Phenotypic Diversity as a Tool to Guide and Optimize Random Strain Improvement Approaches

by

Daniel Klein-Marcuschamer

B.S., Chemical Engineering
University of Texas at Austin, 2005

Submitted to the Department of Chemical Engineering
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Chemical Engineering
at the
Massachusetts Institute of Technology

April 15th, 2009
[JUNE]

Signature of Author ________
Daniel Klein-Marcuschamer
Department of Chemical Engineering

Certified by____________
Gregory Stephanopoulos
Professor of Chemical Engineering
Thesis Supervisor

Accepted by____
William Deen
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

- Blank Page -

Phenotypic Diversity as a Tool to Guide and Optimize Random Strain Improvement Approaches

by

Daniel Klein-Marcuschamer

Submitted to the Department of Chemical Engineering on April 15th, 2009

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Chemical Engineering

Abstract

A sustainable economy will depend, if only partly, on efficient renewable-feedstock conversion to chemicals and fuels, and advances in that direction have relied and will continue to rely on strain engineering. Traditional methods comprising directed genetic modifications (i.e. targeting specific genes) have been quite successful in improving several phenotypes of industrial interest. Evolutionary approaches have also contributed much to these efforts and are gaining attention in particular for addressing complex phenotypes. Most commonly, mutagenesis and selection has been the method of choice, but many other random search-based approaches for phenotypic alteration have been developed in recent years. One such method, transcriptional engineering, relies on transcriptome-wide modifications that can be exploited to better complex traits. The initial aim of this work was to build upon the idea of transcriptional engineering in bacteria, which had been tried in our laboratory through mutagenesis of the principal sigma factor, sigma D.

Initially, we explored new targets for transcriptional engineering. Using error-prone PCR, we constructed libraries of several stress-related sigma factors in *Escherichia coli* (sigma S, sigma E, and sigma H) and screened them for phenotypes of interest. We also considered the alpha subunit of the RNA polymerase as a tool for phenotypic alteration, and fruitfully used it to improve butanol and solvent tolerance, accumulation of hyaluronic acid, and L-tyrosine production. Carefull assessment of the sigma and alpha libraries for a few phenotypes revealed that not all the targets were equally useful, and

that succeeding at improving one trait does not imply that the same target could be used to improve a different one.

We also extended the use of the already-proven target, sigma D, to a new species, *Lactobacillus plantarum.* We constructed random mutagenesis libraries of this gene and produced a cell library that was selected in conditions relevant to the production of lactic acid. This chemical has attracted attention for its use as a food and pharmaceutical additive, and in the production of specialty chemicals and biodegradable plastics. We isolated two mutants with significantly higher growth rates in media acidified with both lactic and hydrochloric acid, one of which is also better at fermenting lactic acid at low pH.

The mixed results obtained during our target search forced us to re-frame the question of what constitutes a useful library for phenotypic alteration. We hypothesized that the phenotypic diversity of a library could be quantified and used to evaluate the potential of different populations for strain improvement. After developing the conceptual framework to support it, we proposed a metric to estimate phenotypic diversity and showed that it correlates with the usefulness of a library with regard to finding an improved mutant. The metric, termed divergence, can be used to assess the potential of different targets, to prioritize and economize screening experiments, and, as we later proved, to optimize the construction of libraries.

The usefulness of evolutionary methods is often times muddled by the element of chance, and more so because failing to isolate an improved mutant does not suggest a modification to the experimental approach. With this in mind, we tested whether the divergence metric could be used to systematize the construction of new libraries when screening or selection of a previous library fails to deliver mutants improved for a trait of interest. We used the metric for successively modifying the alpha subunit library design until a mutant of interest was isolated. We showed that this effort increases the likelihood of finding desired clones, in our case, a butyrate-tolerant mutant that grows significantly faster in the presence of the toxic chemical compared to the wild-type. An optimized library, in which surface amino acids of the C-terminal domain of alpha were targeted for mutagenesis, was constructed by gathering the information about how modifications to the library design affected the resulting divergence. We repeated the approach with the sigma D libraries, and considerably enhanced the diversity by targeting regions 4.1 and 4.2 of this protein for mutagenesis. We used the novel sigma factor libraries to improve tolerance to the simultaneous stresses of overlimed bagasse hydrolysate and high concentrations of ethanol.

Lastly, we explored the use of our divergence metric to study key determinants of regulatory proteins (residues, regions, structures, or functionalities) that have a high potential for altering phenotype. We modified our divergence quantification protocol to test whether individual amino acids in the alpha subunit could be experimentally considered as determinants for diversity. We showed that not only can single residues be probed individually, but also that, by testing the phenotypic diversity produced by saturation mutagenesis at different positions, we could find regions and functionalities

that are promising for further studies. We proposed this as a novel way for reducing the search space in a particular target for the purpose of increasing the quality of a library.

What started as an effort to improve upon transcriptional engineering, soon evolved into a general approach to optimize random search-based methods for isolating traits of interest. We demonstrated the use of this approach for guiding the construction of transcriptional engineering libraries, and in addition outlined the conceptual framework for extending this work to any genetic library. As such, the work of this thesis served both theoretical and practical goals, and furthered the understanding of how evolution can be exploited in the laboratory.

Thesis Supervisor:

Gregory Stephanopoulos, W.H. Dow Professor of Chemical Engineering and

Biotechnology

Acknowledgements

I thank my mother for teaching me the joy of searching and asking questions, and for applauding my nonconformity and rational thinking even at the peril of her patience. I thank my sister, Ilana, and my brother, Samuel, for their unconditional support and for always showing interest out of pure solidarity. I extend my gratitude to my uncles, Salomon and Isaac, for unknowingly encouraging me to work my hardest. To my friends – and you know well who you are – for sharing laughs and cries during these years, for listening to my endless complaints, and for insisting that there is life outside the confines of MIT.

To Prof. Greg Stephanopoulos for his support and for his vision. To Prof. Kris Prather for investing her time in my education much beyond her responsibility. To Prof. Daniel Wang for being strict as a teacher, and friendly as a counselor. To Prof. Dane Wittrup for his sharp and helpful insights. To Prof. Jeff Tester, for treating me as one of his own students and for his unwavering desire to help. To Prof. George Georgiou, whose role as a mentor continues after years of leaving his lab. To many Professors and teachers throughout my life who motivated me to think, even when it was the unpopular thing to do.

To Hal Alper, for being a good friend and a good teacher (and good company during those long days discussing "the webs"). To Curt Fischer, for showing me that getting at the bottom of things is both valuable and fun, and, along Andy Peterson and Tracy Matthews, for making me part of their interest for renewable energy. To Ajikumar Parayil, for responding with a smile when I lost my temper. To Christine Santos, for

reminding me that one can be productive and still remain calm, and for sharing her passion for tyrosine. To Tanguy Chau and Lily Tong, for reminding me that the lab can be a social activity. To Huimin Yu, for staying late with me and for including me in her projects. To Christie Peebles and Franz Hartner, for helping me make beer out of biomass. To all other members of the Stephanopoulos lab who, though did not necessarily contribute to this work directly, were always there for procrastination purposes and to offer advice (and for answering questions on a daily basis): Jose Aleman, Keith Tyo, Ben Wang, Hang Zhou, Jason Walther, Jamey Young, Mark Styczynski, Kyle Jensen, Mitchell Tai, Tina Lutke-Eversloh, Adel Ghaderi, and Vikram Yadav. I would have liked to thank them individually, for all taught me different lessons, but that would necessitate a Chapter of its own.

To the administrative help without which it would have been impossible to navigate through the labyrinth that is our Department: Rosangela Dos Santos, Suzanne Easterly, Mary Wesolowski, and Iris Chang. Finally, to all those overachieving, self-motivated, energetic MIT students and alumni, who have inspired me with their achievements to live up to their accomplishments.

Contents

ABSTRACT **3**

ACKNOWLEDGEMENTS **6**

LIST OF FIGURES **17**

LIST OF TABLES **20**

1. INTRODUCTION **21**

1.1 MOTIVATION 21

1.2 OBJECTIVES AND APPROACH 23

1.2.1 *Aim 1: Find New Targets for Transcriptional Engineering* *23*

1.2.2 *Aim 2: Develop a Quantitative Metric for Evaluation of Transcriptional Engineering and other Evolutionary Approaches for Strain Improvement* *24*

1.2.3 *Aim 3: Optimize Transcriptional Engineering Libraries with the Use of the Divergence Metric* *25*

1.2.4 *Aim 4: Find and Study Key Interactions of Transcriptional Engineering Targets for Phenotypic Alteration* *25*

1.3 THESIS ORGANIZATION 26

2. PHENOTYPIC IMPROVEMENT AS A GOAL FOR METABOLIC ENGINEERING **28**

2.1 MOTIVATION FOR PHENOTYPIC IMPROVEMENT 29

2.2 RATIONAL APPROACHES 31

2.2.1 *Host Selection* *31*

2.2.2 *Genetic Modification* *33*

2.2.3 *Flux determination* *36*

2.3 EXAMPLES 37

2.4 LIMITATIONS OF RATIONAL APPROACHES 40

3. RANDOM SEARCHES FOR PHENOTYPIC IMPROVEMENT 44

3.1 EVOLUTIONARY PRINCIPLES CAN BE EXPLOITED FOR PHENOTYPIC MODIFICATION 45

3.1.1 *Generation of Diversity* *47*

3.1.1.1 Whole-cell Mutagenesis 48

3.1.1.2 Region-wide Mutagenesis 50

3.1.1.3 Position-specific Mutagenesis 51

3.1.1.4 Other Strategies for Diversity Generation 52

3.1.2 *Purifying Selection* *53*

3.1.2.1 Selection Based on Growth 53

3.1.2.2 Screening 55

3.2 RANDOM SEARCH STRATEGIES AND TOOLS FOR PHENOTYPIC IMPROVEMENT... 57

3.2.1 *Challenges of Random Search Strategies* *57*

3.2.2 *Example Library Designs* *59*

3.3 TRANSCRIPTIONAL ENGINEERING 62

3.3.1 *Background* *63*

3.3.2 *Artificial Transcription Factor Libraries for Alteration of Complex Phenotypes* *64*

3.3.3 *Native Transcription Factors and Global Transcription Machinery Engineering (gTME)* *66*

3.4 THE LIMITED NATURE OF STRAIN IMPROVEMENT 69

4. TARGETS FOR TRANSCRIPTIONAL ENGINEERING IN BACTERIA 71

4.1 TRANSCRIPTION IN BACTERIA 72

4.1.1 *The RNA Polymerase: Structure and Function* *73*

4.1.2 *Transcription Regulation* *76*

4.2 PRINCIPAL SIGMA FACTOR (SIGMA D) 78

4.2.1 *Use of sigma D for Phenotypic Improvement in E. coli* *80*

4.2.1.1 Hyaluronic Acid Production 80

4.2.2 *Use of sigma D for Phenotypic Improvement in L. plantarum* *84*

4.2.2.1 Library Construction 88

4.2.2.2 Mutants with Improved Tolerance to Lactic Acid and Low pH 89

4.2.2.3 Sequence Analysis 92

4.2.2.4 Fermentations 94

4.3 ALTERNATIVE SIGMA FACTORS 96

4.3.1 *Use of sigma S for Phenotypic Improvement in E. coli* *100*

4.3.1.1 Hyaluronic Acid Production 100

4.3.1.2 Carbon Dioxide Tolerance 102

4.3.2 *Use of sigma E and sigma H for Phenotypic Improvement in E. coli* *103*

4.3.2.1 Heat and Ethanol Tolerance 104

4.4 ALPHA SUBUNIT OF THE RNA POLYMERASE 108

4.4.1 *Use of the alpha Subunit for Phenotypic Improvement in E. coli* *111*

4.4.1.1 Tolerance to n-butanol 111

4.4.1.2 Tolerance to Other Solvents 115

4.4.1.3 Tolerance to Ion Leakage of Butanol-Tolerant Mutant 117

4.4.1.4 Enhanced Production of L-Tyrosine 118

4.4.1.5 Enhanced Production of Hyaluronic Acid 122

4.4.1.6 Location of the *rpoA* Mutations 123

4.4.2 *Phenotypic Specificity of the alpha Subunit Mutants and Comparison to Alternative Approaches for Strain Improvement* *125*

5. PHENOTYPIC DIVERSITY AS A METRIC FOR EVALUATION OF RANDOM STRAIN IMPROVEMENT LIBRARIES 128

5.1 MOTIVATION FOR DEVELOPING A METRIC 129

5.1.1 *Comparing Different Experimental Approaches for Building Libraries* *130*

5.1.2 *Optimizing the Construction of Libraries* *132*

5.1.3 *Infiniteness of Sequence Space* *133*

5.2 THE CHOICE OF PHENOTYPIC DIVERSITY AS A METRIC 134

5.3 DIVERGENCE: A NORMALIZED PHENOTYPIC DIVERSITY METRIC 135

5.4 QUANTIFICATION OF DIVERGENCE USING GROWTH IN SOLID MEDIA 139

5.4.1 *Validation of Growth as a Phenotype for Diversity Quantification* *139*

5.4.2 *Determination of Divergence Using Growth* *142*

5.4.3 *Comparison between Libraries Constructed via Transcriptional Engineering and Classical Strain Improvement* *145*

5.5 CORRELATION OF THE DIVERGENCE METRIC WITH THE PROBABILITY OF FINDING AN IMPROVED MUTANT 147

5.5.1 *Implications* *148*

5.6 OTHER COMPLEX PHENOTYPES FOR QUANTIFICATION OF DIVERSITY 152

5.6.1 *Validation of Intracellular pH for Phenotypic Diversity Quantification* *154*

5.7 THE RELATIVE NATURE OF THE DIVERGENCE METRIC 154

6. OPTIMIZATION OF TRANSCRIPTIONAL ENGINEERING LIBRARIES USING DIVERGENCE **157**

6.1 OPTIMIZATION AS A REDUCTION OF THE SEARCH SPACE 158

6.2 OPTIMIZATION OF ALPHA SUBUNIT LIBRARIES FOR IMPROVING BUTYRATE TOLERANCE 160

6.2.1 *Optimization of Random Search-Based Strain Improvement Programs* *162*

6.2.2 *Utilization of Previously Constructed Libraries* *163*

6.2.3 *Improvement of Libraries by Reducing the Search Space to the C-terminal Domain* *166*

6.2.4 *Isolation and study of Butyrate-Tolerant Mutants* *169*

6.2.4.1 Sequence Analysis 170

6.2.4.2 Promoter Replacement 171

6.3 RELATIONSHIP BETWEEN DIVERGENCE AND POSTERIOR PROBABILITY OF FINDING IMPROVED MUTANTS 172

6.4 EXPLOITING INFORMATION DERIVED FROM DIVERGENCE FOR CONSTRUCTING AN OPTIMIZED ALPHA SUBUNIT LIBRARY 174

6.4.1 *Design and Construction of the αCTD*t Library* *174*

6.4.2 *Effect of Promoter Strength, Mutagenesis Rate, and Library Size in Divergence* *178*

6.4.3 *Use of aCTD*t Library for Improving Ethanol Productivity in Ethanologenic KO11 E. coli strain* *181*

6.4.3.1 Cellulosic Ethanol as a Biofuel 181

6.4.3.2 A Traditional Metabolic Engineering Approach for Developing an Ethanologenic *E. coli* Strain 183

6.4.3.3 The Challenge of Fermenting Biomass Hydrolysates 184

6.4.3.4 Selection and Isolation of Improved Mutants 187

6.5 DIVERGENCE IN A SINGLE-LOCUS PHENOTYPE 190

6.6 OPTIMIZATION OF SIGMA D LIBRARIES 192

6.6.1 *Targeted Mutations in Promoter-Contacting Residues* *193*

6.6.2 *Targeted Mutagenesis of Promoter-Binding Regions* *197*

6.6.3 *Use of sigma D Optimized Libraries for Improving Survivability in High-Ethanol Overlimed Hydrolysate* *201*

7. THE DIVERGENCE METRIC AS A TOOL FOR STUDYING INTERACTIONS THAT ALTER PHENOTYPE 207

7.1 DETERMINANTS OF DIVERSITY 208

7.2 SINGLE RESIDUES AS DETERMINANTS OF DIVERSITY 209

7.3 AMINO ACIDS IN THE ALPHA SUBUNIT C-TERMINAL DOMAIN (αCTD) AS A TEST CASE 211

7.3.1 *The Difference between Residues with Low and High Regulatory Flexibility 211*

7.3.2 *A More Complete Account of Function-Diversity Relationships* *212*

7.4 KEY INTERACTIONS BETWEEN THE RNA POLYMERASE AND THE PROMOTER FOR PHENOTYPIC ALTERATION 215

7.4.1 *Residue D259 and the 261 Determinant of the αCTD* *216*

7.4.2 *The -35-binding region of sigma D* *218*

8. RECOMMENDATIONS AND CONCLUSIONS 220

8.1 SUMMARY 220

8.2 CONCLUSIONS 222

8.3 RECOMMENDATIONS AND FUTURE WORK 226

9. MATERIALS AND METHODS 230

9.1 REAGENTS AND ENZYMES 230

9.2 METHODS FOR *L. PLANTARUM* 231

9.2.1 *Bacterial Strains, Plasmids and Growth Conditions* *231*

9.2.2 *DNA Extraction and Purification* *233*

9.2.3 *Transformation by Electroporation* *234*

9.2.4 *Library Construction and Phenotype Selection* *235*

9.2.5 *Measurement of Colony Size of* L. plantarum *for Diversity Quantification* *235*

