomyces lividans. *Proc Natl Acad Sci U S A* **100**, 6133-6138 (2003).
85. Sola-Landa, A., Rodriguez-Garcia, A., Franco-Dominguez, E. & Martin, J.F. Binding of PhoP to promoters of phosphate-regulated genes in Streptomyces coelicolor: identification of PHO boxes. *Mol Microbiol* **56**, 1373-1385 (2005).
86. Hesketh, A., Chen, W.J., Ryding, J., Chang, S. & Bibb, M. The global role of ppGpp synthesis in morphological differentiation and antibiotic production in Streptomyces coelicolor A3(2). *Genome Biol* **8**, R161 (2007).
87. Price, B., Adamidis, T., Kong, R. & Champness, W. A Streptomyces coelicolor antibiotic regulatory gene, absB, encodes an RNase III homolog. *Journal of Bacteriology* **181**, 6142-6151 (1999).
88. Karoonuthaisiri, N., Weaver, D., Huang, J., Cohen, S.N. & Kao, C.M. Regional organization of gene expression in Streptomyces coelicolor. *Gene* **353**, 53-66 (2005).
89. Hesketh, A.R. et al. Primary and secondary metabolism, and post-translational protein modifications, as portrayed by proteomic analysis of Streptomyces coelicolor. *Mol Microbiol* **46**, 917-932 (2002).

90. Vohradsky, J., Branny, P. & Thompson, C.J. Comparative analysis of gene expression on mRNA and protein level during development of Streptomyces cultures by using singular value decomposition. *Proteomics* **7**, 3853-3866 (2007).
91. Jayapal, K.P. et al. Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor. *PLoS ONE* **3**, e2097 (2008).
92. Mehra, S. personal communications. (2002).
93. Aggarwal, S. What's fueling the biotech engine-2007. *Nat Biotechnol* **26**, 1227-1233 (2008).
94. Wurm, F.M. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* **22**, 1393-1398 (2004).
95. Tjio, J.H. & Puck, T.T. Genetics of somatic mammalian cells II. Chromosomal constitution of cells in tissue culture. *Journal of Experimental Medicine* **108**, 259-268 (1958).
96. Urlaub, G. & Chasin, L.A. Isolation of Chinese hamster cell mutants deficient in dihydrofolate reductase activity. *Proc Natl Acad Sci U S A* **77**, 4216-4220 (1980).
97. Potter, M. & Boyce, C.R. Induction of plasma cell neoplasms in strain BALB/c mice with mineral oil and mineral oil adjuvants. *Nature* **193**, 1086-1087 (1962).
98. Barnes, L.M., Bentley, C.M. & Dickson, A.J. Advances in animal cell recombinant protein production: GS-NS0 expression system. *Cytotechnology* **32**, 109-123 (2000).
99. Browne, S.M. & Al-Rubeai, M. Selection methods for high-producing mammalian cell lines. *Trends Biotechnol* **25**, 425-432 (2007).
100. Kim, S.H. & Lee, G.M. Down-regulation of lactate dehydrogenase-A by siRNAs for reduced lactic acid formation of Chinese hamster ovary cells producing thrombopoietin. *Appl Microbiol Biotechnol* **74**, 152-159 (2007).
101. Figueroa, B., Jr. et al. Enhanced cell culture performance using inducible anti-apoptotic genes E1B-19K and Aven in the production of a monoclonal antibody with Chinese hamster ovary cells. *Biotechnol Bioeng* **97**, 877-892 (2007).
102. Ku, S.C., Ng, D.T., Yap, M.G. & Chao, S.H. Effects of overexpression of X-box binding protein 1 on recombinant protein production in Chinese hamster ovary and NS0 myeloma cells. *Biotechnol Bioeng* (2007).
103. Griffin, T.J., Seth, G., Xie, H., Bandhakavi, S. & Hu, W.S. Advancing mammalian cell culture engineering using genome-scale technologies. *Trends Biotechnol* **25**, 401-408 (2007).
104. Wlaschin, K. et al. EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotech. Bioeng.* **91**, 592-606 (2005).
105. Bernstein, J.A., Khodursky, A.B., Lin, P.H., Lin-Chao, S. & Cohen, S.N. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A* **99**, 9697-9702 (2002).
106. De Leon Gatti, M., Wlaschin, K.F., Nissom, P.M., Yap, M. & Hu, W.S. Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. *J Biosci Bioeng* **103**, 82-91 (2007).
107. Seth, G. et al. Molecular portrait of high productivity in recombinant NS0 cells. *Biotechnol Bioeng* **97**, 933-951 (2007).
108. Yee, J.C., de Leon Gatti, M., Philp, R.J., Yap, M. & Hu, W.S. Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment. *Biotechnol Bioeng* **99**, 1186-1204 (2008).
109. Lee, M.S., Kim, K.W., Kim, Y.H. & Lee, G.M. Proteome analysis of antibody-expressing CHO cells in response to hyperosmotic pressure. *Biotechnol Prog* **19**, 1734-1741 (2003).

110. Swiderek, H. & Al-Rubeai, M. Functional genome-wide analysis of antibody producing NS0 cell line cultivated at different temperatures. *Biotechnol Bioeng* **98**, 616-630 (2007).