9.2.6 *Fermentations* *236*

9.2.7 *Primers for Amplification and Cloning* *236*

9.2.8 *Primers for Sequencing* *237*

9.3 METHODS FOR *E. COLI* 238

9.3.1 *Construction of sigma D and sigma S Libraries* *238*

9.3.1.1 Error-prone PCR Libraries 238

9.3.1.2 Targeted and Position-Specific Libraries of sigma D 239

9.3.2 *Construction of sigma E-sigma H Libraries* *240*

9.3.3 *Construction of the alpha Subunit Libraries* *240*

9.3.3.1 Error-prone PCR Libraries 240

9.3.3.2 Saturation Mutagenesis Libraries of the L-tyrosine Producing Mutant (rpoA14) 241

9.3.3.3 Targeted and Position-Specific Libraries of the αCTD 241

9.3.3.4 Saturation Mutagenesis of Individual Amino Acids in αCTD 242

9.3.4 *Hyaluronic Acid Methods* *243*

9.3.4.1 Host Strain 243

9.3.4.2 High-Throughput Screen and Quantification with Alcian Blue 244

9.3.4.3 HA Quantification by HPLC 244

9.3.5 *L-Tyrosine Methods* *245*

9.3.5.1 Host Strain and Screening 245

9.3.5.2 Quantification of L-Tyrosine 246

9.3.6 *Selection Experiments* *246*

9.3.6.1 Carbon Dioxide Tolerance 246

9.3.6.2 Heat and Ethanol Tolerance 247

9.3.6.3 Butanol and Solvent Tolerance 248

9.3.6.4 Butyrate Tolerance 248

9.3.6.5 Ethanol and Hydrolysate Tolerance of KO11 Derivative 249

9.3.7 *Intracellular pH Determination for Measuring Ion Leakage of Butanol-Tolerant Mutant* *250*

9.3.8 *Intracellular pH Determination for Phenotypic Diversity Quantification*. *250*

9.3.9 *Phenotypic Diversity Quantification at a Single Locus using GFP* *252*

9.3.10 *Primers for Amplification and Cloning* *252*

9.3.10.1 Sigma D and Sigma S 252

9.3.10.2 Sigma E and Sigma H 253

9.3.10.3 Alpha Subunit 253

9.3.11 *Sequencing Primers* *255*

9.3.11.1 Sigma D 255

9.3.11.2 Sigma S 255

9.3.11.3 Sigma E 256

9.3.11.4 Sigma H 256

9.3.11.5 Alpha Subunit 256

10. REFERENCES 258

List of Figures

FIGURE 3.1-1. EVOLUTIONARY APPROACHES FOR STRAIN IMPROVEMENT 46

FIGURE 4.1-1. TYPES OF INTERACTIONS OF THE RNAP HOLOENZYME AT THE PROMOTER 75

FIGURE 4.2-1. SCREENING OF RECOMBINANT SIGMA D AND SIGMA S LIBRARIES FOR HYALURONIC ACID ACCUMULATION 83

FIGURE 4.2-2. TEST FOR ORGANIC AND INORGANIC ACID TOXICITY IN *LACTOBACILLUS* 87

FIGURE 4.2-3. GROWTH PROFILE OF RETRANSFORMED STRAINS IN DIFFERENT STRESSES 90

FIGURE 4.2-4. SEQUENCE ALIGNMENT OF E. COLI AND L. PLANTARUM SIGMA D 93

FIGURE 4.2-5. FERMENTATION EXPERIMENTS 95

FIGURE 4.3-1. SURVIVABILITY ASSAY FOR SIGMA E -- SIGMA H MUTANTS 107

FIGURE 4.4-1. GROWTH ASSAY IN BUTANOL 113

FIGURE 4.4-2. TEST OF TOXICITY OF VARIOUS ALCOHOLS 115

FIGURE 4.4-3. GROWTH ASSAY FOR MUTANT AND WILD-TYPE IN VARIOUS ALCOHOL SOLVENTS 116

FIGURE 4.4-4. ION LEAKAGE ASSAY FOR MUTANT AND WILD-TYPE 118

FIGURE 4.4-5. L-TYROSINE PRODUCTION OF PARENTAL STRAIN AND MUTANTS 120

FIGURE 4.4-6. ALCIAN BLUE QUANTIFICATION OF HA PRODUCTION BY SELECTED RPOA MUTANTS 123

FIGURE 4.4-7. SCHEMATIC MAPPING OF THE MUTATIONS ON THE RPOA PROTEIN 124

FIGURE 5.1-1. EXAMPLE METABOLIC MAP 132

FIGURE 5.3-1. SCHEMATIC ILLUSTRATION OF THE DIVERGENCE NORMALIZATION METHOD IN 2-D 138

FIGURE 5.4-1. VALIDATION EXPERIMENTS FOR GROWTH AS A PHENOTYPE FOR DIVERSITY QUANTIFICATION 140

FIGURE 5.4-2. EXPERIMENTAL PROCEDURE FOR PHENOTYPIC DIVERSITY QUANTIFICATION 143

FIGURE 5.4-3. DISTRIBUTIONS OF THE AVERAGE PHENOTYPIC DISTANCES 145

FIGURE 5.4-4. DIVERGENCE OF THE LACTOBACILLUS LIBRARIES 146

FIGURE 5.5-1. ESTIMATE OF THE PROBABILITY OF FINDING NOVEL PHENOTYPES 148

FIGURE 6.2-1. TEST OF TOXICITY OF BUTYRATE 161

FIGURE 6.2-2. BACTERICIDAL EFFECTS OF BUTYRATE 162

FIGURE 6.2-3. SELECTION EXPERIMENTS IN BUTYRATE 165

FIGURE 6.2-4. DIVERGENCE OF RPOA LIBRARIES DURING OPTIMIZATION 168

FIGURE 6.2-5. GROWTH IN BUTYRATE OF WILD-TYPE AND MUTANTS 170

FIGURE 6.4-1. DESIGN OF THE ACTD*T LIBRARY 176

FIGURE 6.4-2. DISTRIBUTION OF AMINO ACID CHANGES IN DIFFERENT ACTD*T LIBRARY DESIGNS 178

FIGURE 6.4-3. SIMPLE VARIATIONS TO A LIBRARY DESIGN 181

FIGURE 6.4-4. ROUTES FOR ISOLATING AN ETHANOL- AND HYDROLYSATE-TOLERANT MUTANT 188

FIGURE 6.4-5. FED-BATCH HYDROLYSATE FERMENTATION OF THREE IMPROVED MUTANTS AND CONTROL 189

FIGURE 6.5-1. DIVERGENCE IN A SIMPLE PHENOTYPE 191

FIGURE 6.6-1. SEQUENCE ALIGNMENT OF DIFFERENT SIGMA FACTORS IN E. COLI 194

FIGURE 6.6-2. INITIAL SIZE OF -10-BINDING REGION LIBRARIES 196

FIGURE 6.6-3. DESIGN OF THE LIBRARIES TARGETED TO THE 2.4 AND 4.2 REGIONS OF SIGMA D 198

FIGURE 6.6-4. DIVERGENCE OF THE LIBRARIES TARGETED TO THE 2.4 AND 4.2 REGIONS OF SIGMA D 199

FIGURE 6.6-5. LIBRARY SELECTION IN ETHANOL AND OVERLIMED HYDROLYSATE 202

FIGURE 6.6-6. QUALITATIVE ILLUSTRATION OF LIBRARY SELECTION IN ETHANOL AND OVERLIMED HYDROLYSATE 203

FIGURE 6.6-7. IMPROVED MUTANT FROM OUR SELECTION IN ETHANOL AND OVERLIMED HYDROLYSATE 204

FIGURE 6.6-8. SEQUENCE ANALYSIS OF ETHANOL AND HYDROLYSATE TOLERANT MUTANT 205

FIGURE 7.3-1. DIVERGENCE OF INDIVIDUAL AMINO ACIDS IN THE αCTD 213

FIGURE 7.3-2. DIVERGENCE OF THE αCTD AVERAGED BY DETERMINANT 213

FIGURE 7.3-3. LOCATION OF THE DETERMINANTS MUTATED IN αCTD 215

FIGURE 9.2-1. SCHEMATIC ILLUSTRATION OF PLASMID PDK12D 233

List of Tables

TABLE 1. GROWTH CHARACTERISTICS FOR LACTOBACILLUS MUTANTS IN DIFFERENT STRESSES 92

TABLE 2. *E. COLI* SIGMA FACTORS 97

TABLE 3. HYALURONIC ACID PRODUCTION CHARACTERISTICS OF ISOLATED MUTANTS ... 101

TABLE 4. PHENOTYPIC SPECIFICITY OF *RPOA* MUTANTS 127

TABLE 5. POSTERIOR PROBABILITY CALCULATION FOR FINDING THE IMPROVED MUTANT (S299T) IN DIFFERENT LIBRARIES 174

TABLE 6. SURVIVABILITY OF DIFFERENT MUTANTS IN 40 G/L ETHANOL AT 48 HR 188

Chapter 1

1. Introduction

1.1 Motivation

In the most fundamental of ways, humankind owes its subsistence to the exploitation of natural resources. Most notable are crops and animals, and a close look to our relationship with the species that we have strived to improve for millennia evidences the influence crops and animals have had in our history. Exploiting nature through domestication is the clearest example and in some sense the definition of phenotypic improvement. In this context, metabolic engineering emerges from our relationship with the species that surround us as an enabling science that aims at systematizing efforts to manipulate traits of interest. The basis of the science lies in the mapping between genotype and metabolism –defined as the network of biochemical reactions that sustain life – and between metabolism and phenotype.

One main application of the principles of metabolic engineering has been the improvement of industrially-relevant microorganisms. In fact, the revolution heralded by the advent of recombinant DNA technology was almost immediately applied to improving traits in single-celled organisms, such as bacteria and yeast. Simple genetic modifications, such as gene overexpression or deletion, have been exploited for metabolic engineering when the mapping between the genotype and phenotype is straightforward. For all their simplicity, directed genetic modifications are the basis of phenomenal achievements in the production of fermentation-based chemicals.

In spite of the successes of directed genetic changes for phenotypic improvement, not all the traits of industrial interest can be tackled with these techniques, because some phenotypes are complex and poorly understood. Following the commonly-used recipe of imitating nature, scientists have tried to accelerate and direct the process of evolution to confront these sorts of challenges. Evolutionary methods consist of creating libraries of genotypic variants and searching in them an individual that is improved with respect to the traits of interest; therefore, they are said to be based on random searches. The classical evolutionary approach for strain improvement, whole-cell mutagenesis and selection, predates the advances and tools of molecular biology and continues to be useful today.

As the challenges of industrial biotechnology become more ambitious and the achievements of evolutionary approaches become more widespread, so grows the list of experimental methods for random search-based phenotypic alteration. One such technique was being developed with great promise in our laboratory not long before the work of this thesis began, based on mutagenesis of components of the native transcriptional machinery (Alper & Stephanopoulos, 2007; Alper *et al.*, 2006). The approach stemmed from earlier findings that artificial transcription factors could be used for strain improvement (Lee *et al.*, 2004; Park *et al.*, 2005a; Park *et al.*, 2003; Park *et al.*, 2005b), and the goal became to improve upon these tools, collectively called transcriptional engineering.

How to improve upon transcriptional engineering quickly raises the question of what it means to improve upon an evolutionary approach in general: it is not possible to improve upon something if one does not know how good it is to begin with. Therefore, the majority of the present thesis aims at developing tools to evaluate and compare evolutionary approaches for strain improvements, and it applies them in particular to the optimization of transcriptional engineering.

1.2 Objectives and Approach

The overarching goal of improving upon transcriptional engineering was progressively decomposed in different specific objectives.

1.2.1 *Aim 1: Find New Targets for Transcriptional Engineering*

The transcriptional engineering approach is based on global perturbation of the transcriptome in order to alter many cellular functions simultaneously, but cellular physiology is regulated at several levels and by different nodes of the network (Ma *et al.*, 2004; Martinez-Antonio *et al.*, 2008). With this in mind, we explored the use of targets for phenotypic alteration other than the principal sigma factor of *E. coli*, which was the focus of previous studies in our laboratory (Alper & Stephanopoulos, 2007).

As a first step, we constructed mutagenesis libraries of sigma factors other than the housekeeping factor (sigma D, coded by *rpoD*), in particular, we focused on the stress-associated factors sigmas S, E, and H, and later on the alpha subunit of the RNA polymerase. We also extended the use of *rpoD* libraries in a different bacterial species, *L. plantarum*. Each of the libraries was tried for the improvement of traits of interest in order to test whether these targets held promise for global transcriptomic modification.

1.2.2 *Aim 2: Develop a Quantitative Metric for Evaluation of Transcriptional Engineering and other Evolutionary Approaches for Strain Improvement*

While examining targets for transcriptional engineering, we realized that not all targets are useful for improving phenotype. To complicate matters, we observed that some targets were successful for some phenotypes but not others. Therefore, we began to develop the conceptual grounds for evaluating randomized libraries, which allowed us to define a statistical metric based on the quantification of the phenotypic diversity of different populations.

We then demonstrated the use of this metric (termed divergence) for comparing different populations in *L. plantarum*, either based on error-prone PCR libraries of the principal sigma factor or on whole-cell chemical mutagenesis. We tested the ability of this metric to inform us on the probability that a new phenotype will be found in a population of cells.

1.2.3 *Aim 3: Optimize Transcriptional Engineering Libraries with the Use of the Divergence Metric*

At that point of the thesis, we had relied on undirected changes in the sequence of genes of interest (accomplished by error-prone PCR, epPCR, of the entire coding region), which theoretically can deliver a large proportion of variants of the wild-type sequence with potentially useful phenotypic characteristics. Because epPCR hinges on creating combinatorial arrangements of many nucleotides, the number of variants that can be constructed is virtually infinite.

We hypothesized that optimization of libraries for strain improvement aimed at reducing the search space could increase the probability of finding a desired mutant. The tradeoff of such a reduction is that potentially useful mutations are forgone by delimiting the nucleotide regions that can be changed, and we believed that our divergence metric could aid in estimating the impact of this tradeoff. Guiding the design of libraries by sequential probing the search space is an especially valuable time- and resource-saving effort when an improved mutant is not readily isolated. We chose one such case to test our optimization algorithm.

1.2.4 *Aim 4: Find and Study Key Interactions of Transcriptional Engineering Targets for Phenotypic Alteration*

The reduction of the search space that resulted from optimization suggested that some regions of a target are more important than others for their capacity to alter phenotype. With this in mind, we extended the concept of phenotypic diversity and used it to search for and to examine regions in our transcriptional engineering targets with most potential for delivering traits of interest. We applied the method for quantification of divergence to individual amino acids in the alpha subunit of the RNA polymerase, and we found that some residues are more promising than others as targets for mutagenesis. Furthermore, we tried to understand the potential of different amino acids and regions in the context of their functions in transcriptional regulation.

1.3 Thesis Organization

The overall organization of this thesis follows closely the specific aims outlined above. Chapter 2 establishes the motivation for phenotypic improvement and describes the traditional approaches in metabolic engineering. It discusses the successes and limitations of these approaches, and emphasizes the need for random search-based evolutionary methods for strain improvement. Chapter 3 expands on the description and underlying principles of evolutionary methods, and provides the background for the general idea of transcriptional engineering.

This preamble leads to Chapter 4, which deals with the use of previously-found and novel targets for transcriptional engineering and draws on the results of specific aim #1. It reports on the use of these targets for improving several industrially-relevant phenotypes. Chapter 5 sets the conceptual grounds for the development of the divergence metric and its use to quantify the potential of strain improvement libraries. This part also illustrates the correlation between the newly-developed metric and its use to compare the potential of different libraries, corresponding to specific aim #2. These two chapters set the stage for the optimization efforts that constitute a central theme of the thesis.

Chapter 6 describes the experimental approaches for optimizing transcriptional engineering libraries of the alpha and sigma D subunits of the RNA polymerase. It reports on the improvement of additional phenotypes of practical significance, and it epitomizes the results of specific aim #3. Chapter 7 concerns with the last of the aims, and exemplifies the use of the divergence metric to study key interactions that hold potential for building better libraries. It also gives the conceptual basis for using the divergence metric to guide the construction of libraries by reducing the search space.

Finally, Chapter 8 concludes with a summary and a list of the most important findings, and puts forth recommendations for future work. The experimental methods are described in Chapter 9, and references are given in Chapter 10.

Chapter 2

2. Phenotypic Improvement as a Goal for Metabolic Engineering

Traditionally, metabolic engineering has combined genetic modification, flux determination, and phenotype evaluation to rationally improve traits of interest (Raab *et al.*, 2005; Stephanopoulos, 1999). More recently, the toolset for phenotypic modification has been enriched by the use of evolutionary and combinatorial approaches (Santos & Stephanopoulos, 2008a). In this Chapter, we discuss the motivation for modifying phenotypes and illustrate the traditional approaches for achieving this goal. The principles and the many tools that have been the bedrock of metabolic engineering are described in sufficient detail so as to emphasize the significance of phenotypic improvement; we then turn to the limitations of these traditional methods, which have encouraged the use and development of the evolutionary approaches described in the next Chapter.

2.1 Motivation for Phenotypic Improvement

The mere observation that wild species have properties that are measurably different from those that have been in contact with humans from immemorial times is a testament to our incessant drive to manipulate nature. Regardless of how rich and diverse it may seem, our environment does not readily provide resources optimally suited for human use and, in essence, this is the motivation for phenotypic improvement. Both the cause and the cure for this discrepancy are contained in the concept of evolution. Selective breeding has been the traditional tool for guiding evolution of a few species away from their untamed counterparts in order to preserve or enhance the properties that serve us.

Because microorganisms are invisible to the naked eye, they remained undiscovered for most of human history, and thus they have not been a principal target for selective breeding. Nonetheless, several species of microorganisms have evolved closely with humans, and some degree of phenotypic modification has resulted from this rather passive form of domestication. Staples of Western civilization such as wine, bread, and cheese, are possible because of the fermentative metabolism of microorganisms (many examples in other civilizations exist, but most are unfamiliar to the author). The microscope allowed formalizing the interest in a previously hidden world into what today is the field of microbiology. A better understanding of microbial physiology then enabled a more scientific approach to selective breeding of microorganisms.