111. Korke, R. et al. Large scale gene expression profiling of metabolic shift of mammalian cells in culture. *J Biotechnol* **107**, 1-17 (2004).
112. Dinnis, D.M. et al. Functional proteomic analysis of GS-NS0 murine myeloma cell lines with varying recombinant monoclonal antibody production rate. *Biotechnol Bioeng* **94**, 830-841 (2006).
113. Jeffery, I.B., Higgins, D.G. & Culhane, A.C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics* **7**, 359 (2006).
114. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).
115. Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G. & Davis, R.W. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A* **102**, 12837-12842 (2005).
116. Hannah, M.A., Redestig, H., Leisse, A. & Willmitzer, L. Global mRNA changes in microarray experiments. *Nat Biotechnol* **26**, 741-742 (2008).
117. Curtis, R.K., Oresic, M. & Vidal-Puig, A. Pathways to the analysis of microarray data. *Trends Biotechnol* **23**, 429-435 (2005).
118. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
119. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273 (2003).
120. Ben-Shaul, Y., Bergman, H. & Soreq, H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics* **21**, 1129-1137 (2005).
121. Tian, L. et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* **102**, 13544-13549 (2005).
122. Bild, A.H. et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353-357 (2006).
123. Smales, C.M. et al. Comparative proteomic analysis of GS-NSO murine myeloma cell lines with varying recombinant monoclonal antibody production rate. *Biotechnology & Bioengineering* **88**, 474-488 (2004).
124. Yee, J.C., Gerdtzen, Z.P. & Hu, W.S. Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells. *Biotechnol Bioeng* **(in press)** (2008).
125. Dinnis, D.M. & James, D.C. Engineering mammalian cell factories for improved recombinant monoclonal antibody production: lessons from nature? *Biotechnol Bioeng* **91**, 180-189 (2005).
126. Salonen, J.M., Valmu, L., Ronnholm, G., Kalkkinen, N. & Vihinen, M. Proteome analysis of B cell maturation. *Proteomics* **6**, 5152-5268 (2006).
127. Ollila, J. & Vihinen, M. Immunological systems biology: Gene expression analysis of B-cell development in ramos B-cells. *Molecular Immunology* **44**, 3537-3551 (2007).
128. Shaffer, A.L. et al. XBP1, downstream of Blimp-1, expands the secretory apparatus and other organelles, and increases protein synthesis in plasma cell differentiation. *Immunity* **21**, 81-93 (2004).

129. Lee, J. in Chemical Engineering and Materials Science, Vol. PhD 161 (University of Minnesota, Twin Cities; 2005).
130. Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P. From data mining to knowledge discovery: an overview. *Advances in knowledge discovery and data mining table of contents*, 1-34 (1996).
131. Cheung, J.T.Y. & Stephanopoulos, G. Representation of process trends- Part II. The problem of scale and qualitative scaling. *COMP. CHEM. ENG.* **14**, 511-539 (1990).
132. Cheung, J.T.Y. & Stephanopoulos, G. Representation of process trends--part I. A formal representation framework. *COMP. CHEM. ENG.* **14**, 495-510 (1990).
133. Bakshi, B.R. & Stephanopoulos, G. Representation of Process Trends. 4. Induction of Real-Time Patterns from Operating Data for Diagnosis and Supervisory Control. *Computers & Chemical Engineering* **18**, 303-332 (1994).
134. Bakshi, B.R. & Stephanopoulos, G. Representation of process trends—3. Multi-scale extraction of trends from process data. *Computers and Chemical Engineering*, 267-302 (1994).
135. Moult, J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos Trans R Soc Lond B Biol Sci* **361**, 453-458 (2006).
136. Tompa, M. et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**, 137-144 (2005).
137. Huang, J., Nanami, H., Kanda, A., Shimizu, H. & Shioya, S. Classification of fermentation performance by multivariate analysis based on mean hypothesis testing. *J Biosci Bioeng* **94**, 251-257 (2002).
138. Kamimura, R.T., Bicciato, S., Shimizu, H., Alford, J. & Stephanopoulos, G. Mining of biological data I: identifying discriminating features via mean hypothesis testing. *Metab Eng* **2**, 218-227 (2000).
139. Sakoe, H. & Chiba, S. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43-49 (1978).
140. Keogh, E., Chakrabarti, K., Pazzani, M. & Mehrotra, S. Locally adaptive dimensionality reduction for indexing large time series databases. *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* **30**, 151-162 (2001).
141. Keogh, E. & Ratanamahatana, C.A. Exact Indexing of dynamic time warping. *Knowledge and Information Systems* **7**, 358-386 (2005).
142. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).
143. Buck, K.K., Subramanian, V. & Block, D.E. Identification of critical batch operating parameters in fed-batch recombinant E. coli fermentations using decision tree analysis. *Biotechnol Prog* **18**, 1366-1376 (2002).
144. Coleman, M.C., Buck, K.K. & Block, D.E. An integrated approach to optimization of Escherichia coli fermentations using historical data. *Biotechnol Bioeng* **84**, 274-285 (2003).
145. Stephanopoulos, G., Locher, G., Duff, M.J., Kamimura, R. & Stephanopoulos, G. Fermentation database mining by pattern recognition. *Biotechnol Bioeng* **53**, 443-452 (1997).