The industrial revolution brought about the systematization and scale-up of production practices, and this school of thought eventually infiltrated into the food manufacturing and processing business. Fermentation processes were not left behind, and they soon became the target for extensive optimization. New products, from solvents

(Woods, 1995) to antibiotics (Kumazawa & Yagisawa, 2002), were soon being produced at unprecedented scales. The microbial strains responsible for these products, even those that had been traditionally used for the small-scale version of the processes, had to perform in the new conditions. Industrial microbiologists could not wait for evolution to usher the necessary adaptive changes, and thus the development of methodological approaches for phenotypic improvement became a key goal of overall process optimization. With the advent of recombinant DNA technology and the molecular biology revolution, more precise genetic manipulations for trait modification were possible, which is the basic premise behind what is now the field of metabolic engineering.

Three parameters are of main interest to the metabolic engineer in the context of industrial-scale bioprocesses: yield, titer, and productivity. The first refers to the amount of product synthesized per unit of substrate; the second refers to the concentration of the product of interest in the fermentation broth; and the third to the rate at which the product of interest is synthesized (Marten *et al.*, 2002; Shuler & Kargi, 2002; Stephanopoulos *et al.*, 1998). All three are tightly linked to the economic profitability of a production process, and thus, all are relevant in the context of phenotypic improvement.

The strict requirement of economic viability has for a long time biased the use of biochemical processes for the production of value-added compounds. Although we have only been able to tap into a small part of nature's "warehouse", several products have been exploited as pharmaceuticals, agrochemicals, pigments, and raw materials for polymer synthesis (Demain & Fang, 2000; Kayser & Quax, 2007). Bioactive natural products are often intriguingly complicated, making chemical synthesis economically infeasible (Zhong & Yue, 2005) and ensuring that the biochemical or microbiological routes remain competitive.

Increasingly, the general appeal of sustainability has directed attention to commodity chemicals derived from renewable resources. These products are characterized by small profit margins and large scales, and thus present a different set of pressures and constraints on the process designers and engineers. In particular, yields must be high because the lignocellulosic feedstock is usually a major cost driver; titers and productivities are inherently limited because both the substrate and product mixes tend to be toxic to the fermenting microorganism (Lynd *et al.*, 1999; Lynd *et al.*, 2005). Much has been accomplished in recent years through traditional and new approaches in metabolic engineering. The following section examines traditional approaches for phenotypic improvement in greater detail.

2.2 Rational Approaches

2.2.1 *Host Selection*

The first and most obvious requirement for initiating a phenotypic improvement program is selecting a strain to modify. Finding an optimal host for biochemical production is the foundation of any metabolic engineering effort, but is far from trivial. In the context of commodity chemical manufacturing, the ideal host would degrade lignocellulosic components, ferment the resulting sugars (both hexoses and pentoses) with high yield, and tolerate high titers of the end-product and other toxins at high temperatures (to avoid cooling costs). The remarkably high biodiversity, especially for microorganisms, may suggest that an ideal host for production in these conditions already exists, so that the real challenge is finding it. Since there is no clear path for solving this needle-in-a-haystack problem, most researchers have opted to engineer optimal hosts by combining desirable characteristics into a host using recombinant DNA technology. Such efforts usually entail alteration of phenotypes dictated by multiple genes. The systematic implementation of this approach requires substantial knowledge of the host to be modified, which has favored the use of "laboratory strains" that are research-friendly, but not necessarily robust enough for industrial applications. The characteristics that make laboratory strains most appealing are also to a large extent those that will be needed for a long-term commitment to any production strain.

The first requisite is genetic competence, the ability of a strain to accept foreign DNA in a controllable fashion. High transformation efficiencies are desirable, especially for the successful use of combinatorial libraries for screening genes or gene variants that confer a particular phenotype of interest (see discussion in following Chapter). Even though transformation protocols have been used for a long time, they tend to be host-specific and based on empirical observations rather than on underlying principles. Furthermore, their success seems to depend on a variety of factors such as the activity of host restriction and DNA-modification systems (Alegre *et al.*, 2004; Matsushima *et al.*, 1989), genetic background (Umemoto *et al.*, 1996), origin of replication and marker of the vector (Aukrust *et al.*, 1995), to name a few.

A second trait that makes laboratory strains attractive is the availability of well-characterized metabolic engineering "modules" that allow manipulation of the genotype in different ways. Examples include promoters of various strengths (Alper *et al.*, 2005; Hammer *et al.*, 2006; Jensen & Hammer, 1998a), termination sequences, repressor-inducer systems, plasmids (with known copy number, replication mechanism, compatibility with other plasmids, etc.), chromosome integration cassettes for building knockouts or stable replication of genes, etc.

A third feature is the availability of "omics" platforms and algorithms for genome-wide characterization of cellular responses to different manipulations and environments (microarrays, metabolic network models, etc.). A fourth, and most understated feature of all, is the great amount of accumulated knowledge on the physiology of laboratory strains provided by generations of researchers. Because most of these studies were initiated with divergent goals, which at times converge in solving practical problems, similar circumstances are hard to replicate for strains that will be *ad hoc* designed for production of a particular chemical, even with significant monetary resources.

2.2.2 *Genetic Modification*

After a host with some or all of the above properties has been recognized and selected, metabolic engineering programs can begin, in accordance to overall process considerations. Since genotype-phenotype maps are complex and largely unknown, metabolic engineering relies heavily on rigorous phenotype evaluation (with or without flux determination) following each genetic modification. This constitutes the traditional path for metabolic engineering. Much of the effort in strain improvement lies in trying to explore the interconnectivity of this map, because a basic premise of metabolic engineering is that phenotype arises from the biochemical interactions between gene products and metabolites and not only from the gene products themselves. In addition, recombinant DNA technology opened the door for transferring genetic determinants from one organism into another in order to effect changes in metabolism, effectively expanding the native genotype-phenotype map of an organism and the resulting metabolic network.

There are various ways by which genetic-level modulations can be indtroduced into an organism. In the context of metabolic engineering, they can be categorized in three classes: (i) efforts that aim at altering the level of a gene product; (ii) efforts that aim at altering the interactions between the gene product and its targets; and (iii) efforts that aim at introducing heterologous gene products into the host. All of these genetic manipulations rely on a handful of molecular biology tools and protocols, some of which are explained in the Materials and Methods Chapter (if relevant to the present thesis), or are covered elsewhere (Sambrook & Russell, 2001b; Sambrook *et al.*, 2006).

Manipulations that belong to the first class are based on the idea that the concentration of gene products may impact phenotype by changing the relative significance of the nodes that constitute the metabolic network. For example, if the gene product is an enzyme, altering its concentration may have an effect in flux, depending on its control coefficient at the base level (Stephanopoulos *et al.*, 1998), making the reaction more or less important relative to others. If the gene product is a regulator, it may affect phenotype by forcing a response that does not occur at the base level (or by eliminating the response altogether). An increase or decrease in concentration, or complete elimination of gene products can be accomplished in one of several ways. Replacing the native promoter with a weaker or stronger one has been used to alter the flux through pathways of interest and to balance intermediate pools (Alper *et al.*, 2005; Hammer *et al.*, 2006; Jensen & Hammer, 1998a; Klein-Marcuschamer *et al.*, 2007). Alteration of copy number through the use and engineering of extrachromosomal DNA vectors has also been used (Jones *et al.*, 2000; Tao *et al.*, 2005). Knockouts or gene deletions have been a great tool for elimination of reactions that compete with the formation of products of interest and regulatory responses that elicit unwanted phenotypes (Cirz *et al.*, 2007; Green *et al.*, 1996; Shams Yazdani & Gonzalez, 2008; Sillers *et al.*, 2008).

Manipulations that belong to the second category are based on the fact that interactions between enzymes and metabolites or between regulators and their targets (DNA, RNA, proteins, etc.) determine the flow of information between the nodes of the metabolic network. For example, rendering rate-limiting or controlling enzymes resistant to feedback inhibition has been used to unlatch the flux through metabolic pathways for the production of valuable metabolites (Lutke-Eversloh & Stephanopoulos, 2005; Lutke-Eversloh & Stephanopoulos, 2007; Malumbres & Martin, 1996; Sahm *et al.*, 1996). Although the initial isolation of these enzymes has been commonly achieved using random search strategies (see next Chapter), only few simple and directed manipulations are needed to replicate the results in the same or closely-related strains (and, thus, they can be regarded as simple phenotypes). Another example of this type of manipulation is the introduction of engineered versions of native enzymes with favorable kinetics, or with altered substrate or product specificity (el Hawrani *et al.*, 1996; Lunzer *et al.*, 2005; Munir *et al.*, 1993; Yoshikuni *et al.*, 2006a; Yoshikuni *et al.*, 2006b). Yet another example is the alteration of the interaction between a regulator and its target, interesting instances being specialized ribosomes for the controlled translation of specific transcripts

(Brink *et al.*, 1995; Hui & de Boer, 1987) and the manipulation of gene noncoding regions in order to alter the signal transduction pathways that control to their expression (Wei *et al.*, 2008).

Manipulations that belong to the third category are based on the fact that, by introducing non-native gene products, one may add nodes to the metabolic network that did not exist in the wild-type host. This has allowed a wide range of applications, from the production of fuels in easy-to-manipulate hosts (Atsumi *et al.*, 2008a; Atsumi *et al.*, 2008b; Hanai *et al.*, 2007) to transferring phenotypes like heat-shock protection from one species to another (Liu *et al.*, 2007).

2.2.3 *Flux determination*

Because metabolism is defined by the set of biochemical reactions in the cell, and reactions are dynamic, the fluxes through the metabolic pathways are a main source of information in the process of translating genetic modifications into observable phenotypes. Reaction fluxes can be computed by measuring the *in vivo* distribution of isotopic tracers through the metabolic network, using one or several analytical chemistry tools, such as gas chromatography and mass spectroscopy (Antoniewicz *et al.*, 2007; Young *et al.*, 2008).

A key requirement for flux determination to be useful is that a model of the metabolic network must be available. Because of the nature of the phenotypes that were considered in the present thesis, these techniques were not used, so we omit the details for brevity. A more complete account of the applications of flux determination for gathering information about the metabolic network can be found elsewhere in the metabolic engineering literature.

2.3 Examples

Since its formal inception nearly two decades ago, metabolic engineering has achieved sterling success in the development of novel microbial strains for use in sustainable and cost-competitive bioprocesses. Some notable examples include the production of bulk chemicals such as citric acid, lactic acid, propanediol, ethanol and biopolymers such as poly(hydroxybutyrate) (PHB) and other poly(hydroxyalkanoates) (PHAs), as well as fine chemicals such as synthetic drug intermediates, lycopene and lysine (Klein-Marcuschamer *et al.*, 2007; Raab *et al.*, 2005). Let us describe some of these in greater detail to provide a more complete description of the traditional metabolic engineering framework.

1. Citric acid is a common flavor and acidifying additive of extensive use in the food industry, with a worldwide market in the order of millions of tones per year (Forster *et al.*, 2007a). A strain of the yeast *Yarrowia lipolytica* has been recently engineered for the production of citric acid from sucrose to compete with a less environmentally-friendly process that employs the fungus *Aspergillus niger*. The approach combines the introduction of a heterologous invertase enzyme from *Saccharomyces cerevisiae*, which allows the utilization of sucrose, with overexpression of the native isocitrate lyase, which minimizes the flux to the competing product isocitrate (Forster *et al.*, 2007a; Forster *et* *al.*, 2007b). This approach exemplifies the use of two of the three classes of genetic modifications described in Section 2.2.2.

2. Propanediol presents another instance that illustrates the success of metabolic engineering (we consider 1,2-propanediol in particular). This achievement, added to the fact that propanediol has been traditionally derived from petroleum, has fueled significant interest in metabolic engineering for advancing sustainable processes. Propanediol is employed as antifreeze and as a feedstock in the production of polyester resins, cosmetics, pharmaceuticals, household products, among others (Cameron *et al.*, 1998). Introduction in *E. coli* of an aldose reductase from rat was shown to divert methylglyoxal to 1,2-propanediol , which is otherwise not measurable in the fermentation broth (Cameron *et al.*, 1998). Subsequent optimization, which involved testing other enzymes with reductase activity and changing the fermentation parameters, yielded improvements in the recombinant process (Altaras & Cameron, 1999; Altaras & Cameron, 2000).

3. A final example is the production of polyhydroxyalkanoates (PHAs) in *E. coli.* These are polymers or co-polymers of hydroxyacyl units; the polyester formed by 3-hydroxybutyrate monomers, PHB, has received most attention for its potential use as a biodegradable plastic. PHB is a clear, brittle compound, synthesized from acetyl-coenzyme A (acetyl-CoA) in three steps. The reactions are catalyzed by the enzymes 3-ketothiolase, acetoacetyl-CoA reductase, and poly(3-hydroxybutyrate) synthase, or their homologues (Anderson & Dawes, 1990). Although several bacterial species naturally produce PHB from sugars, recombinant production in *E. coli* has increased product yields and simplified downstream purification steps (Nikel *et al.*, 2006; Tyo *et al.*, 2006). The engineered host has been constructed by several research groups through the introduction of the three necessary enzymes from one of several species, such as *Ralstonia eutropha, Cupriavidus necator, Alcaligenes latus,* and *Streptomyces aureofaciens* (Choi *et al.*, 1998; Mahishi *et al.*, 2003; Nikel *et al.*, 2006). The resulting strain carries out the conversion of acetyl-CoA, an abundant intermediate of central carbon metabolism, to PHB and accumulates the product intracellularly. Cost reduction of PHB production has been accomplished by the use of inexpensive carbon sources such as biomass hydrolysates (Keenan *et al.*, 2006; Lee, 1998), but the fact that aerobic conditions are needed implies high energy consumption and thus negates many of the benefits offered by PHB (Harding *et al.*, 2007). A recent metabolic engineering effort aimed at circumventing this limitation by placing the operon under the control of anaerobic promoters (Wei *et al.*, 2008). Further genetic modifications and process improvements could render PHB competitive with synthetic plastics, even at low oil prices.

Apart from bulk chemicals, specialty and therapeutic compounds have also attracted the attention of metabolic engineers. Such is the case of the development of production platforms for lycopene and other carotenoids (Klein-Marcuschamer *et al.*, 2007), glycosylated proteins (Hamilton *et al.*, 2003) and other biotherapeutics (Gerngross, 2004). While most of the early success almost exclusively relied on introducing a particular enzyme or set of enzymes into a host cell (Betengaugh & Bentley, 2008), the recent explosion in the volume of gene and protein data significantly improved understanding of cellular metabolism and genetic regulation. Advances in microbial genetics and plant biotechnology have emboldened metabolic engineers to take on grander and more pressing challenges such as energy, climate change, human health, and others.

Admittedly, undertaking these tasks is easier said than done, but monumental as these challenges may seem, the incessant technological developments – most notably, recent innovations in gene sequencing (Shendure *et al.*, 2004), *de novo* oligonucleotide synthesis (Tian *et al.*, 2004), *in silico* enzyme design and protein engineering (Dwyer *et al.*, 2004), "omics" tools (Park *et al.*, 2008), and synthetic biology (Benner & Sismour, 2005; Sprinzak & Elowitz, 2005) – provide several reasons for metabolic engineers to remain optimistic.

2.4 Limitations of Rational Approaches

The vast literature available about the progress of metabolic engineering using rational approaches may suggest at first glance that phenotypic improvement can generally be achieved by simple and directed genetic modifications. For such efforts to be implemented, the researcher must have at minimum a working hypothesis, if not a clear understanding, about the biosynthetic reaction network, along with its kinetics and regulation. Not only is a hypothesis about what genetic modifications are likely to result in the desired phenotype crucial, but it is also imperative that the trait of interest depends on no more than a few genetic determinants. These requirements stem from the fact that the list of genetic modifications (including gene deletions, overexpressions, alteration of regulatory signals, relief of feedback inhibition processes, etc.), are infinite even when considering only a few targets for manipulation. In addition, the time it takes to consider even a handful of such modifications is long enough that simultaneous consideration of many hypotheses is impractical. In hosts that have not been "domesticated" enough, genetic modifications may be impossible altogether. In summary, for traditional approaches to be adequate, the phenotype to be engineered must be mechanistically simple and relatively well understood.

Out of the three parameters that are targeted for optimization – yield, titer, and productivity – yield is most amenable to traditional approaches consisting of directed and rationally-selected genetic changes. Yield depends mainly on the structure of the metabolic reaction network, and, once the network is described, mathematical and computational tools can be used to optimize its structure to maximize the amount of substrate that ends up as product (Burgard *et al.*, 2003; Durot *et al.*, 2009). Once an optimal network is constructed *in silico*, the working hypotheses can be tested *in vivo*.

Titer and productivity, on the other hand, depend on interactions between the entire metabolic network and the conditions of growth (temperature, pH, media composition, etc.), which complicates matters because the properties of *all* biomolecules and their chemical interactions are affected by those same parameters. Therefore, high titers and productivities in a particular set of conditions are usually, though not always, complex phenotypes. Environmental tolerance in particular becomes a target for improvement when the process conditions that are needed for profitability differ significantly from those in which the strain of interest is naturally found. If possible, the metabolic engineer should select a host that has evolved under the conditions similar to those of interest, as this would lower the likelihood of limitations arising from environmental toxicity.

The prerequisites for host selection described in Section 2.2.1, added to the need for a robust microorganism, has stimulated the development of phenotypic improvement tools for dealing with complex and poorly understood traits. Most of these, as will be explained in greater detail in the next Chapter, are based on random modifications of the genetic material and subsequent isolation of the variants that show improvements in phenotype. In some cases, directed genotypic changes have been used to improve phenotypes that are usually reckoned as complex. Conceptually, this is not surprising, given that cellular systems are themselves complex with a hierarchical regulatory structure that allows them to use a few nodes to control many others simultaneously (Jeong *et al.*, 2000; Martinez-Antonio *et al.*, 2008). For example, overexpressing a regulator that coordinates the response to a certain stress may improve the tolerance phenotype because the regulator itself coordinates a complex set of reactions. In other cases, one or a few genes have evolved for protecting against exactly the environmental condition that limits growth, and, therefore, we can replicate the protecting effect with a few modifications.

One instance is illustrated by Fiocco *et al.*, who overexpressed heat shock proteins in *L. plantarum* to alleviate growth inhibition at higher-than-normal temperatures (Fiocco *et al.*, 2007). Examples similar to this are few and tend to be the exception rather than the rule. More importantly, even when they deliver initial improvements, these are commonly modest. When no additional changes are obvious, further optimization requires the use of random methods.