146. Tai, Y.C. & Speed, T.P. A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics* **34**, 2387-2412 (2006).

147. Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K. & Jaakkola, T.S. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci U S A* **100**, 10146-10151 (2003).
148. Ringner, M. What is principal component analysis? *Nat Biotechnol* **26**, 303-304 (2008).
149. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).
150. Kamimura, R.T., Bicciato, S., Shimizu, H., Alford, J. & Stephanopoulos, G. Mining of biological data II: assessing data structure and class homogeneity by cluster analysis. *Metab Eng* **2**, 228-238 (2000).
151. Agrawal, R. & Srikant, R. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases, VLDB* **1215**, 487499 (1994).
152. Seno, M. & Karypis, G. LPMiner: An algorithm for finding frequent itemsets using length-decreasing support constraint. *Proceeding of the 2001 IEEE International Conference on Data Mining*, 505-512 (2001).
153. Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31**, 264-323 (1999).
154. Ahlberg, C. Spotfire: an information exploration environment. *ACM SIGMOD Record* **25**, 25-29 (1996).
155. D'Haeseleer, P. How does gene expression clustering work? *Nat Biotechnol* **23**, 1499-1501 (2005).
156. Lapointe, J. et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816 (2004).
157. Duda, R.O., Hart, P.E. & Stork, D.G. Pattern Classification, Edn. 2nd. (Wiley-Interscience, 2000).
158. Slonim, N., Atwal, G.S., Tkacik, G. & Bialek, W. Information-based clustering. *Proc Natl Acad Sci U S A* **102**, 18297-18302 (2005).
159. Lin, J., Keogh, E., Lonardi, S. & Chiu, B. A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, 2-11 (2003).
160. Glassey, J., Montague, G.A., Ward, A.C. & Kara, B.V. Enhanced supervision of recombinant E. coli fermentations via artificial neural networks. *Process Biochemistry* **29**, 387-398 (1994).
161. Glassey, J., Montague, G.A., Ward, A.C. & Kara, B.V. Artificial neural network based experimental design procedures for enhancing fermentation development. *Biotechnol Bioeng* **44**, 397-405 (1994).
162. Bachinger, T., Riese, U., Eriksson, R.K. & Mandenius, C.F. Electronic nose for estimation of product concentration in mammalian cell cultivation. *Bioprocess and Biosystems Engineering* **23**, 637-642 (2000).
163. Vlassides, S., Ferrier, J.G. & Block, D.E. Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information. *Biotechnol Bioeng* **73**, 55-68 (2001).
164. Coleman, M.C. & Block, D.E. Retrospective optimization of time-dependent fermentation control strategies using time-independent historical data. *Biotechnol Bioeng* **95**, 412-423 (2006).
165. Kirdar, A.O., Conner, J.S., Baclaski, J. & Rathore, A.S. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. *Biotechnol Prog* **23**, 61-67 (2007).
166. Vapnik, V.N. Statistical learning theory. (Wiley, New York; 1998).

167. Vapnik, V.N. The Nature of Statistical Learning Theory, Edn. 2nd. (Springer, 2000).
168. Li, Y. & Long, P.M. The Relaxed Online Maximum Margin Algorithm. *Machine Learning* **46**, 361-387 (2002).
169. Weinberger, K., Blitzer, J. & Saul, L. Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems* **18**, 1473-1480 (2006).
170. Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I. & Noble, W.S. A statistical framework for genomic data fusion. *Bioinformatics* **20**, 2626-2635 (2004).
171. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., El Ghaoui, L. & Jordan, M.I. Learning the Kernel Matrix with Semidefinite Programming. *The Journal of Machine Learning Research* **5**, 27-72 (2004).
172. Brown, M.P. et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267 (2000).
173. Tong, S. & Koller, D. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research* **2**, 45-66 (2002).
174. Tong, S. & Chang, E. Support vector machine active learning for image retrieval. *Proc of the ninth ACM Intl Conf on Multimedia* **9**, 107-118 (2001).
175. Noble, W.S. What is a support vector machine? *Nat Biotechnol* **24**, 1565-1567 (2006).
176. Vapnik, V.N. The nature of statistical learning theory. (Springer, New York; 1995).
177. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* **2**, 121-167 (1998).
178. Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. & Williamson, R.C. Estimating the Support of a High-Dimensional Distribution. *Neural Computation* **13**, 1443-1471 (2001).
179. Weston, J. & Watkins, C. Support vector machines for multi-class pattern recognition. *Proceedings of the Seventh European Symposium On Artificial Neural Networks* **4**, 6 (1999).
180. Kerr, M.K. & Churchill, G.A. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc Natl Acad Sci USA* **98**, 8961-8965 (2001).
181. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning* **52**, 91-118 (2003).
182. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* **2**, 1137-1145 (1995).
183. Ambroise, C. & McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **99**, 6562-6566 (2002).
184. Price, M.N., Arkin, A.P. & Alm, E.J. The life-cycle of operons. *PLoS Genet* **2**, e96 (2006).
185. Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. & Collado-Vides, J. Operons in Escherichia coli: genomic analyses and predictions. *Proc Natl Acad Sci U S A* **97**, 6652-6657 (2000).