Thus, comprehensive metabolic engineering programs consist of (i) constructing a strain that catalyzes the needed biochemical conversion efficiently using directed genetic modifications (traditional approaches), and (ii) improving its fitness to perform under the required process conditions. The order of these two steps could probably be reversed without much consequence to the final result, although evidence is lacking in this regard. During the course of the research presented by this thesis, we improved several complex phenotypes (production of lactic acid, L-tyrosine, and ethanol) in strains that were already competent in the production of the compound of interest, either naturally or through directed genetic modifications. In any case, it is important to emphasize that random approaches for fitness improvement are pursued in combination with traditional metabolic engineering approaches, and it is important to know the promises and limitations of both to apply them productively.

Chapter 3

3. Random Searches for Phenotypic Improvement

The rational approaches described in the previous Chapter may be the trademark of metabolic engineering, but random approaches based on evolutionary principles have also generated considerable interest. The list of experimental methods that belong to this category is rapidly expanding, encouraging metabolic engineers to develop supporting tools and to gain a deeper appreciation of the fundamentals behind them. We now turn to explain and examine these methods in light of the general goal of the thesis.

The evolutionary process is random and, given that humans have for long depended on natural variation for successful use of selective breeding, random searches for phenotypic improvement are rarely a new enterprise. The scientific counterpart to this process has consisted in both trying to understand and exploit the molecular basis of variation and inheritance, and in developing reliable techniques for selecting the variants of interest from a diverse population. These concepts have been used once and again for engineering everything from biomolecules to entire cells. Although the term 'directed evolution' has been traditionally used in the context of protein engineering, the basic premises behind it can be applied to improvement of cellular phenotypes, which is the connotation given in the present discussion.

In this Chapter, we first describe the principles that allow random searches for phenotypic improvement, we then provide a set of examples that illustrate the increasing interest in applying these principles, and finally focus on transcriptional engineering as a particular case belonging to this family of tools. This discussion sets the stage for the hypotheses and experiments that follow in subsequent Chapters.

3.1 Evolutionary Principles Can be Exploited for Phenotypic Modification

Evolution has been defined in different ways throughout history, although with respect to biological systems, it specifically refers to the change in inheritable traits produced by the combination of three processes: variation, selection, and reproduction. Selective breeding for improving species of interest has been conventionally based on natural variation, partner selection, and reproduction. As such, it is the artificial character of the selection pressure which has delivered the desired results. This process is slow, but has delivered notable outcomes considering we have applied it for no more than a few millennia. The molecular basis of evolution remained unknown for most of our history, but a period of intense scientific inquiry enlightened us to the point we now apply evolutionary principles to engineer a wide variety of biological systems.

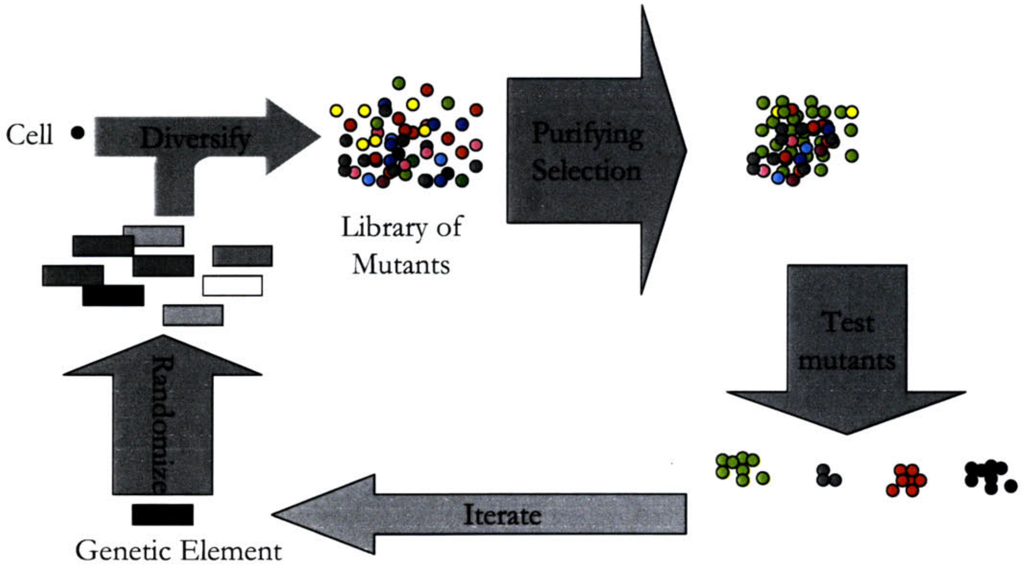

Figure 3.1-1. Evolutionary approaches for strain improvement

Genotypic diversity can be introduced into a cell by randomizing genetic elements, resulting in a library of mutants. One can then search for phenotypes of interests (different phenotypes are shown by different colors) by purifying selection. Individual mutants can then be tested separately to confirm the presence of the trait. Iteration, in lighter gray, is optional, and is used to refine the search.

Simply put, the evolutionary approach consists of the iteration of two steps: the creation of genotypic diversity and the isolation of the variants that have interesting properties (**Figure 3.1-1**). As these overarching principles began to emerge, scientists progressively relaxed the original attributes of the concept of evolution (which is tied to the process of speciation). For example, molecules have become analogous to species, and *in vitro* amplification akin to reproduction. Variation is no longer naturally-engendered nor is the result of sequentially unfaithful replication of the genetic material, but is introduced in a highly parallel manner by the experimenter. Such mental framework has allowed the development of techniques such as SELEX (systematic evolution of ligands by exponential enrichment), in which randomly generated and successively amplified RNA variants are selected for their ability to bind a ligand (Ellington & Szostak, 1990), and gene shuffling (Stemmer, 1994).

In the context of metabolic engineering, complex cellular phenotypes, such as those described in Section 2.4, have become a major target for the application of evolutionary methods. These are appropriate when one has the experimental tools to substitute the effort of gaining deep mechanistic understanding of the traits of interest by (i) the creation of cell populations (libraries) with meaningful phenotypic diversity, and (ii) the development of a relevant selection or screening scheme that allows isolation of improved variants.

Environmental tolerance phenotypes are particularly suited for these approaches, since they remain poorly understood and the selection step can be associated with growth. Metabolite overproduction has also been a major area where library-based methods have been employed (Baltz, 2001; Demain & Solomon, 1986). Nonetheless, as will be discussed later, selection or screening procedures are time- and resource-intensive, even when based on growth (Bonomo *et al.*, 2008; Demain *et al.*, 1999; McDaniel *et al.*, 2001). We now examine some of the techniques commonly used for generating genotypic diversity and for isolation of strain variants with improved traits.

3.1.1 *Generation of Diversity*

There is a long list of techniques that have been developed for and devoted to the generation of genotypic diversity. When based on mutagenesis, which will be the focus of this discussion (others will be described later), the methods differ in the location of the mutations, the mutation bias or identity (i.e. what bases are most commonly altered, whether transversions are less common than transitions, etc.), the type of host in question, and others.

3.1.1.1 Whole-cell Mutagenesis

The most general way of effecting mutagenesis, and historically the most common, is whole-cell mutagenesis, which implies that all the DNA content of the cell is targeted for modification (when followed by selection, the protocol is usually referred to as 'classical strain improvement,' or CSI). This approach allows the researcher to treat the system as a 'black box', and is appropriate when (i) even basic mechanistic understanding about the phenotype to be improved is lacking (or the supposed mechanism is intricate) so that one cannot guess where mutations are likely to influence it; (ii) the host is hard to manipulate genetically (see Section 2.2.1); or (iii) a combination of (i) and (ii).

Whole-cell mutagenesis can be accomplished by physical or chemical means. Undoubtedly, the most widespread method for the former is exposure to ultraviolet (UV) radiation. UV-rays cause a variety of photochemical reactions in DNA, such as the intrastrand dimerization of adjacent pyrimidine bases, which result in mismatch repair and eventually to fixation of the mutations (Witkin, 1976). Chemical mutagenesis is most frequently implemented by DNA-methylating substances (e.g. methyl methane sulfonate, dimethyl sulfate, *N*-methyl-*N*-nitrosourea, and *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine or simply nitrosoguanidine) or by ethidium bromide. Methylating agents are electrophilic and readily react with various nucleophilic positions in DNA, forming N-methyl or O-methyl adducts (Wyatt & Pittman, 2006). Some of these adducts are mutagenic because they cause mispairing and some because they lead to the formation of abasic positions following depurination (Wyatt & Pittman, 2006). Ethidium bromide, on the other hand, is a DNA intercalating agent that induces mutagenesis by obstructing topoisomerase function and by causing DNA fragmentation (Schneider-Berlin *et al.*, 2005; Turner & Denny, 1996).

Whole-cell mutagenesis continues to be a common technique, being conceptually straightforward and enjoying an extensive record of successes. One of the legendary instances of this approach boasts a 4000-fold improvement in penicillin production (Parekh *et al.*, 2000). Another, more recent and modest (yet impressive) example of the use of this technique for improving titers of the enzyme glucose oxidase in *Aspergillus niger* fermentations was reported by Singh (Singh, 2006). Chand *et al.* used a similar approach to augment the production of cellulases in a fungal strain, an application that has immediate impact in the production of commodity chemicals from biomass-derived sugars (Chand *et al.*, 2005).

For all its successes, whole-cell mutagenesis has also several limitations. The sequence space (i.e. the set of all sequences that are theoretically possible in a collection of variants) for undirected mutagenesis of entire genomes, which are usually millions of bases long, is for all practical purposes infinite. This implies that the probability of success is low, or that the library to be screened must be large. As will become apparent in the next section, large library sizes present challenges to the isolation step. Another disadvantage of whole-cell mutagenesis is the accumulation of deleterious base changes as the evolutionary cycle is iterated (**Figure 3.1-1**). The result is the isolation of specialized, but overall unhealthy, variants that are likely to underperform when placed in the mixture of stresses inherent of most bioprocesses (Sauer, 2001).

Genome shuffling was recently applied to circumvent the drawbacks imposed by the appearance of crippling mutations. Although based on a combination of old concepts and techniques, the method has awakened renewed interest. After an initial round of mutagenesis and purifying selection, the strains are pooled together and fused via protoplast formation. Subsequent rounds of selection, mutagenesis, and fusion are performed until the desired result is obtained. According to the authors, this amounts to multi-parental sexual exchange of genetic material, so that deleterious mutations can be eliminated while favorable ones are retained and become fixed (Patnaik *et al.*, 2002; Zhang *et al.*, 2002). This technique has been replicated for a growing list of applications in the recent years (Hou, 2009; Kalia & Purohit, 2008; Shi *et al.*, 2009; Wang *et al.*, 2007; Yu *et al.*, 2008b).

3.1.1.2 Region-wide Mutagenesis

When the researcher has some clues regarding where mutations are likely to influence phenotype, more targeted approaches can be pursued. For example, if one or a few enzymes are known to be limiting, mutagenesis can be directed to the subregion of the genome that codes for these enzymes. This route has been followed for relieving feedback inhibition at enzymes that are known to be controlled by the accumulation of a downstream metabolite (Lutke-Eversloh & Stephanopoulos, 2005). Mutagenesis of the feedback-inhibited enzyme may result in a resistant variant that frees the flux of the pathway and allows overproduction of the suppressive metabolite.

The advent of molecular biology brought with it tools for effecting mutagenesis aimed at a known stretch of DNA. One of the most popular remains to be error-prone PCR, based on *in vitro* replication of DNA either with mutagenic polymerases or with conditions that compromise the ability of conventional polymerases to carry out faithful incorporation of bases (i.e. according to complementarity). For example, engineering of a *Pyrococcus furiosus* thermostable polymerase rendered it unable to perform 3'→ 5' exonuclease-mediated proofreading, which, added to other modifications resulted in an enzyme with overall lower fidelity (Biles & Connolly, 2004). Using nucleotide analogues, adding manganese ions (Mn^{2+}), or altering the ratio of bases in the PCR mixture may also produce a mutagenic reaction (Wang *et al.*, 2006).

One limitation of PCR-based methods is that the sequence diversity that results is restricted and biased. With single base mutations per codon – a common assumption with most protocols – only 5.7 amino acids are accessible per position on average, and in most cases the resulting set of amino acids does not accurately represent the spectrum of physicochemical properties of naturally-occurring residues (Miyazaki & Arnold, 1999).

3.1.1.3 Position-specific Mutagenesis

Although epPCR is a very common technique for relatively targeted mutagenesis, in some instances the researcher can guess with more precision the nature or location of the required mutations. For example, one can mutate the amino acids near or at the active site of an enzyme in order to increase the activity, change the substrate specificity, or the product spectrum (Ohnuma *et al.*, 1996; Yoshikuni *et al.*, 2006a; Yoshikuni *et al.*, 2006b).

Targeting changes to specific positions is usually accomplished with the use of synthetic DNA. In contrast to amplification-based methods, synthetic DNA technology for library construction allows, at the very least, specifying the location where mutations are possible (e.g. saturation mutagenesis), and ultimately permits completely designing the desired sequence diversity. A challenge for such high-resolution targeted mutagenesis is that detailed knowledge regarding where genotype changes are likely to affect phenotype is needed. Such knowledge can be obtained experimentally, through preliminary rounds of error-prone PCR, or computationally, based on structural information (Voigt *et al.*, 2001).

In the extreme case where the sequence diversity is entirely specified, the evolutionary effort is reduced to finding the mutant with most fitting properties. The reduction and design of the search space is therefore a key motivation for using synthetic DNA libraries instead of epPCR for creating sequence diversity. Synthesis technologies based on sequential elongation of DNA molecules with codon-sized fragments will likely become more popular for these applications (Van den Brulle *et al.*, 2008), but oligonucleotide-based methods have remained most popular.

3.1.1.4 Other Strategies for Diversity Generation

The three general categories of genotypic diversity generation – whole-cell, region-wide, and position-specific mutagenesis – are all explored at different times and for different library designs in the present thesis. Even though so far we have only discussed mutagenesis-based methods for generation of genotypic diversity, other ways exist. For example, a gene overexpression library, in which a genome is fragmented randomly and placed under control of a relatively strong promoter, is a way of introducing diversity that does not depend on mutagenesis. As such, a library so constructed and the subsequent screening or selection step constitute a random search strategy for phenotypic improvement. Many others exist, as will be illustrated in a slightly different context by a few examples found later in this chapter. We now turn to describing the technical considerations of the second step of evolutionary approaches for phenotypic improvement: purifying selection.

3.1.2 *Purifying Selection*

Once diversity is introduced into a population, the variant or variants with desired traits must be found. We assume, at least for now, that the mutant of interest is in fact present in our diverse population, and that the remaining task is isolating it from the rest. The term 'purifying selection' will be used to refer to any experimental technique that aims at enriching the variant of interest with respect to other variants. Most of the literature on the topic usually distinguishes, though rather implicitly, between the terms 'selection' and 'screening', the former referring to enrichment using a growth advantage of the improved mutant, and the latter referring to enrichment using discrepant performance of mutants with respect to a phenotypic assay (e.g. metabolite analysis, cytometry, etc.). Since both protocols serve the same purpose, we bundle them together in the concept of 'purifying selection' for the purposes of this discussion.

3.1.2.1 Selection Based on Growth

Selection based on a growth advantage of the improved mutants can be carried out whenever the phenotype of interest can be associated with the cell's ability to reproduce, as when dealing with environmental tolerance. For example, if a strain needed for the production of a certain compound has been optimized for yield but grows poorly under the actual process conditions, subjecting the library of mutants to those conditions may enrich for tolerant variants. If the compound of interest is toxic, so that final titers remain low, selection under high concentrations of the compound may deliver a mutant that attains higher titers. Growth advantage may be used for metabolite or enzyme overproduction applications by enriching in media containing a metabolic inhibitor called antimetabolite. The antimetabolite forces the pathway in question to be hyperactive if it is to allow faster growth. These techniques are common, so it is constructive to illustrate them with a few examples.

1. Ethanol has gained significant attention as a potential biofuel following the recent increase in oil prices, the need for energy security, and flourishing arguments on the importance of reducing carbon dioxide emissions. The most common route to ethanol today is fermentation of sugars by the yeast *S. cerevisiae*. To avoid cooling costs and curb contamination, the bioprocess should be run at relatively high temperatures, but the wild-type yeast grows optimally at 30 °C. At least two groups have successfully implemented genome shuffling for improving thermotolerance in this microorganism. After rounds of mutagenesis and shuffling (as described in Section 3.1.1), improved mutants were selected for their ability to grow at high temperatures (in fact, both temperature and ethanol were used as challenges for selection). The isolates from the two studies (Hou, 2009; Shi *et al.*, 2009) were able to grow at 50 and 42 °C, respectively.

2. In order to improve the assimilation of starch by a solvent-producing strain of the genus *Clostridium*, Annous and Blaschek used nitrosoguanidine-mediated mutagenesis to generate a diverse library. Then, they enriched the mutated population for strains with enhanced amylolytic activity by growing them in the presence of the antimetabolite 2-deoxyglucose, a glucose analogue that cannot undergo full glycolysis (Annous & Blaschek, 1991). The improved strain was reported to produce nearly 2-fold higher amylolytic enzyme, compared to the parental.

Selection strategies based on growth seem, at first glance, uncomplicated and dependable as means to increase titer and productivity. Unfortunately, the conditions during selection are not always relevant for the phenotype we try to improve; for example, our unpublished observations show that enrichment under high concentrations of ethanol may deliver better ethanol-producing strains in some cases, but not in others. Other experimenters have pointed out that growth-based selection is hard to control and poorly understood (Bonomo *et al.*, 2008). Experiments frequently result in false-positives and the outcomes can be difficult to reproduce.

3.1.2.2 Screening

Enrichment based on phenotypic assays is performed when an analytical method for quantifying the differential performance of improved mutants with respect to a trait of interest is available. Because of the large population sizes of most libraries, the quantification method must be high-throughput, that is, it must be fast, use small volumes, and be adaptable for studying individual colonies. Compared to the case of selection by growth, screening is even more application-specific, and a method based on the goal to be accomplished must be developed.