186. Ermolaeva, M.D., White, O. & Salzberg, S.L. Prediction of operons in microbial genomes. *Nucleic Acids Res* **29**, 1216-1221 (2001).
187. Price, M.N., Huang, K.H., Alm, E.J. & Arkin, A.P. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* **33**, 880-892 (2005).

188. Sabatti, C., Rohlin, L., Oh, M.K. & Liao, J.C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res* **30**, 2886-2893 (2002).
189. Westover, B.P., Buhler, J.D., Sonnenburg, J.L. & Gordon, J.I. Operon prediction without a training set. *Bioinformatics* **21**, 880-888 (2005).
190. Wang, L., Trawick, J.D., Yamamoto, R. & Zamudio, C. Genome-wide operon prediction in Staphylococcus aureus. *Nucleic Acids Res* **32**, 3689-3702 (2004).
191. Moreno-Hagelsieb, G. & Collado-Vides, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl 1**, S329-336 (2002).
192. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. in Proc 17th Int Conf Machine Learning (Morgan Kaufmann, Stanford, CA; 2000).
193. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. A probabilistic learning approach to whole-genome operon prediction. *Proc Int Conf Intell Syst Mol Biol* **8**, 116-127 (2000).
194. Zhang, G.Q., Cao, Z.W., Luo, Q.M., Cai, Y.D. & Li, Y.X. Operon prediction based on SVM. *Comput Biol Chem* **30**, 233-240 (2006).
195. Tjaden, B., Haynor, D.R., Stolyar, S., Rosenow, C. & Kolker, E. Identifying operons and untranslated regions of transcripts using Escherichia coli RNA expression analysis. *Bioinformatics* **18**, S337-S344 (2002).
196. De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N. & Miyano, S. Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. *Pac. Symp. Biocomput* **9**, 276–287 (2004).
197. Bockhorst, J., Craven, M., Page, D., Shavlik, J. & Glasner, J. A Bayesian network approach to operon prediction. *Bioinformatics* **19**, 1227-1235 (2003).
198. Mehra, S. et al. A framework to analyze multiple time series data - a case study with streptomyces coelicolor. *J Ind Microbiol Biotechnol* **33**, 159-172 (2006).
199. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
200. Huang, J. et al. Cross-regulation among disparate antibiotic biosynthetic pathways of Streptomyces coelicolor. *Mol Microbiol* **58**, 1276-1287 (2005).
201. Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O. & Salzberg, S.L. Prediction of transcription terminators in bacterial genomes. *J Mol Biol* **301**, 27-33 (2000).
202. Joachims, T. in Advances in kernel methods - support vector learning. (eds. B. Scholkofp, C.J.C. Burges & A.J. Smola) 169-184 (M.I.T. Press, 1999).
203. Lee, J.H. & Lin, C.J. (Department of Computer Science and Information Engineering, National Taiwan University, Taipei; 2000).
204. Takeuchi, T., Sawada, H., Tanaka, F. & Matsuda, I. Phylogenetic analysis of Streptomyces spp. causing potato scab based on 16S rRNA sequences. *Int J Syst Bacteriol* **46**, 476-479 (1996).
205. Leblond, P., Redenbach, M. & Cullum, J. Physical map of the Streptomyces lividans 66 genome and comparison with that of the related strain Streptomyces coelicolor A3(2). *J Bacteriol* **175**, 3422-3429 (1993).
206. Kohavi, R. in Proceedings of Fourteenth International Conference on Artificial Intelligence 1137-1143Montreal, CA; 1995).
207. Fawcett, T. 38 (HP Laboratories, Palo Alto; 2004).
208. Bockhorst, J. et al. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics* **19 Suppl 1**, i34-43 (2003).

209. Takano, H., Obitsu, S., Beppu, T. & Ueda, K. Light-induced carotenogenesis in Streptomyces coelicolor A3(2): identification of an extracytoplasmic function sigma factor that directs photodependent transcription of the carotenoid biosynthesis gene cluster. *J Bacteriol* **187**, 1825-1832 (2005).
210. Lee, E.J., Cho, Y.H., Kim, H.S. & Roe, J.H. Identification of sigmaB-dependent promoters using consensus-directed search of Streptomyces coelicolor genome. *J Microbiol* **42**, 147-151 (2004).
211. Cho, Y.H., Lee, E.J., Ahn, B.E. & Roe, J.H. SigB, an RNA polymerase sigma factor required for osmoprotection and proper differentiation of Streptomyces coelicolor. *Mol Microbiol* **42**, 205-214 (2001).
212. Hu, D.S., Hood, D.W., Heidstra, R. & Hodgson, D.A. The expression of the trpD, trpC and trpBA genes of Streptomyces coelicolor A3(2) is regulated by growth rate and growth phase but not by feedback repression. *Mol Microbiol* **32**, 869-880 (1999).
213. Sevcikova, B. & Kormanec, J. Activity of the Streptomyces coelicolor stress-response sigma factor sigmaH is regulated by an anti-sigma factor. *FEMS Microbiol Lett* **209**, 229-235 (2002).
214. Sevcikova, B., Benada, O., Kofronova, O. & Kormanec, J. Stress-response sigma factor sigma(H) is essential for morphological differentiation of Streptomyces coelicolor A3(2). *Arch Microbiol* **177**, 98-106 (2001).
215. Kormanec, J., Sevcikova, B., Halgasova, N., Knirschova, R. & Rezuchova, B. Identification and transcriptional characterization of the gene encoding the stress-response sigma factor sigma(H) in streptomyces coelicolor A3(2). *FEMS Microbiol Lett* **189**, 31-38 (2000).
216. Bralley, P. & Jones, G.H. Organization and expression of the polynucleotide phosphorylase gene (pnp) of Streptomyces: Processing of pnp transcripts in Streptomyces antibioticus. *J Bacteriol* **186**, 3160-3172 (2004).
217. Chang, S.A., Bralley, P. & Jones, G.H. The absB Gene Encodes a Double Strand-specific Endoribonuclease That Cleaves the Read-through Transcript of the rpsO-pnp Operon in Streptomyces coelicolor. *J Biol Chem* **280**, 33213-33219 (2005).
218. Bucca, G., Brassington, A.M., Hotchkiss, G., Mersinias, V. & Smith, C.P. Negative feedback regulation of dnaK, clpB and lon expression by the DnaK chaperone machine in Streptomyces coelicolor, identified by transcriptome and in vivo DnaK-depletion analysis. *Mol Microbiol* **50**, 153-166 (2003).
219. Fornwald, J.A., Schmidt, F.J., Adams, C.W., Rosenberg, M. & Brawner, M.E. Two promoters, one inducible and one constitutive, control transcription of the Streptomyces lividans galactose operon. *Proc Natl Acad Sci U S A* **84**, 2130-2134 (1987).
220. Vierling, S., Weber, T., Wohlleben, W. & Muth, G. Transcriptional and mutational analyses of the Streptomyces lividans recX gene and its interference with RecA activity. *J Bacteriol* **182**, 4005-4011 (2000).
221. Tieleman, L.N., van Wezel, G.P., Bibb, M.J. & Kraal, B. Growth phase-dependent transcription of the Streptomyces ramocissimus tuf1 gene occurs from two promoters. *J Bacteriol* **179**, 3619-3624 (1997).
222. Laing, E., Mersinias, V., Smith, C.P. & Hubbard, S.J. Analysis of gene expression in operons of Streptomyces coelicolor. *Genome Biol* **7**, R46 (2006).
223. Yada, T., Nakao, M., Totoki, Y. & Nakai, K. Modeling and predicting transcriptional units of Escherichia coligenes using hidden Markov models. *Bioinformatics* **15**, 987-993 (1999).

224. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96**, 2896-2901 (1999).
225. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. & Maltsev, N. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol* **1**, 93-108 (1999).
226. Riley, M. Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**, 862-952 (1993).
227. Limauro, D., Avitabile, A., Cappellano, C., Puglia, A.M. & Bruni, C.B. Cloning and characterization of the histidine biosynthetic gene cluster of Streptomyces coelicolor A3(2). *Gene* **90**, 31-41 (1990).
228. Carere, A., Russi, S., Bignami, M. & Sermonti, G. An operon for histidine biosynthesis in Streptomyces coelicolor. I. Genetic evidence. *Mol Gen Genet* **123**, 219-224 (1973).
229. Fink, D., Weissschuh, N., Reuther, J., Wohlleben, W. & Engels, A. Two transcriptional regulators GlnR and GlnRII are involved in regulation of nitrogen metabolism in Streptomyces coelicolor A3(2). *Mol Microbiol* **46**, 331-347 (2002).
230. Xiao, G., Martinez-Vaz, B., Pan, W. & Khodursky, A.B. Operon information improves gene expression estimation for cDNA microarrays. *BMC Genomics* **7**, 87 (2006).
231. Kuramochi, M. & Karypis, G. in Proceedings of the IEEE 2nd International Symposium on Bioinformatics and Bioengineering Conference 191-200Bethesda, MD; 2001).
232. Salgado, H. et al. RegulonDB (version 3.0): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res* **28**, 65-67 (2000).
233. Banerjee, S., Chalissery, J., Bandey, I. & Sen, R. Rho-dependent transcription termination: more questions than answers. *J Microbiol* **44**, 11-22 (2006).
234. Brendel, V. & Trifonov, E.N. Computer-aided mapping of DNA-protein interaction sites. *CODATA Bulletin*, 17-20 (1984).
235. Brendel, V. & Trifonov, E.N. A computer algorithm for testing potential prokaryotic terminators. *Nucleic Acids Research* **12**, 4411-4427 (1984).
236. Unniraman, S., Prakash, R. & Nagaraja, V. Conserved economics of transcription termination in eubacteria. *Nucleic Acids Research* **30**, 675-684 (2002).
237. Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J. & Kasif, S. Computational Identification of Operons in Microbial Genomes. *Genome Res* **12**, 1221-1230 (2002).
238. Noble, W.S. in Kernel methods in computational biology. (eds. B. Scholkofp, K. Tsuda & J. Vert) 71-92 (M.I.T. Press, 2004).
239. Braga-Neto, U.M. & Dougherty, E.R. Is cross-validation valid for small-sample microarray classification? *Bioinformatics* **20**, 374-380 (2004).