Screening based on multi-well fermentation of clones and subsequent analysis of the fermentation broth has been a popular scheme for isolating mutants with improved production of extracellular metabolites (Demain *et al.*, 1999; Isett *et al.*, 2007; Kittell *et al.*, 2005). Other methods include visual inspection or spectrophotometric assays, but, once again, their applicability is highly dependent upon the properties of the compound to be studied (Alper & Stephanopoulos, 2007; Baltz & Seno, 1981; Pfleger *et al.*, 2007; Smolke *et al.*, 2001).

For example, Zhang and coworkers aimed at improving the production of the polyketide antibiotic tylosine by sequential rounds of genome shuffling. Screening was done spectrophotometrically by measuring the absorbance of prepared supernatants at 290 nm (Zhang *et al.*, 2002). A different group used GFP-producing mevalonate auxotrophs as biosensors for high-throughput visual inspection-based screening of mevalonate-overproducing strains (Pfleger *et al.*, 2007). More recently, members of our group reported a colorimetric method for identification of L-tyrosine overproducer strains based on the black pigment melanin (Santos & Stephanopoulos, 2008b). L-tyrosine, a colorless amino acid, is a substrate in the melanin-production pathway, and thus this colored compound can be used to spot mutants with high tyrosine production capability.

Being more widely applicable and general in scope, multi-well-based fermentations with subsequent analysis via HPLC, GC-MS, or similar analytical techniques is still the method of choice. In many cases, high-throughput screening methods may be automated, but they remain a very costly and time-consuming step; when performed manually, the capital cost incurred may decrease, but continues to be a major expense of phenotypic improvement programs (Demain & Solomon, 1986; Demain *et al.*, 1999; McDaniel *et al.*, 2001).

3.2 Random Search Strategies and Tools for Phenotypic Improvement

The two key steps of evolutionary approaches for phenotypic improvement – diversity generation and purifying selection – can be associated one-to-one to the words in the term 'random search'. The undirected nature of the genotypic diversity generation step constitutes the basis of the term 'random', while the fact that one must find an improved variant from a large population is the root of the term 'search'. Furthermore, the word 'search' alludes to the possibility that one may not find an improved variant; the approach is, by nature, open-ended. This is still true even if a variant that has precisely the phenotype of interest is theoretically attainable given a certain library design. Let us examine a few reasons why this may be the case.

3.2.1 *Challenges of Random Search Strategies*

Firstly, search spaces in most evolution-based experimental designs are astronomically large. For example, the number of variants that can be constructed with a DNA sequence of length N is 4^N (for a protein with N residues, the number is 20^N), a space that is impossible to cover experimentally unless N is very small. This implies that, if there are only a few sets of base combinations that will deliver a trait of interest (for a sufficiently large N), the probability of obtaining an improved individual is negligible. Consider an epPCR library for a 300-residue enzyme that is being engineered to exhibit a new functionality. If the only possible improved variant contains three specific mutations (e.g.

S20T, A187G, M201R), the probability of finding it in an optimally designed library (one with an adjusted mutagenesis rate that results in an average of three mutations per sequence) is about 1 in 10^{12}. Since the epPCR library sizes attainable (in *E. coli*) are in the order of 10^{6}, we would need to construct and screen millions of libraries in order to be certain to find this mutant. In fact, even if this three-mutation variant enjoyed several-fold improvement compared to the parental, the low probability of success suggests that this enzyme should be considered a poor choice for engineering the novel functionality. If we now consider whole-cell mutagenesis for improving a complex phenotype that hinges on changes in a few distant genes subject to epistatic interactions (Applebee *et al.*, 2008), the challenges associated with the size of the search space becomes even more pronounced.

Secondly, the experimental protocols for building and screening libraries are subject to a variety of stochastic effects. Let us continue with our epPCR example. The desired mutant may be present after the mutagenic amplification, but it may not be properly cleaved during the restriction reaction or not be correctly ligated to the vector. The mutant may cause an indirect effect in physiology so that it is subject to a negative selection pressure, diluting its presence in the final library pool even before screening begins. The mutant may not perform during screening as it would during the conditions that are ultimately of interest, obscuring its presence in the library. The mutant may be enriched significantly during screening, but it may still remain unnoticed when testing individual clones. In summary, the fact that a mutant is theoretically attainable with a particular library design is no guarantee that it will be eventually isolated.

Fortunately, the sum of all these factors does not prevent the discovery of improved phenotypes, as evidenced by the overt popularity of random searches for phenotypic alteration. We can interpret this success as a suggestion that there is a sizable subset of library designs for which the probability of finding phenotypes of interest is within experimental reach. It must be noted that not all library designs will serve for improving a particular phenotype, but there are a few designs that serve for improving many traits (e.g. whole-cell mutagenesis, genome shuffling, knockout libraries, among others). Let us now turn to more examples of random search-based library designs that, albeit they suffer from the aforementioned limitations, their variety adds to the potential of the evolutionary approach in general.

3.2.2 *Example Library Designs*

A few library designs have been covered in Section 3.1.1, when discussing the generation of genotypic diversity via mutagenesis. These techniques were described in great detail for two reasons. First, mutagenesis-based methods for library construction are most relevant to the present thesis. Second, if mutagenesis is analyzed on purely conceptual grounds, all library designs could be theoretically based on this principle, if we allowed the construction and searching of infinite spaces (for this to be true, we must also permit the introduction of additional stretches of DNA, which would allow for designs that depend on plasmid-borne genotypic determinants). Therefore, mutagenesis presents an ideal framework for explaining many other library designs. Let us consider a few examples.

Knockout libraries can be constructed via random insertion of an antibiotic marker cassette, with the aid of the enzyme transposase (Santos & Stephanopoulos, 2008a). The result is a collection of mutants with disruptions throughout the genome, allowing inactivation of genes or operons, alteration of their regulatory features, or a combination of both. Even for the case of a finite and relatively manageable number of genes in a genome, transposon libraries can be infinitely genotypically diverse (assuming no insertional bias), thus potentially suffering from the challenges associated with large search spaces outlined previously. However, this is in practice a relatively inconsequential problem, and transposon libraries have been used fruitfully to deliver phenotypic improvements. For example, this approach was used to identify targets for enhancing lycopene production in a recombinant strain of *E. coli* (Alper & Stephanopoulos, 2008). Such an approach may uncover deletions that are not predicted by *in silico* modeling, and yet can affect the phenotype in question due to epistatic or regulatory effects.

Gene overexpression libraries provide an example that is seemingly opposite to the knockout library design delineated above. A genome is randomly cut with restriction enzymes and the inserts are cloned downstream of a relatively strong promoter (Jin & Stephanopoulos, 2007; Lynch *et al.*, 2007). The genotypic diversity introduced is based on (i) the cloned gene (or genes) being expressed at a higher level compared to the wild-type; (ii) the disruption or alteration of regulatory mechanisms to which the gene (or genes) are subjected to; and (iii) the merodiploid nature of the resulting population, which may become pertinent if new mutations arise. Gill and coworkers have recently used this method for studying and optimizing selection strategies, and were able to identify genes that confer tolerance to 3-hydroxypropionic acid and 1-naphtol, among other stresses (Gall *et al.*, 2008; Warnecke *et al.*, 2008).

The library designs so-far described are based on common genetic modifications, but the list of random search-based phenotypic alteration strategies has grown significantly in recent years, becoming ever more creative in both methods and applications. For example, Wang *et al.* reported a method for constructing libraries of randomized short-hairpin-loop RNAs (shRNAs), which can alter the phenotype globally by targeting multiple genes both through activating and silencing pathways (Wang *et al.*, 2008). The method was used to enhance survivability of murine pro-B FL5.12, an interleukin-3-dependent cell line. The improved variant was able to survive nearly two-fold better than the parental after IL3 withdrawal (Wang *et al.*, 2008). Even though silencing pathways have been thought to act in a directed fashion, this and other research groups (e.g. (Jackson *et al.*, 2003)) have described off-target regulation by small interfering RNAs (siRNAs), implying that randomized siRNA libraries may be helpful to alter phenotype globally. Such library designs may therefore present interesting prospects for engineering complex phenotypes.

One last example of library design concerns the less explored, but potentially valuable, method based on randomized ribozymes. Similar to siRNAs, ribozymes can target transcripts for silencing by catalyzing RNA cleavage, and until now have mainly been used for gene discovery applications (Miyagishi *et al.*, 2005). However, similar to siRNAs, they can conceivably be used for phenotypic improvement: libraries of randomized ribozymes have been used to introduce genotypic diversity, and, as long as they produce phenotypically diverse populations, they can also be used for evolutionary approaches for cell engineering (more of this topic in Chapter 5).

3.3 Transcriptional Engineering

Among this plethora of random search-based methods for strain improvement, there is a sub-category that concerns to those in which diversity is introduced at the level of the transcriptome. Naturally, given that for all evolutionary approaches the determinant that causes an alteration in phenotype must be inheritable, this transcriptomic diversity should be based in genotypic changes. The term transcriptional engineering, as defined here, refers to methods and techniques based on this general premise, although the level and scope of the transcriptomic changes may vary widely. This definition does not exclude directed modifications to the transcriptome, but we hereby only refer to random search strategies that aim at global transcriptional alteration when using this term.

Transcriptional engineering, in its most general way, is not a new concept. It is based on the idea that the transcriptome is closer to phenotype than the genome is, an observation that logically emerges from the structure of information flow in the cell. Moreover, it is known that non-coding regions of the genome evolve faster than coding regions, because changes in the former tend to be more forgiving to cellular physiology (Molina & van Nimwegen, 2008). Transcriptomic information is known to vary widely in different growth conditions and during different stages of the life cycle. A recent study showed that rapid speciation can be partly attributed to a divergence in transcription factor binding patterns, suggesting that changes in regulation lead to measurable phenotypic changes in yeast (Borneman *et al.*, 2007).

Although the sizable amount of evidence supporting the use of transcriptomic changes for altering phenotype is not all recent, the application of this concept to engineering complex traits is relatively novel. We now describe the research that led onto the development of some of the methods put forth by the present thesis.

3.3.1 *Background*

Probably the first attempts to engineer phenotype using transcriptional regulators were based on the use of artificial zinc-finger proteins (ZFP). The zinc-finger domain, sometimes called Cys2-His2 or C_2H_2, is the most common DNA-binding motif in eukaryotes, and has been exploited for the design of artificial transcription factors (Beerli & Barbas, 2002). ZFPs are made of DNA-binding modules and can be easily fused with activator or repressor domains, depending on whether one aims to upregulate or downregulate the target genes (Segal *et al.*, 2003).

Artificial ZFPs have been primarily applied in eukaryotic systems, and mainly with the admitted intention of targeting specific locations in chromosomal DNA (Beerli & Barbas, 2002). For example, Zhang and coworkers reported on a set of ZFPs that target and activate the human erythropoietin gene endogenously (Zhang *et al.*, 2000). A similar experiment showed the successful activation of vascular endothelial growth factor A, an important inducer for the formation of blood vessels that may result in embryonic lethality if expressed at low levels (Liu *et al.*, 2001). Other studies have illustrated similar achievements using artificial ZFPs (Bartsevich & Juliano, 2000; Falke *et al.*, 2003; Ren *et al.*, 2002).

A review of the ZFP literature suggests that what has driven research in this area is the optimization of binding specificity, affinity, and overall stability. Maximizing specificity is helpful for effecting directed transcriptomic modifications, simplifying many experimental protocols in eukaryotic genetics (Jamieson *et al.*, 2003). This school of thought has temporarily delayed the use of artificial ZFPs for engineering complex phenotypes, but, as we will see next, such tools have proven effective for altering many genes at once.

3.3.2 *Artificial Transcription Factor Libraries for Alteration of Complex Phenotypes*

The study that most probably was responsible for setting the stage for the use of transcriptional engineering applied to improving complex phenotypes was that of Park and coworkers (Park *et al.*, 2003). The study describes the use of ZFP-based artificial transcription factor libraries to improve thermotolerance, ketoconazole-resistance, and osmotolerance in *S. cerevisiae*. The method involved random assembly of ZFP domains and subsequent fusing with activator or repressor domains, which resulted in a collection of artificial transcription factors (ATFs) that were transformed into yeast cells. The so-constructed library was then used to isolate improved clones in different conditions; the responsible ATFs were identified by sequencing.

For the case of thermotolerance, the group was able to increase the survivability of yeast from 0.04% for the wild-type after 2 hr at 52 °C, to 10% in some of the isolates transformed with the ATFs. The phenotype was transferrable and also dependent upon induction of the ATF. For the case of osmotolerance, the selection conditions were 100 mM LiCl, and an up to 100-fold improvement in survivability was observed in some cases. Finally, the libraries were exposed to 35 μM ketoconazole, an antifungal agent that is widely used but for which resistance may develop. In an effort to understand the mechanism of resistance, they performed microarray analysis of three ATFs that conferred ketoconazole-resistance. Four open reading frames (ORFs) showed more than 2-fold overexpression compared to the wild-type across the three isolates. A fraction of the phenotype could be recovered by overexpression of one of these ORFs, identified as YLL053C, allowing the authors to hint at a possible mechanism.

This very complete study opened the possibility for phenotypic improvement by introducing global transcriptomic changes. It parted from the previous approach of using ATFs for directed modifications, and focused on complex traits. Although apparently not integral to their original hypothesis, the researchers present enough evidence to support that the improved phenotypes arise from orchestrated changes in transcriptome. For instance, they found that the ketoconazole-resistance phenotype was not due to sequence-specific interactions of the ATFs. Even though they identified four ORFs that were activated by the three studied ATFs, only one conferred resistance when overexpressed, and only a fraction of the phenotype was recovered. Furthermore, some genes known for their action against the drug were activated by some, but not all, of the isolated ATFs, implying that there is more than one mechanism responsible for the observed phenotype. All this suggests that the improvement was caused by the ability of the ATFs to make many *simultaneous* changes in the transcriptome.

Ensuing studies with ZFP-based ATFs supported their use for global transcriptomic modification for strain improvement in various systems. Lee *et al.* improved taxol-resistance in a HeLa cell line using ATF libraries and various cycles of purifying selection (Lee *et al.*, 2004). Microarray analysis of two of the isolated studies revealed nearly 200 differentially-regulated genes compared to the wild-type, providing further proof that these ATFs act at many targets simultaneously. Out of these, 37 genes were found in both microarrays. Even if it was only this set of genes which bestowed taxol-resistance, engineering this trait with directed genetic modifications would still present an intractable experimental endeavor.

The system was also applied to prokaryotic systems, in particular to *E. coli* (Lee *et al.*, 2008; Park *et al.*, 2005a). In the more complete study, Lee *et al.* explore the use of ZFP-based ATFs to improve tolerance to heat, cold shock, or osmotic pressure (Lee *et al.*, 2008). They not only identify ATFs that can confer the improved phenotypes in a transferrable fashion, but also attempt to explain the mechanism behind the thermotolerant isolate. Their results complement the previous arguments stating that (i) the improvement arises from the coordinated change in the transcription (up and downregulation) of many genes; (ii) the trait cannot be reproduced by a few directed genetic manipulations; and (iii) the mechanism behind the improvement is complex and resists simplification.

3.3.3 *Native Transcription Factors and Global Transcription Machinery Engineering (gTME)*

The work with ATFs provided a basis and a motivation to further the transcriptional engineering approach, but it had overlooked the potential of manipulating the natural regulatory mechanisms of the cell. In particular, transcription is coordinated by a few nodes in the physiological network (Isalan *et al.*, 2008; Martinez-Antonio *et al.*, 2008), and those can be targeted for altering complex phenotypes.

With this in mind, Alper and Stephanopoulos considered the use of the principal sigma factor of *E. coli*, sigma D (coded by *rpoD*), for transcriptional engineering (Alper & Stephanopoulos, 2007). Several studies had shown that mutations in this protein alter the specificity of the RNA polymerase (RNAP) for its target promoters (Gardella *et al.*, 1989; Siegele *et al.*, 1988; Siegele *et al.*, 1989; Waldburger *et al.*, 1990), and, considering the centrality of *rpoD* in the physiological network (Martinez-Antonio *et al.*, 2008), the change in specificity would alter the transcriptome globally. They termed this approach global transcription machinery engineering (gTME). Even though most of our description of this method is not expounded until the next Chapter, we briefly discuss some results in this Section to provide some background on the early achievements of the gTME approach.

The initial study focused on ethanol tolerance, lycopene overproduction, and simultaneous ethanol-SDS tolerance in *E. coli*. Ethanol tolerance was significantly improved after three rounds of epPCR-based mutagenesis and growth selection (see Section 3.1). The resulting mutant was analyzed using DNA microarrays, showing nearly 100 differentially-regulated genes in the absence of stress and similarly-complex responses in the presence of it. For the case of lycopene, several *rpoD* mutants were isolated that individually conferred a higher improvement in production than previously-explored directed genetic modifications. Finally, the study explored different strategies for enhancing tolerance to ethanol and SDS simultaneously, and found that co-expression of independently isolated mutants was the most promising approach.

In a related study, the TATA-binding protein of *S. cerevisiae* was targeted for mutagenesis, and the resulting libraries were selected under high ethanol and high glucose stresses (Alper *et al.*, 2006). Mechanistic studies determined that the exhibited improvements were a complex function of the three mutations present in the best-performing variant and of the many transcriptional changes elicited by it.

Working with the native cellular machinery has some distinguishing features compared to ATFs. First, the approach resembles natural evolution of novel phenotypes, as it relies exclusively on native genes. Second, isolated variants may act by molecular mechanisms not directly related to transcription; for example, the previously-discussed *rpoD* variant that conferred the highest level of ethanol tolerance was a truncated version of the gene that most likely affects the cellular response by acting on a capacity different from that of a transcription factor (Alper & Stephanopoulos, 2007). Third, even when the variants act by modulating transcription, simultaneous and diverse effects may contribute to the observed traits; for example, competition between the wild-type, chromosome-borne factor and the mutated, plasmid-borne one for available RNAP may play a role in the development of the improved phenotype.

The findings of these two seminal gTME studies are in tune with those based on ZFP-based ATFs described in 3.3.2, and opened the door for finding new routes for engineering complex traits by effecting global changes to the intracellular environment. Before we continue on this topic extensively in Chapter 4, we must put forth a disclaimer regarding all strain improvement approaches.

3.4 The Limited Nature of Strain Improvement

Until now, we have emphasized the potential of strain improvement through the use of both traditional and evolutionary approaches. Before continuing with the description of some experimental achievements of strain improvement, one last remark about the inherent limitations of these approaches is worth stressing. Although obvious, the fact that any genetic changes are made upon an existing genome is sometimes ignored when pursuing a strain improvement effort. This fact implies that there is a theoretical limit for improvement given by the maximum number of changes that can be implemented experimentally, either in a directed or random fashion. This limitation is especially relevant for engineering complex traits.

Let us take the engineering of thermotolerance as an example. This stress is known to unleash a series of events that could kill the cell: unfolding and aggregation of proteins (Villaverde & Carrio, 2003), redirection of metabolic pathways (since kinetics depend on temperature (Moreno-Sanchez *et al.*, 1999)), fluidization of the cellular membrane (Shigapova *et al.*, 2005), among others. Even if chaperones and proteases are overexpressed, thermophilic versions of key enzymes are introduced, membrane properties are modified, etc. with the goal of engineering thermotolerance, the makeup of an organism that makes it tolerant to heat is a property of the system in its entirety. As such, it defies reductionistic approaches. This does not imply that tolerance cannot be improved with respect to that of the wild-type, but that the practical relevance of the limit for improvement depends on the phenotype we choose to engineer. Therefore, the prospects of finding an ideal strain for production of commodity chemicals reside on a balance between choosing the right host as a starting point and choosing which properties to change.

Chapter 4

4. Targets for Transcriptional Engineering in Bacteria

In the preceding Chapters, we have introduced the universal motivation behind phenotypic improvement, we have provided an overview of traditional approaches for modification of phenotype, we have highlighted the many tools that have been developed for applying these approaches, and we have discussed the limitations of traditional methods that invited the adoption of random search-based evolutionary strategies for strain engineering. In this way, we have progressed from the general to the particular, and have arrived at the technique that constitutes the focus of this thesis: transcriptional engineering. We will use this technique both as a subject of study and optimization, and as an example of a random search strategy for testing and exploring general evolutionary principles.

As outlined in Chapter 1, the overarching goal of this thesis has been to improve upon transcriptional engineering in the context of other random search strategies for phenotypic improvement. One way in which such effort was materialized consisted in finding and evaluating new targets for global alteration of the transcriptome. In addition, we proved that old targets (sigma D in this case) can be used in more than one species, by extending the approach to *L. plantarum*.

As we will see momentarily, transcription is an intricately regulated process, coordinated by a sequence of chemical and physical interactions centered in several molecular complexes. Because these interactions are codified by the amino acid sequence of the protein subunits of such complexes, mutagenesis of these proteins allows alteration of the regulation process. As a result, the transcriptome of the cell carrying the mutated regulator can be manipulated in the hope of educing an improvement in a phenotype of interest. In this light, the present Chapter serves two objectives which are simultaneously considered: first, it describes the process of transcriptional regulation and some of its key players; and, second, it illustrates the experiments and results that demonstrate the use of new targets for transcriptional engineering.

4.1 Transcription in Bacteria

Transcription is the first committed step in gene expression and a key step for regulating phenotype in bacteria. The former fact is probably a reason for the latter, since, in order to save resources, the cell should only produce the transcripts for which a product is needed at any one condition. Therefore, transcription initiation ought to be a focal process for manipulation of cellular phenotype, and, accordingly, will serve as a theme throughout this Section.

Transcription in bacteria has been studied most closely in *E. coli*. The process is executed by a single DNA-dependent RNA polymerase (RNAP), which encompasses both the ability to catalyze RNA synthesis and the ability to interact with DNA and protein effectors (activators or repressors) (Browning & Busby, 2004). It is this set of interactions which allows differential expression of genes, thus it is this same set which we aim at modifying during transcriptional engineering.

4.1.1 *The RNA Polymerase: Structure and Function*

The bacterial RNAP has a subunit composition given by $\alpha_2\beta\beta'\omega$ and is capable of carrying out all steps of transcript synthesis except for promoter binding and initiation (Ebright, 2000). The so-called core enzyme assembles first by dimerization of α, and then by aggregation of the β and β' subunits; finally, one of several σ-factors (seven for *E. coli*) binds to the core enzyme to form the holoenzyme, which may bind to promoters and begin transcription (Gourse *et al.*, 2000). The alpha-subunit is composed of two domains, the amino- and carboxy-terminal domains (αNTD and αCTD, respectively); the former is bound to the rest of the polymerase and the latter is tethered to the former by a flexible linker and interacts with different elements at the promoter site (**Figure 4.1-1**).

The holoenzyme is responsible for integrating the vast array of signals into a single output at each promoter, that is, a transcript (or lack thereof). Before we detail the source of these signals, let us describe the general process of RNA synthesis. The material here presented is extremely basic and can be found in any of several reviews that cover different aspects of transcription (Browning & Busby, 2004; Ebright, 2000; Featherstone, 2002; Gourse *et al.*, 2000; Gruber & Gross, 2003; Ishihama, 2000; McClure, 1985; Roberts *et al.*, 2008; Schauer *et al.*, 1996). The first step in transcription is the binding of RNAP holoenzyme to DNA; stronger promoters can recruit the polymerase better than weaker ones, and can also stabilize it for longer once bound. Positioning of the RNAP at the promoter is followed by “melting”, a process in which about 10-14 base pairs upstream of the transcription initiation site, inclusive, unwind and form the so-called open complex. If the open complex is stable enough, the bubble formed by melting grows and allows elongation, or polymerization of the newly-formed transcript; otherwise, the open complex dissolves in a process known as abortive initiation.

The elongation reaction is most plainly described as a succession of three steps: (i) the incoming nucleoside triphosphate (NTP), which is complementary to the DNA counterpart being “read”, positions itself in the catalytic site of the polymerase; (ii) the 3’ hydroxyl group in the growing RNA strand reacts with the NTP, resulting in the addition of an NMP to the transcript and the concomitant release of pyrophosphate (the energy of the reaction comes from the breakage of the NMP-pyrophosphate bond); and (iii) the RNAP translocates onto the next DNA position.

Once stable elongation is underway, the polymerization reaction occurs rapidly (at about 50-100 nucleotides per second (Roberts *et al.*, 2008)) and usually proceeds until termination. However, the RNAP is known to pause during synthesis depending, among other things, on sequence features and interactions with antiterminators. Antiterminators are proteins that interact with the RNAP at genetically-specific sites and allow it to bypass terminator sequences and inhibit elongation pausing. Transcription ends with termination, a process that may or may not require an RNA translocase (e.g. the rho terminator), depending on the gene or operon being transcribed.

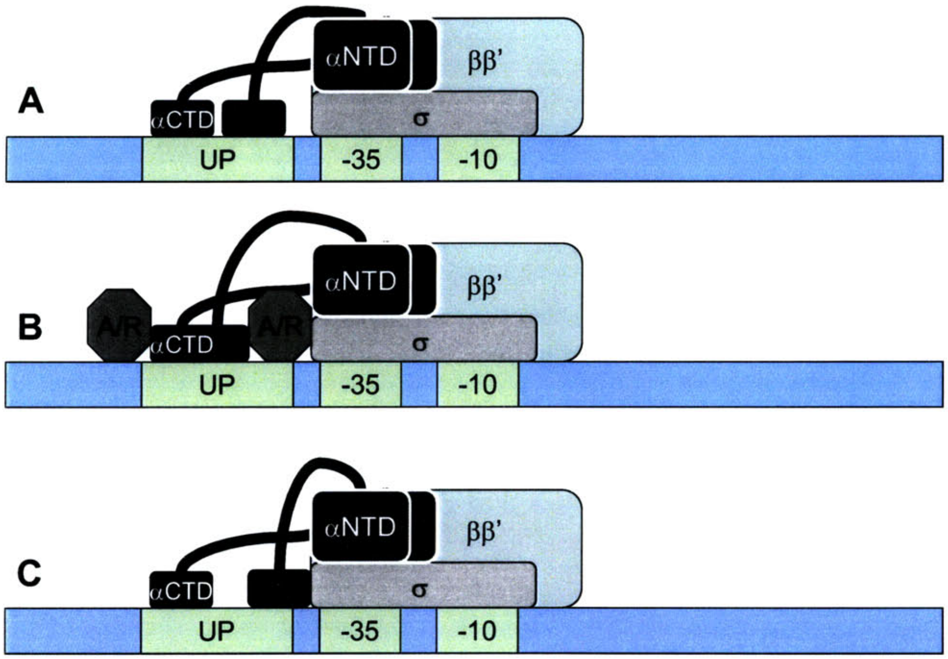

Figure 4.1-1. Types of interactions of the RNAP holoenzyme at the promoter

The canonical promoter regions and subunits are indicated (the amino- and carboxy-terminal domains of the α-subunit are indicated as NTD or CTD, respectively), A/R indicates an effector, i.e., an activator or repressor. (A) Simple promoter, no activators or repressors present. (B) The positions at which effectors may bind and their interactions are indicated. Most effector-binding promoters do not have both effectors as shown (although this is possible (Busby & Ebright, 1994)); instead, they have one or the other and are categorized as Class I or II depending on the location (see text). (C) Interaction between the αCTD and the σ-subunit is indicated. Figure adapted from (Busby & Ebright, 1994).

This oversimplification of the process presents the sequence of events that conjunctively lead to the transcriptome, but does not suggest routes for its manipulation. In order to do that, we must understand how the cell uses each of these steps, and transcription initiation in particular, to regulate its physiology. This is the subject of the next section.

4.1.2 *Transcription Regulation*

Plainly, phenotypic regulation through transcription is the result of differential gene expression, since it is largely the relative level of proteins in the cell which determines its physiological response to the environment. Such differential expression comes about mainly from the integration of signals at the interface between the RNAP holoenzyme and the promoter. For reasons of cellular economy, it makes sense to integrate most of these signals prior to or at transcription initiation, that is, prior to open complex formation. Because RNAP is in limiting amounts, not all operons can be simultaneously transcribed; instead, different sites "compete" for RNAP complexes with varying degrees of efficacy, giving rise to the differential expression alluded to earlier (Ishihama, 2000). A recent review focused on the five molecular mechanisms that are responsible for the differential distribution of RNAPs throughout the genome: promoters, effectors (activators or repressors), sigma-factors, ligands, and local DNA structure (Browning & Busby, 2004).

The promoter signals interact with the RNAP holoenzyme in different ways (**Figure 4.1-1**): the -10 hexamer (located at or near the -10 position with respect to the transcription start site), the -10 extended site, and the -35 hexamer all bind sigma D or similar factors. In addition, the UP promoter element binds the αCTD. At some loci, transcription factors or effectors stimulate or inhibit transcription by interacting with either the alpha- or sigma-subunits of RNAP. At class I promoters, the effector binds upstream of the -35 hexamer and near the UP promoter site and interacts with αCTD, while at class II promoters, the binding site overlaps with the -35 hexamer and the effector interacts with sigma and one or both alpha domains (Browning & Busby, 2004; Niu *et al.*, 1996; Savery *et al.*, 2002). As will be discussed later in sections 4.2 and 4.3, the sigma factor that is bound to the RNAP core enzyme determines much of its specificity, and thus is a key factor for differential gene expression.

Small molecule ligands can also affect the affinity of the RNAP holoenzyme at some promoters, especially under certain environmental conditions; for example, the ligand ppGpp (guanosine tetraphosphate) can bind the RNA polymerase and either modulate its interaction with promoters or the competition between sigma factors (see Section 4.3) (Jishage *et al.*, 2002; Paul *et al.*, 2004). Other small molecules can operate indirectly through the action of transcription factors, e.g., by allosterically stimulating or inhibiting DNA-binding or affecting their overall conformation (Botsford & Harman, 1992; Dalbow & Bremer, 1975). Finally, some regulation is effected by the local DNA-structure, although this mechanism is only partly understood.

In general, the signals that are distributed throughout the cell or that are non-specific are hard to exploit for engineering complex phenotypes. From the remaining list, the subunits of the RNAP holoenzyme are good examples of possible targets. We now turn our attention to describing them and their use for strain improvement.

4.2 Principal Sigma Factor (sigma D)

Bacterial genomes may encode for one (e.g. *Mycoplasma genitalium*) or many (e.g. *Bacillus subtilis*) sigma factors (Browning & Busby, 2004). Regardless of the actual number, there is one principal or housekeeping sigma which can transcribe at most promoters when bound to the core RNAP, in particular under exponential growth or non-stressful conditions. For *E. coli*, as it has been previously described, this housekeeping factor is referred to as sigma D and is coded by the gene *rpoD*; the same is true for another species that will be considered here, *L. plantarum*.

Sigma D belongs to a category of bacterial sigma factors known as the sigma 70 family, named after the molecular weight of the *E. coli* housekeeping factor; members of this family share many sequence and structural features (Paget & Helmann, 2003). Homology analysis across the sigma 70 family reveals four regions, which can be divided further in sub-regions.

Region 1 is poorly conserved, but in some species is known to have DNA-like binding properties, such that it functions as a protective flap for regions 2 and 4 in the free (not RNAP-bound) factor (Dombroski *et al.*, 1992). Region 2, and sub-region 2.4 in particular, is highly conserved, consisting in a series of α-helices responsible for recognizing the -10 promoter hexamer, binding the core RNAP, and melting the promoter at the open complex (see Section 4.1) (Campbell *et al.*, 2002). Region 3 is, similarly to region 1, poorly conserved, and is sometimes even absent altogether; when present, it is responsible for binding the -10 extended sequence at some promoters (Paget & Helmann, 2003). Finally, region 4 is highly conserved and has two pairs of α-helices, it contacts the

-35 hexamer, and the αCTD or effectors at many loci (see below) (Campbell *et al.*, 2002; Dove *et al.*, 2003).

Sigma D, and that of *E. coli* in particular, is known to regulate transcription not only through direct protein-DNA interactions, but also through protein-protein contacts (this fact was mentioned earlier, but rather briefly). For example, work with the bacteriophage λ cI protein (λcI) and *E. coli* sigma D has shown a direct activating interaction between E34 in λcI and R588 in sigma region 4, and one more between D38 in λcI and R596 in the same region (Li *et al.*, 1994; Nickels *et al.*, 2002). Other studies have shown direct contacts involved in transcription regulation between the αCTD and sigma region 4.2 (Ross *et al.*, 2003).

The many structural features of sigma factors involved in RNAP-promoter interaction suggest that these proteins are central to transcriptional regulation. Indeed, several groups who have studied the organization of transcriptional networks in *E. coli* place sigma factors high in the hierarchy (Ma *et al.*, 2004; Martinez-Antonio *et al.*, 2008).

In one way or another, it is the regulatory functions which we try to exploit during transcriptional engineering, although it is not obvious where (i.e. in which region(s)) mutagenesis would have the greatest impact in phenotype. During the first set of experiments of this thesis, we assumed that non-directed mutagenesis of the entire sigma D protein could be used for finding improved mutants. As we see in the next few sections, this assumption turned out to be correct, at least to a first approximation.

4.2.1 *Use of sigma D for Phenotypic Improvement in E. coli*

Previously, in Section 3.3.3, we discussed the use of sigma D for improving ethanol tolerance, lycopene production, and resistance to various stresses (referring to previous work in our laboratory (Alper & Stephanopoulos, 2007)), proving the use of this target for transcriptional engineering. An initial project of the thesis, and one that brought further proof that sigma D was a good target for phenotypic engineering, was in the application of a high-throughput screening method for isolating strains with enhanced hyaluronic acid production.

4.2.1.1 Hyaluronic Acid Production

Hyaluronic acid (Hyaluronan, HA) is a valuable functional biopolymer, its importance stemming from its structural, rheological, physiological, and biological properties. Similar to other biomaterials with comparable attributes, it possesses a wide range of applications in the health, cosmetic and clinical fields (Goa & Benfield, 1994; Lauren, 1998). Microbial production of HA using the group C or group A *Streptococcus* has been pursued as an alternative to chemical extraction from chicken combs (Kim *et al.*, 1996; Ogrodowski *et al.*, 2005). Recently, new methods using novel recombinant production strains that utilize inexpensive media and avoid pathogenicity have been developed as substitute of *Streptococcus* fermentation. The construction of a recombinant gram-positive *Bacillus subtilis* has been reported (Widner *et al.*, 2005), and more recently, our group developed a recombinant *E. coli*, Top10/pMBAD-*sseABC* to produce HA (Yu & Stephanopoulos, 2008).

The above methods applied traditional metabolic engineering (Section 2.2) in constructing the HA producing strains (Stephanopoulos & Sinskey, 1993;

Stephanopoulos & Kelleher, 2001), addressing issues of precursor supply and pathway kinetics and regulation (Stephanopoulos & Simpson, 1997; Vallino & Stephanopoulos, 1994). While such methods have yielded nontrivial improvements in many instances, especially when combined with bioreactor operation and optimization strategies (Follstad *et al.*, 1999; Kiss & Stephanopoulos, 1991; San & Stephanopoulos, 1984), recombinant synthesis of HA is limited by many factors.

For example, intermediates in the pathway (such as glucose-6-phosphate and N-acetyl-glucosamine) are also needed for important cellular functions, such as glycolysis and cell wall synthesis; thus HA production directly competes with cell growth and viability (Yu & Stephanopoulos, 2008). In addition, HA is highly viscous and may interfere with the transport of nutrients and gases, imposing an additional stress on the cell (Widner *et al.*, 2005). Thus, HA productivity is likely limited by several factors and random search-based strain improvement approaches become ideally suited for increasing HA production. A key element to the implementation of such search strategies is the availability of high throughput screens for isolating cells capable of high product accumulation. Therefore, our laboratory developed and tested a screen for the case of HA production in *E. coli.*

The two-step screen consisted of translucent colony identification followed by alcian blue staining. High HA-producing colonies are viscous and appear clearer compared to low HA-producers (Kim *et al.*, 1996; Widner *et al.*, 2005). Colony morphology is only adequate to discern large differences in HA productivity, so a subsequent, more quantitative approach was necessary to measure incremental improvements in HA accumulation.