240. Baba, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006 0008 (2006).
241. Ishii, N. et al. Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science* **316**, 593-597 (2007).
242. Gust, B., Challis, G.L., Fowler, K., Kieser, T. & Chater, K.F. PCR-targeted Streptomyces gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc Natl Acad Sci U S A* **100**, 1541-1546 (2003).
243. Akutsu, T., Miyano, S. & Kuhara, S. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*, 17-28 (1999).
244. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601-620 (2000).
245. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17 Suppl 1**, S215-224 (2001).

246. de Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* **9**, 67-103 (2002).
247. McAdams, H.H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A* **94**, 814-819 (1997).
248. McAdams, H.H. & Arkin, A. Simulation of prokaryotic genetic circuits. *Annu Rev Biophys Biomol Struct* **27**, 199-224 (1998).
249. Liang, S., Fuhrman, S. & Somogyi, R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, 18-29 (1998).
250. Ideker, T.E., Thorsson, V. & Karp, R.M. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac Symp Biocomput*, 305-316 (2000).
251. Murphy, K. & Mian, S. (University of California, Computer Science Division, Berkeley; 1999).
252. Zou, M. & Conzen, S.D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **21**, 71-79 (2005).
253. Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166-176 (2003).
254. Bar-Joseph, Z. et al. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**, 1337-1342 (2003).
255. Xue, H. et al. A modular network model of aging. *Mol Syst Biol* **3**, 147 (2007).
256. Barabasi, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-113 (2004).
257. Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**, 450-461 (2007).
258. Takano, E., Chakraburtty, R., Nihira, T., Yamada, Y. & Bibb, M.J. A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in Streptomyces coelicolor A3(2). *Mol Microbiol* **41**, 1015-1028 (2001).
259. Mehra, S., Charaniya, S., Takano, E. & Hu, W.S. A bistable gene switch for antibiotic biosynthesis: the butyrolactone regulon in Streptomyces coelicolor. *PLoS ONE* **3**, e2724 (2008).
260. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet* **31**, 64-68 (2002).
261. Prill, R.J., Iglesias, P.A. & Levchenko, A. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol* **3**, e343 (2005).
262. Gautier, L., Cope, L., Bolstad, B.M. & Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004).
263. Affymetrix 2001).
264. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).
265. Hardin, J., Mitani, A., Hicks, L. & VanKoten, B. A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics* **8**, 220 (2007).
266. Jayapal, K.P., Lian, W., Glod, F., Sherman, D.H. & Hu, W.S. Comparative genomic hybridizations reveal absence of large Streptomyces coelicolor genomic islands in Streptomyces lividans. *BMC Genomics* **8**, 229 (2007).
267. Mersinias, V. in Department of Biomolecular Sciences, Vol. PhD (University of Manchester 2004).

268. Lee, E.J. et al. A master regulator sigma governs osmotic and oxidative response as well as differentiation via a network of sigma factors in Streptomyces coelicolor. *Mol Microbiol* **57**, 1252-1264 (2005).
269. Elliot, M.A. et al. The chaplins: a family of hydrophobic cell-surface proteins involved in aerial mycelium formation in Streptomyces coelicolor. *Genes Dev* **17**, 1727-1740 (2003).
270. Fong, R. et al. Characterization of a large, stable, high-copy-number Streptomyces plasmid that requires stability and transfer functions for heterologous polyketide overproduction. *Appl Environ Microbiol* **73**, 1296-1307 (2007).
271. San Paolo, S., Huang, J., Cohen, S.N. & Thompson, C.J. rag genes: novel components of the RamR regulon that trigger morphological differentiation in Streptomyces coelicolor. *Mol Microbiol* **61**, 1167-1186 (2006).
272. Martin, D. et al. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol* **5**, R101 (2004).
273. Resnik, P. in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 14 448-453Montreal; 1995).
274. Lin, D. in Proceedings of the 15th International Conference on Machine Learning 296-304San Francisco, CA; 1998).
275. Jiang, J.J. & Conrath, D.W. in Proceedings of the International Conference on Research in Computational Linguistics, Vol. 33Taiwan; 1998).
276. Frohlich, H., Speer, N., Poustka, A. & Beissbarth, T. GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics* **8**, 166 (2007).
277. Pertea, M., Ayanbule, K., Smedinghoff, M. & Salzberg, S.L. OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res* (2008).
278. Ahn, B.E. et al. Nur, a nickel-responsive regulator of the Fur family, regulates superoxide dismutases and nickel transport in Streptomyces coelicolor. *Mol Microbiol* **59**, 1848-1858 (2006).
279. Hoskisson, P.A., Rigali, S., Fowler, K., Findlay, K.C. & Buttner, M.J. DevA, a GntR-like transcriptional regulator required for development in Streptomyces coelicolor. *J Bacteriol* **188**, 5014-5023 (2006).
280. Borovok, I., Gorovitz, B., Schreiber, R., Aharonowitz, Y. & Cohen, G. Coenzyme B12 controls transcription of the Streptomyces class Ia ribonucleotide reductase nrdABS operon via a riboswitch mechanism. *J Bacteriol* **188**, 2512-2520 (2006).
281. Shin, J.H., Oh, S.Y., Kim, S.J. & Roe, J.H. The zinc-responsive regulator Zur controls a zinc uptake system and some ribosomal proteins in Streptomyces coelicolor A3(2). *J Bacteriol* **189**, 4070-4077 (2007).