Alcian blue is a water soluble copper-phthalocyanine dye, $C_{56}H_{68}C_{l4}CuN_{16}S_4$, which can be used for the staining of sulfated and carboxylated acid mucopolysaccharides (Penney *et al.*, 2002). It is believed to form salt linkages with the acid groups of mucopolysaccharides due to the presence of copper in the molecule, which decreases the blue color. Since HA is a mucopolysaccharide, it was feasible to establish an alcian blue staining method for quantifying the HA concentration.

The above screen was tested in an HA-producing recombinant *E. coli* host transformed with libraries of the *rpoD* gene. The gene was first cloned into a low-copy number vector (with a pSC101 origin of replication), and then amplified with epPCR with three different mutation frequencies (described in (Alper & Stephanopoulos, 2007)). The two-step high throughput screening was then applied to the resulting libraries.

Using the first identification step (translucency), 74 *rpoD* mutants were selected from thousands of colonies on solid plates, and subsequently tested for HA accumulation by the alcian blue method (sigma S libraries were screened simultaneously, but the sigma S results shown in **Figure 4.2-1** will not be discussed in detail until Section 4.3.1). The parental strain carrying only the plasmid for HA synthesis (Top10/pMBAD-*sseABC*) was cultured in parallel and used as a control. The selection results of both libraries are plotted in **Figure 4.2-1**, showing several improved mutants; for example, the D72 strain showed a significant increase of HA concentration relative to the control (100% line).

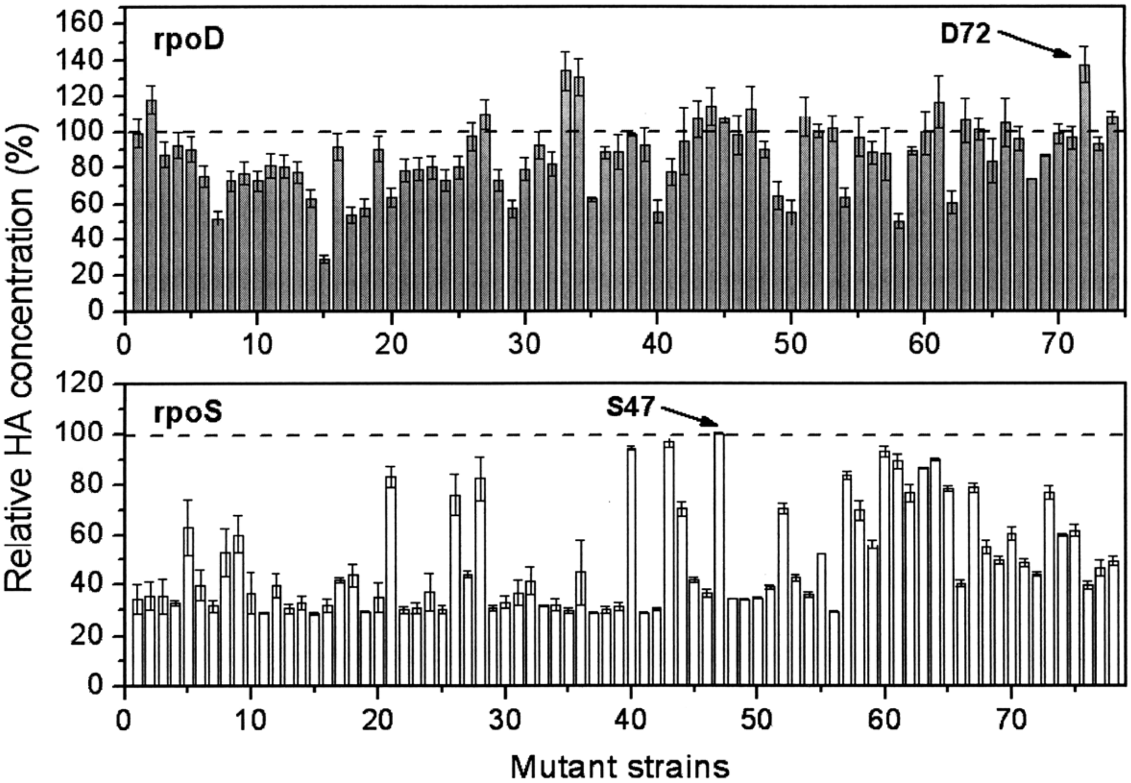

Figure 4.2-1. Screening of recombinant sigma D and sigma S libraries for hyaluronic acid accumulation

The relative HA concentration was determined using alcian blue quantification. Control strain (dashed line, 100%) is a Top10 /pMBAD-*sseABC*, and the mutants that were further studied are indicated with an arrow. All samples were measured in duplicate.

The most promising mutants obtained from the primary screening described above were further studied in shake flask cultures; the culture volume was scaled to 40 ml medium in a 250 ml flask. The results from these experiments are shown in Table 3, found in Section 4.3.1. From them, we can conclude that, compared to an isogenic, but unmutated control (D0), mutant D72 shows an improvement in both HA titer and productivity, evidencing the value of sigma D for engineering this phenotype. Other conclusions are found in the aforementioned Section.

Other phenotypes were improved using sigma D during the course of this thesis, although using a different library design than that described for HA. The details will be found in Section 6.6.3, in which we describe the use of optimized sigma D libraries for improving the survivability of an ethanologenic *E. coli* strain to the combined stress of overlimed hydrolysate and ethanol.

4.2.2 *Use of sigma D for Phenotypic Improvement in L. plantarum*

The use of sigma D libraries in *E. coli* proved that transcriptional engineering cannot only be effected by zinc finger protein-based artificial transcription factors (Section 3.3.2), but with native factors as well. However, it was difficult to generalize the use of the principal sigma factor as a target for transcriptional engineering without presenting evidence from other species. We thus turned our attention to *L. plantarum*, a member of the family of lactic acid bacteria. This family of microorganisms is routinely employed in the industrial production of precisely that organic compound, and, since it is the interest in this chemical what motivated our study, let us delve more into the details of its manufacture.

Lactic acid is a 3-carbon organic acid with a pK_a value of 3.85, commercially produced mostly by fermentation at a scale of about 150,000 tonnes per year (in 2007) (Sauer *et al.*, 2008). Its uses span a wide range of applications: (i) as a preservative for food, pharmaceuticals, cosmetics, textiles and leather; (ii) as a flavor additive; (iii) as a chemical feedstock or intermediate for other synthesis processes; and (iv) in the production of polylactate, a valuable, biodegradable plastic (Hofvendahl & Hahn-Hagerdal, 2000).

In *L. plantarum*, categorized as a facultative heterofermentative strain, lactic acid can be produced either by the Embden-Meyerhof-Parnas (EMP) or the phosphoketolase pathways, resulting in a homolactic or heterolactic product mixture, respectively (Kleerebezem *et al.*, 2003). In the EMP pathway, two pyruvate molecules are formed during glycolysis, both of which are converted into lactic acid by the enzyme lactate dehydrogenase. In the phosphoketolase pathway, the hexose (e.g. glucose) is broken into CO_2, pyruvate, and acetyl-phosphate (acetyl-P); pyruvate is then converted into lactate (as in the EMP), and acetyl-P into either acetate or ethanol (Kleerebezem *et al.*, 2003; Murphy *et al.*, 1985).

The process characteristics of most interest, i.e. lactic acid yield, titers, and productivities, are highly dependent on choices such as the strain, carbon and nitrogen sources, temperature, pH, fermentation mode, among others (Hofvendahl & Hahn-Hagerdal, 2000). One of the key control parameters is the pH: fermentations that are run at low pH (at or below the pK_a) are more economical than those at high pH, since the free form of the acid can be readily separated using organic extraction and the salt form is expensive and cumbersome to process (Patnaik *et al.*, 2002). This fact, along with the very high titers of lactic acid achievable in fermentations (>150 g/L, (Hofvendahl & Hahn-Hagerdal, 2000)), implies that an industrial strain must perform well at low pH and high lactic acid concentrations. Unfortunately, the combination of these stresses presents a particularly big challenge.

Lactic and other organic acids are thought to hinder growth via different mechanisms. Toxicity is pronounced at low pH because this condition favors the protonated, uncharged form of the acid that can be transported freely through the membrane. In the cytosol, the acid dissociates following the Henderson-Hasselbach equilibrium, lowering the cytosolic pH and increasing the concentration of the anionic species. This sequence of events results, first, in the partial dissipation of the proton gradient across the cell membrane, which leads to energetic inefficiencies (McDonald *et al.*, 1990); second, in the buildup of protons in the cytoplasm that impact negatively many biochemical processes (Booth, 1985; Kresnowati *et al.*, 2007); and third, in the intracellular accumulation of the anion and accompanying end-product inhibition (Pieterse *et al.*, 2005). This type of inhibition refers to the inability of *L. plantarum* to re-generate NAD^+ effectively as pyruvate accumulates in the cell at high lactic acid conditions (Pieterse *et al.*, 2005). As a consequence of these effects, acidification of the media with inorganic acids presents a milder challenge compared to organic acids (**Figure 4.2-2**).

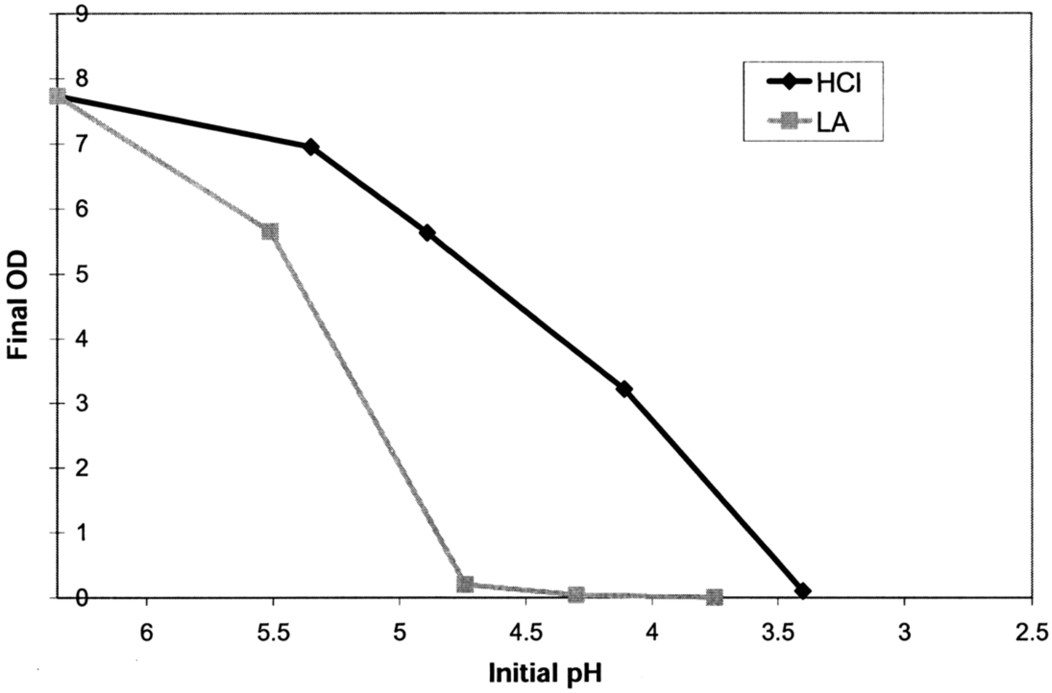

Figure 4.2-2. Test for organic and inorganic acid toxicity in *Lactobacillus*

The graph shows final growth of wild-type *L. plantarum* cells (OD, optical density at 600 nm) at different initial pH values in media acidified with either hydrochloric acid (HCl) or lactic acid (LA).

This list of multiple physical and chemical effects translates in a similarly-lengthy list of cellular responses. Several studies have suggested mechanisms or genes that are involved in coping with low pH and high lactic acid conditions. For example, insertional mutants of four different genes purportedly involved in decarboxylation reactions known to aid during acid shock, were constructed and their ability to survive was tested at low pH. The mutants were more sensitive to acid in some conditions, but not in others, showing that the mechanisms behind tolerance are distributed and complex (Azcarate-Peril *et al.*, 2004). In another study, insertional mutagenesis was used to find 18 new loci that affect the ability of *Lactococcus lactis*, a related lactic acid bacteria, to cope with acid stress (Rallu *et al.*, 2000). Detailed transcriptional analysis in the presence of various conditions faced by the cells during lactic acid fermentations also showed a genome-wide response (Pieterse *et al.*, 2005). Because we aim at engineering tolerance to various, simultaneous stresses (i.e. low pH, high lactic acid, high osmolarity, etc.) that elicit a complex, poorly-understood response, random search-based evolutionary approaches were deemed an ideal choice.

4.2.2.1 Library Construction

In comparison to *E. coli*, *Lactobacillus spp.* are less common laboratory strains, and are harder to engineer due to a relative scarcity of tools and experimental protocols. Of the requirements for an appropriate host selection listed in Section 2.2.1, high transformation efficiencies are particularly essential for constructing sigma factor libraries. We began this study with a *L. casei* strain (ATCC393), given that this species is known for its desirable lactic acid production characteristics (Hofvendahl & Hahn-Hagerdal, 2000). Unfortunately, the transformation protocol consistently failed, even after modifying the conditions and vectors used. Eventually, it was realized that the available type strain is physiologically different than a homologous *L. casei* strain that has been known to be transformable with reasonable efficiencies (Prof. Gaspar Perez-Martinez (CSIC, Valencia, Spain), personal communication; and (Acedo-Felix & Perez-Martinez, 2003)). We thus switched to *L. plantarum*, which was successfully transformed with a chloramphenicol and erythromycin resistance-carrying plasmid, pGK12. This plasmid, which contains a pWVO1 origin of replication, can propagate in both *L. plantarum* and *E. coli* (Kok *et al.*, 1984; Lin & Chung, 1999), although it does not contain a multiple cloning site (MCS).

After inserting an MCS into pGK12 to form pDK12, the gene for the principal sigma factor, *rpoD*, was cloned and expressed from its native promoter. The wild-type version of this plasmid was named pDK12D (**Figure 9.2-1**); from it, three libraries with different mutation frequencies were constructed via epPCR, and transformed into electrocompetent *L. plantarum.* Different mutation frequencies (low, medium, and high) were achieved by varying the amount of template DNA in the reaction. The fraction of mutated-DNA copies and the average number of mutations per copy decrease with increasing amounts of template DNA, as each template strand is replicated fewer times.

4.2.2.2 Mutants with Improved Tolerance to Lactic Acid and Low pH

The resulting transcriptional engineering libraries were screened for improved phenotypes in industrially-relevant conditions. Since low pH fermentations are characterized by high concentrations of free lactic acid and protons, both conditions were explored. We challenged these libraries either in 5.5 g/L of L-lactate at an initial pH=4.6 (LA condition) or at an initial pH of 3.85 adjusted with inorganic acid (HCl condition). The LA condition addresses the end-product stress directly, while the HCl condition does so indirectly (i.e. only as the cells produce lactate). Individual clones were selected after three rounds of subculturing and the plasmids carrying the mutant sigma factors isolated. The latter were sequenced and retransformed by electroporation into cells with a clean genetic background to ensure that the improved phenotype did not arise due to spontaneous mutation of the chromosomal DNA. After confirming the phenotype, the best clones, (mutant S6 under the LA condition and mutant H13 under the HCl condition), were selected for more detailed analysis. **Figure 4.2-3** shows the growth profiles of the retransformed mutants and control under the same stresses used for selection. Mutant S6 grows about 3.5-fold faster and up to a 5-fold higher OD than the control in the LA condition. Mutant H13 reaches 86.4% higher OD and a 25% higher growth rate than the control when grown in the HCl condition.

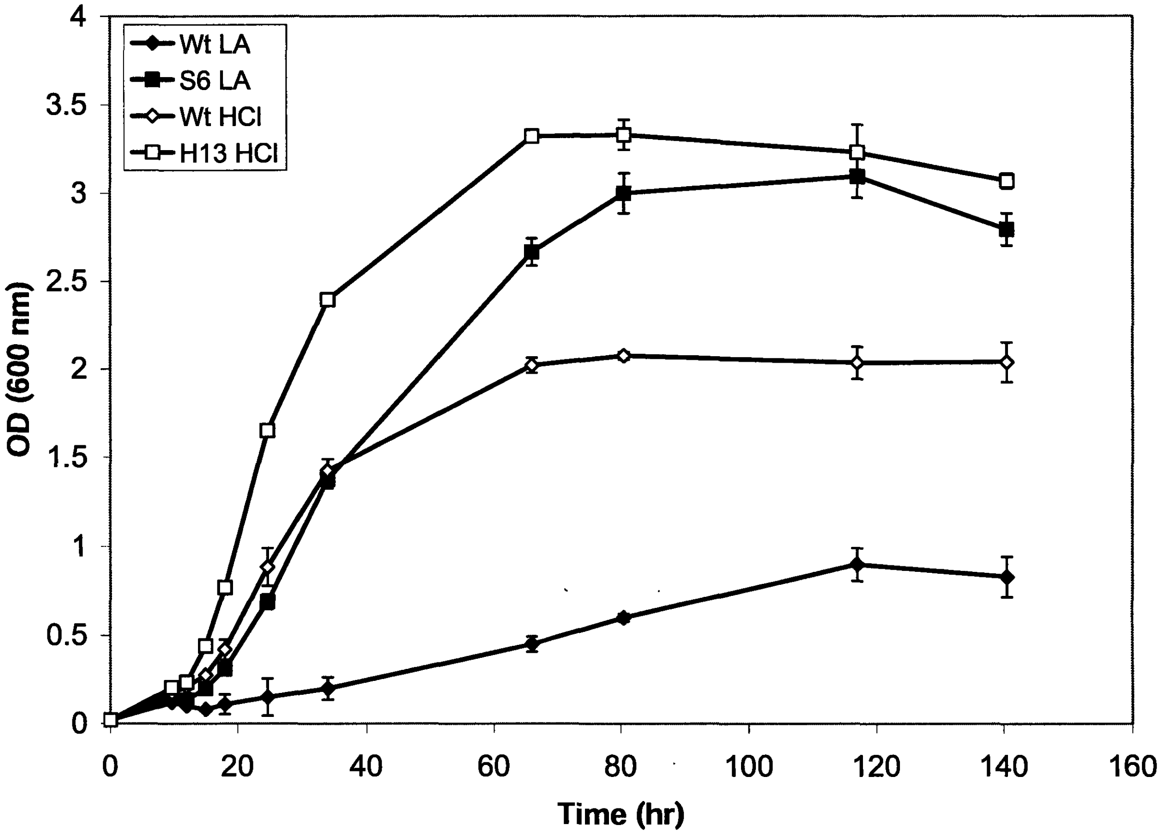

Figure 4.2-3. Growth profile of retransformed strains in different stresses

The mutants bear sigma factor variants S6 and H13 and are compared to a strain bearing the control plasmid (pDK12D). The curves show growth for each mutant in the media used for selection, i.e., LA condition (closed symbols) for S6 and HCl condition (open symbols) for H13. The wild-type (Wt) is shown in both conditions for comparison.