282. Owen, G.A., Pascoe, B., Kallifidas, D. & Paget, M.S. Zinc-responsive regulation of alternative ribosomal protein genes in Streptomyces coelicolor involves zur and sigmaR. *J Bacteriol* **189**, 4078-4086 (2007).
283. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).
284. Basso, K. et al. Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37**, 382-390 (2005).
285. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
286. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651-654 (2000).

287. Yook, S.H., Oltvai, Z.N. & Barabasi, A.L. Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928-942 (2004).
288. Huss, J.W., 3rd et al. A gene wiki for community annotation of gene function. *PLoS Biol* **6**, e175 (2008).
289. Margolin, A.A. & Califano, A. Theory and limitations of genetic network inference from microarray data. *Ann N Y Acad Sci* **1115**, 51-72 (2007).
290. Laing, E., Sidhu, K. & Hubbard, S.J. Predicted transcription factor binding sites as predictors of operons in Escherichia coli and Streptomyces coelicolor. *BMC Genomics* **9**, 79 (2008).
291. Khoo, S.H., Falciani, F. & Al-Rubeai, M. A genome-wide transcriptional analysis of producer and non-producer NS0 myeloma cell lines. *Biotechnol Appl Biochem* **47**, 85-95 (2007).
292. Alete, D.E. et al. Proteomic analysis of enriched microsomal fractions from GS-NS0 murine myeloma cells with varying secreted recombinant monoclonal antibody productivities. *Proteomics* **5**, 4689-4704 (2005).
293. Smales, C.M. et al. Comparative proteomic analysis of GS-NS0 murine myeloma cell lines with varying recombinant monoclonal antibody production rate. *Biotechnol Bioeng* **88**, 474-488 (2004).
294. Goeman, J.J. & Buhlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980-987 (2007).
295. Doniger, S.W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).
296. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann Appl Statist* **1**, 107-129 (2007).
297. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31**, 19-20 (2002).
298. Reich, M. et al. GenePattern 2.0. *Nat Genet* **38**, 500-501 (2006).
299. Bigay, J., Gounon, P., Robineau, S. & Antonny, B. Lipid packing sensed by ArfGAP1 couples COPI coat disassembly to membrane bilayer curvature. *Nature* **426**, 563-566 (2003).
300. Shields, R.L. et al. Lack of fucose on human IgG1 N-linked oligosaccharide improves binding to human Fcgamma RIII and antibody-dependent cellular toxicity. *J Biol Chem* **277**, 26733-26740 (2002).
301. Ross, J.L., Ali, M.Y. & Warshaw, D.M. Cargo transport: molecular motors navigate a complex cytoskeleton. *Curr Opin Cell Biol* (2008).
302. Stamnes, M. Regulating the actin cytoskeleton during vesicular transport. *Curr Opin Cell Biol* **14**, 428-433 (2002).
303. Song, M.S. et al. The tumour suppressor RASSF1A regulates mitosis by inhibiting the APC-Cdc20 complex. *Nat Cell Biol* **6**, 129-137 (2004).
304. Jensen, R.E., Dunn, C.D., Youngman, M.J. & Sesaki, H. Mitochondrial building blocks. *Trends Cell Biol* **14**, 215-218 (2004).
305. Bibb, M.J., Van Etten, R.A., Wright, C.T., Walberg, M.W. & Clayton, D.A. Sequence and gene organization of mouse mitochondrial DNA. *Cell* **26**, 167-180 (1981).
306. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat Genet* **33**, 49-54 (2003).
307. Hooker, A.D. et al. Constraints on the transport and glycosylation of recombinant IFN-gamma in Chinese hamster ovary and insect cells. *Biotechnol Bioeng* **63**, 559-572 (1999).

308. Cai, H., Reinisch, K. & Ferro-Novick, S. Coats, tethers, Rabs, and SNAREs work together to mediate the intracellular destination of a transport vesicle. *Dev Cell* **12**, 671-682 (2007).
309. Ungar, D. et al. Characterization of a mammalian Golgi-localized protein complex, COG, that is required for normal Golgi morphology and function. *J Cell Biol* **157**, 405-415 (2002).
310. Dell'Angelica, E.C. et al. AP-3: an adaptor-like protein complex with ubiquitous expression. *Embo J* **16**, 917-928 (1997).
311. Subramaniam, V.N., Peter, F., Philp, R., Wong, S.H. & Hong, W. GS28, a 28-kilodalton Golgi SNARE that participates in ER-Golgi transport. *Science* **272**, 1161-1163 (1996).
312. Lowe, S.L., Peter, F., Subramaniam, V.N., Wong, S.H. & Hong, W. A SNARE involved in protein transport through the Golgi apparatus. *Nature* **389**, 881-884 (1997).
313. Sollner, T., Bennett, M.K., Whiteheart, S.W., Scheller, R.H. & Rothman, J.E. A protein assembly-disassembly pathway in vitro that may correspond to sequential steps of synaptic vesicle docking, activation, and fusion. *Cell* **75**, 409-418 (1993).
314. Lee, Y.C., Kurosawa, N., Hamamoto, T., Nakaoka, T. & Tsuji, S. Molecular cloning and expression of Gal beta 1,3GalNAc alpha 2,3-sialyltransferase from mouse brain. *Eur J Biochem* **216**, 377-385 (1993).
315. DeBose-Boyd, R.A. et al. Transport-dependent proteolysis of SREBP: relocation of site-1 protease from Golgi to ER obviates the need for SREBP transport to Golgi. *Cell* **99**, 703-712 (1999).
316. Tetsu, O. & McCormick, F. Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature* **398**, 422-426 (1999).
317. Dryden, S.C., Nahhas, F.A., Nowak, J.E., Goustin, A.S. & Tainsky, M.A. Role for human SIRT2 NAD-dependent deacetylase activity in control of mitotic exit in the cell cycle. *Mol Cell Biol* **23**, 3173-3185 (2003).
318. Shivakumar, L., Minna, J., Sakamaki, T., Pestell, R. & White, M.A. The RASSF1A tumor suppressor blocks cell cycle progression and inhibits cyclin D1 accumulation. *Mol Cell Biol* **22**, 4309-4318 (2002).
319. Fang, G., Yu, H. & Kirschner, M.W. The checkpoint protein MAD2 and the mitotic regulator CDC20 form a ternary complex with the anaphase-promoting complex to control anaphase initiation. *Genes Dev* **12**, 1871-1883 (1998).
320. Li, Y. & Benezra, R. Identification of a human mitotic checkpoint gene: hsMAD2. *Science* **274**, 246-248 (1996).
321. Wohlschlegel, J.A. et al. Inhibition of eukaryotic DNA replication by geminin binding to Cdt1. *Science* **290**, 2309-2312 (2000).
322. Sacher, M. et al. TRAPP, a highly conserved novel complex on the cis-Golgi that mediates vesicle docking and fusion. *Embo J* **17**, 2494-2503 (1998).
323. Allan, B.B., Moyer, B.D. & Balch, W.E. Rab1 recruitment of p115 into a cis-SNARE complex: programming budding COPII vesicles for fusion. *Science* **289**, 444-448 (2000).
324. Tisdale, E.J. & Balch, W.E. Rab2 is essential for the maturation of pre-Golgi intermediates. *J Biol Chem* **271**, 29372-29379 (1996).
325. Fucini, R.V., Chen, J.L., Sharma, C., Kessels, M.M. & Stamnes, M. Golgi vesicle proteins are linked to the assembly of an actin complex defined by mAbp1. *Mol Biol Cell* **13**, 621-631 (2002).
326. Schugerl, K. Progress in monitoring, modeling and control of bioprocesses during the last 20 years. *J Biotechnol* **85**, 149-173 (2001).

327. Wlaschin, K.F. & Hu, W.S. Fedbatch culture and dynamic nutrient feeding. *Adv Biochem Eng Biotechnol* **101**, 43-74 (2006).
328. Scholkopf, B., Smola, A.J., Williamson, R.C. & Bartlett, P.L., Vol. 12 1207-1245 (MIT Press, 2000).
329. Chang, C.C. & Lin, C.J. LIBSVM: a library for support vector machines. *Software available at http://www. csie. ntu. edu. tw/cjlin/libsvm* **80**, 604-611 (2001).
330. Beyer, K.S., Goldstein, J., Ramakrishnan, R. & Shaft, U. When is "nearest neighbor" meaningful? *Proceedings of the 7th International Conference on Database Theory*, 217-235 (1999).
331. Kamimura, R., Konstantinov, K. & Stephanopoulos, G. Knowledge-based systems, artificial neural networks and pattern recognition: applications to biotechnological processes. *Curr Opin Biotechnol* **7**, 231-234 (1996).
332. Ignova, M., Montague, G.A., Ward, A.C. & Glassey, J. Fermentation seed quality analysis with self-organising neural networks. *Biotechnol Bioeng* **64**, 82-91 (1999).
333. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-453 (2003).
334. Jeong, D. et al. Blocking of acidosis-mediated apoptosis by a reduction of lactate dehydrogenase activity through antisense mRNA expression. *Biochem Biophys Res Commun* **289**, 1141-1149 (2001).
335. Zhou, W., Rehm, J. & Hu, W.S. High viable cell concentration fed-batch cultures of hybridoma cells through on-line nutrient feeding. *Biotechnol Bioeng* **46**, 579-587 (1995).
336. Wlaschin, K.F. & Hu, W.S. Engineering cell metabolism for high-density cell culture via manipulation of sugar transport. *J Biotechnol* **131**, 168-176 (2007).
337. Rommel, S. & Schuppert, A. Data Mining for Bioprocess Optimization. *Engineering in Life Sciences* **4**, 266-270 (2004).
338. Dantas, G., Sommer, M.O., Oluwasegun, R.D. & Church, G.M. Bacteria subsisting on antibiotics. *Science* **320**, 100-103 (2008).
339. Cohen, K.B. & Hunter, L. Getting started in text mining. *PLoS Comput Biol* **4**, e20 (2008).
340. Sharan, R. & Ideker, T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol* **24**, 427-433 (2006).
341. Chechik, G. et al. Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat Biotechnol* **26**, 1251-1259 (2008).
342. Draghici, S. et al. A systems biology approach for pathway level analysis. *Genome Res* **17**, 1537-1545 (2007).