The mutants were also tested under stresses not used for selection, i.e., mutant S6 was grown in the HCl condition and H13 in the LA condition. Acidification of the media with inorganic acid causes a different transcriptomic response than when lactate is added to the media (Pieterse *et al.*, 2005), although these stresses are inseparable during fermentation. The production of lactic acid is accompanied by acidification of the media, while low pH increases the amount of free lactic acid that may enter the cell and effect toxicity (Giraud *et al.*, 1991; McDonald *et al.*, 1990). Therefore, it is possible that the transcriptome that protects H13 at low pH (HCl condition) also protects it against lactic acid (LA condition). Conversely, the same may be true of mutant S6 in the HCl condition.

To explore this transcriptomic overlap, we tested the mutants and control in both conditions (Table 1). Mutant H13 exhibits improved growth at low pH adjusted with both inorganic and lactic acids, but mutant S6 does not when inorganic acid is used. Most likely, the mutants cope with these stresses differently and the underlying mechanisms result in a convergent phenotype in lactic acid. We tried to exploit both mechanisms by co-expressing the mutant sigma factors in the same cell, as prior work suggested that improvements in phenotype conferred by different sigma factors may be additive (Alper & Stephanopoulos, 2007). The phenotype of the combined mutant in both LA and HCl was similar to that of H13, suggesting that the mechanism of action of this sigma factor is dominant over that of S6 and wild-type (Table 1).

Table 1. Growth characteristics for Lactobacillus mutants in different stresses

Growth rate (μ) and stationary phase optical density (600nm) of the mutants and control under the experimental conditions.

	LA condition		HCl condition	
Strain	**μ (hr^{-1})**	**Stat. OD**	**μ (hr^{-1})**	**Stat. OD**
S6	0.101 ± 0.001	3.0 ± 0.1	0.08 ± 0.01	1.19 ± 0.01
H13	0.057 ± 0.004	3.2 ± 0.2	0.151 ± 0.001	3.33 ± 0.08
Wild-type	0.028 ± 0.005	0.9 ± 0.2	0.12 ± 0.01	2.08 ± 0.02
S6-H13	0.058 ± 0.004	2.9 ± 0.2	0.152 ± 0.002	3.02 ± 0.01

4.2.2.3 Sequence Analysis

We also studied the sequences of the mutant factors that gave rise to the observed characteristics. Mutant S6 has a single nonsynonymous substitution (Q345K). This mutation was responsible for the increased growth in high lactic acid, the higher specific productivity of lactate (between 40 and 60% with respect to control), sensitivity to HCl (Table 1), tolerance to higher salt concentrations (qualitative observation), and probably other traits that remain uncharacterized. The pleiotropic nature of the mutation suggests that it changes the intracellular environment globally. Glutamine 345 is located in a region that is highly conserved across sigma factors of many species involved in the recognition of and interaction with the -35 promoter box, as shown in **Figure 4.2-4** (Campbell *et al.*, 2002). This mutation most likely changes the relative affinity of the RNA polymerase (RNAP) holoenzyme for different promoter regions, similar to what has been previously observed in *E. coli* (Gardella *et al.*, 1989; Siegele *et al.*, 1989), which may result in a global response. As will be seen later in Section 6.6.2, the region of *E. coli* sigma D that contacts the -35 hexamer has overall a large potential for phenotypic alteration. Therefore, this mutation is in tune with later findings of this thesis.

```
TLEEVGKQFDVTRERIRQIEAK 593
TLEEVGKVFGVTRERIRQIEAK 349
******* *.************
```

Figure 4.2-4. Sequence alignment of *E. coli* and *L. plantarum* sigma D

The alignment corresponds to a highly conserved stretch of amino acids in region 4.2 of the sigma factor. The top sequence corresponds to the *E. coli* sequence, and the one below it to the *L. plantarum* equivalent. The asterisks indicate highly conserved residues. Amino acid Q345 is highlighted, marking the location of the mutation in variant S6 isolated in the LA condition.

Mutant H13 has several nonsynonymous substitutions (T44A, R74K, D114A, and S119A) and an insertion that results in a truncated sigma factor that includes all of region 1.1 and part of region 1.2 of the protein. Region 1.1 is relatively unconserved across species (see Section 4.1.1). Many bacterial sigma factors (like that of *E. coli*) have acidic N-termini, presumably to mimic the DNA strand and prevent nonspecific binding of the sigma subunit when not bound to the core RNAP (Dombroski *et al.*, 1992). Others (like that of the cyanobacterium *Thermosynechococcus elongates*) have basic regions that have been suggested to be involved in direct DNA binding (Imashimizu *et al.*, 2006). Given that taxonomic analyses suggest that gram-positives and cyanobacteria are sister groups (Gruber & Bryant, 1997), it is more probable that the *L. plantarum* sigma factor region

1.1 has the latter, rather than the former functionality. To further test this possibility, we analyzed the first 70 amino acids of *L. plantarum,* and found 12 basic residues, contrasting with 3 in the *E. coli* counterpart. This suggests that it is possible that region 1.1 of the H13 sigma subunit binds DNA and that this free form acts as a nonspecific repressor. The 3-D structure of the *Lactobacillus* sigma subunit has not been determined, which precluded us from doing a surface charge analysis to assess this possibility. For both mutants, a more complete explanation of the effect of the altered factors in the transcription process would require a multifaceted study and thus was beyond the scope of the present thesis.

4.2.2.4 Fermentations

Fermentations were carried out to determine the lactic acid productivity of H13 and control. This mutant was tested as it was tolerant to both HCl and LA stresses, whereas S6 was specifically tolerant to the LA condition (Table 1). MRS media was either supplemented with glucose (no stress) or the initial pH was adjusted to 3.85 with no added glucose (HCl condition). Under no stress, H13 and wild-type had similar lactic acid titers. At an initial pH of 3.85, mutant H13 grew better and produced more lactic acid **(Figure 4.2-5**).

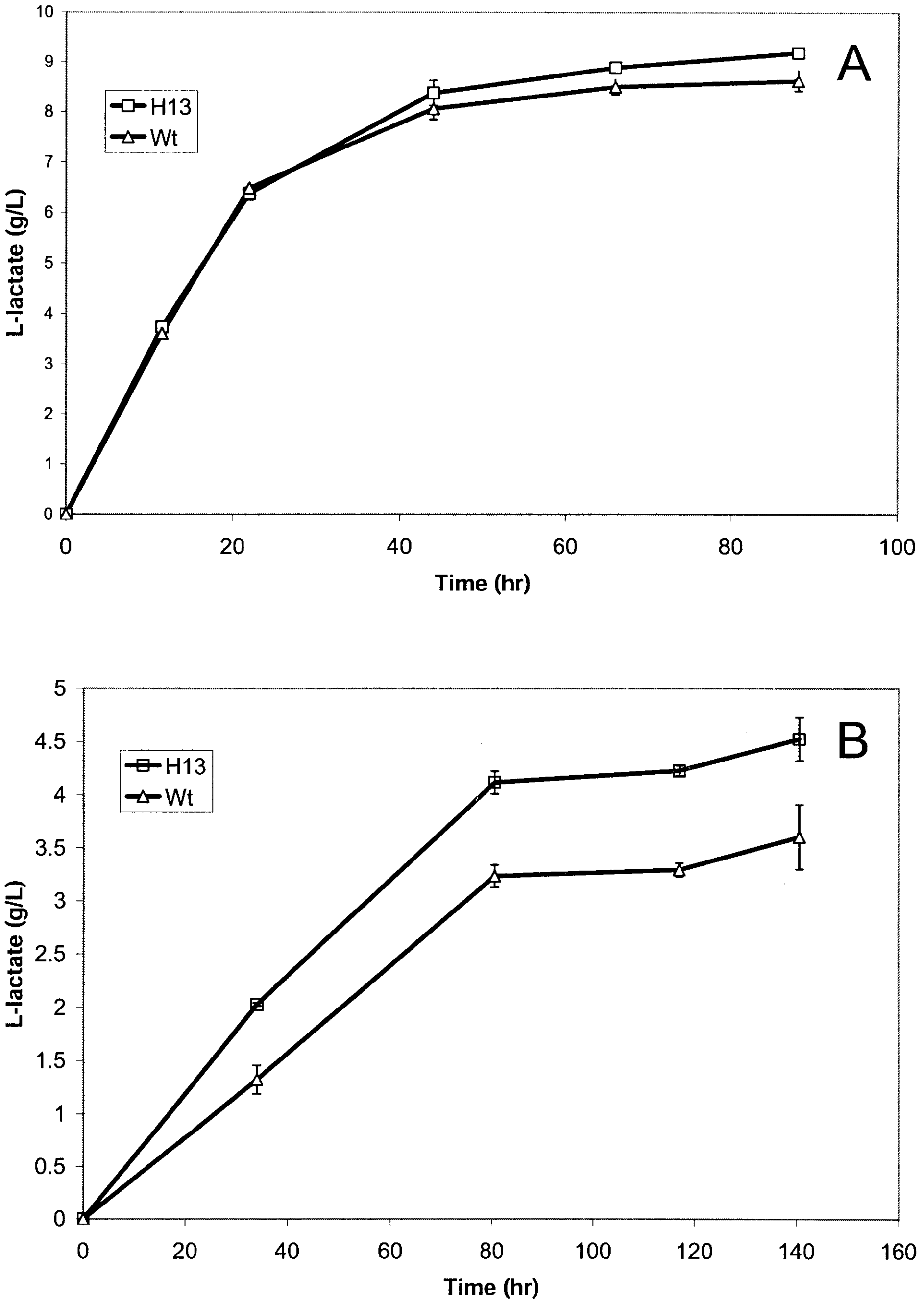

Figure 4.2-5. Fermentation experiments

The graphs show shake-flask fermentation of L-lactate by the H13 mutant and wild-type (Wt) in media supplemented with glucose and unadjusted pH (no stress condition, (A)), or in the HCl condition (initial pH=3.85 ± 0.05, (B)).

4.3 Alternative Sigma Factors

Up to date, seven sigma factors have been identified in *E. coli*, all of which belong to one of two protein families. The previously-mentioned sigma 70 family, named after the housekeeping sigma in *E. coli*, includes the sigma factors D, S, F, E, H, and fecI. The second family, sigma 54, is named after and includes only sigma N, which has little sequence homology with the rest of the factors (Burgess & Anthony, 2001; Gruber & Gross, 2003). Table 2 summarizes the activities coordinated by each of the seven sigma factors (Braun *et al.*, 2006; Mytelka & Chamberlin, 1996; Nystrom, 2004; Reitzer, 2003).

In exponentially growing cells, the RNAP is almost exclusively bound to sigma D, which orchestrates the transcription of genes related to proliferation (see Section 4.2). The Eσ^D complex (core RNAP bound to sigma D) is responsible for transcribing genes involved in DNA replication, cell membrane and ribosome biosynthesis, substrate uptake, etc. (Mooney *et al.*, 2005; Nystrom, 2004). On the other hand, the alternative sigma factors respond to specific stresses, environmental changes, or other growth-limiting conditions (Nystrom, 2004; Vijayakumar *et al.*, 2004). In *E. coli*, sigma S regulates the most number of genes after sigma D; it is induced during common stress responses such as starvation, and coordinates the stationary phase phenotype. The Eσ^S complex directs transcription of genes needed for nutrient scavenging, DNA repair, protein turnover, etc. (Vijayakumar *et al.*, 2004).

Other sigma factors are in charge of sensing the environment in different compartments in the cell. For example, sigmas E and H respond to misfolded proteins in the periplasm and cytoplasm respectively (Ruiz & Silhavy, 2005; Zhao *et al.*, 2005); they react to stresses that denature proteins, such as heat and sub-lethal doses of ethanol (see below).

Table 2. ***E. coli*** **sigma factors**

Sigma factor	Function
D	DNA replication, substrate uptake, membrane synthesis, etc.
S	Stress response, DNA repair, nutrient scavenging, etc.
H	Heat-shock response, cytoplasmic sensor for denatured proteins
F	Flagellum synthesis and cell motility
E	Envelope/periplasm sensor for denatured proteins
fecI	Ferric citrate uptake
N	Nitrogen assimilation, anaerobic or nutrient-limiting growth

Since the cell faces limiting resources, it cannot devote the same energy to both reproduction and survival. Therefore, there is an inherent tradeoff between expressing genes that promote rapid growth and those that protect against stressful conditions. It has been argued that such tradeoff originates in the competition of different sigmas (mainly sigma D and S) for unbound core polymerase (Fang, 2005; Ferenci, 2003). Recall from Section 4.1.2 that RNAP is present in limiting amounts in the cell, so that not all operons can be transcribed simultaneously. Therefore, the cell must regulate the expression and stability of alternative sigma factors, as well as the process of sigma factor competition for free RNAP. Furthermore, the cell should have a mechanism for tailoring the preference of each sigma to its regulon, and, interestingly, this mechanism is not always entirely related to promoter sequence.

The levels of sigma S are mostly regulated post-transcriptionally. A comprehensive review of the mechanisms here enlisted can be found in a recent publication (Hengge-

Aronis, 2002). Translation of the sigma S mRNA is highly controlled, through a wide variety of regulators such as the histone-like protein HN-S, the RNA chaperone Hfq, and several small noncoding RNAs. The level of sigma S protein is further modulated by the action of proteases such as ClpXP.

Because sigma S is very closely related to sigma D, they bind to nearly-identical consensus promoter sequences *in vitro* (Gruber & Gross, 2003). *In vivo*, however, sigma S is known to have a measurably different regulon compared to sigma D (Weber *et al.*, 2005), and a considerable amount of research has gone into finding the cues needed for altering the selectivity inside the cell. Small differences in sequence seem to provide some discriminatory basis for sigma S- and sigma D-bound RNAP (Becker & Hengge-Aronis, 2001). Other interactions – such as those with global effectors – and other features – such as local DNA supercoiling – have also been proposed to affect this selectivity (Typas & Hengge, 2006).

Small ligands have too been implicated in the general competition between sigma D and alternative sigmas. For example, the molecule ppGpp (guanosine tetraphosphate) accumulates during amino acid starvation conditions and binds to RNAP, which both increases sigma S synthesis and eases the formation of the Eσ^S complex (Jishage *et al.*, 2002; Magnusson *et al.*, 2005; Nystrom, 2004). Potassium glutamate is another example of a small ligand that changes the transcription profile based on sigma factor competition (Lee & Gralla, 2004).

Although studied in less detail, the regulation of other sigma factors of interest to this thesis (sigmas E and H) has also been explored. Sigma E, which directs expression of a regulon that protects against misfolded periplasmic proteins is regulated posttranslationally through a sigma/anti-sigma mechanism (Ades, 2004; Ruiz & Silhavy, 2005). This system communicates information about envelope stresses to the cytoplasm, where the response initiates at the transcriptional level. RseA is a membrane-bound protein that, under normal conditions, sequesters sigma E (sigma E binds to RseA with 300-fold greater affinity than to RNAP) (Ades, 2004). The presence of misfolded proteins in the periplasm activates a proteolytic cascade which culminates with the cleavage of RseA, which in turn frees sigma E. The sigma factor then binds RNAP and coordinates the stress response (Ades, 2004; Ruiz & Silhavy, 2005). The response includes the expression of chaperones and proteases that are targeted to the periplasmic space (Ades, 2004). This system is also involved in tolerance to less common envelope stresses, such as high metal concentration (Egler *et al.*, 2005).

Similar to sigma S, the level of sigma H is also regulated post-transcriptionally. This sigma protects against misfolded or aggregated cytoplasmic proteins, and its regulon includes a set of chaperones, proteases, and other ancillary proteins (Yura & Nakahigashi, 1999). Its mRNA forms a series of secondary structures that hinder translation at low temperatures, but that unfold during heat shock. The structures sequester the ribosome-binding site (RBS), which becomes accessible when the base pairings melt (Schlax & Worhunsky, 2003); this allows the mRNA to function as an *in vivo* thermometer. In addition, the sigma H protein is unstable at normal physiological conditions, being rapidly degraded with the aid of the DnaK-DnaJ-GrpE chaperones; at higher temperatures, these are recruited elsewhere, stabilizing sigma H and causing its accumulation (Yura & Nakahigashi, 1999).

It can be concluded from this discussion that the activity of one sigma factor is not independent of the activity of another, and that more than one factor could be targeted for transcriptional engineering. This was a particularly enticing prospect since, based on their function, the relative control of alternative sigmas over the transcriptome should strengthen during stress. That is, these factors should be more influential in the precise conditions of interest for a sizable fraction of strain improvement programs. We now describe some experiments and results regarding the use of sigma S, E and H for improving the environmental tolerance of *E. coli* in different conditions.

4.3.1 *Use of sigma S for Phenotypic Improvement in E. coli*

The general stress response sigma factor, sigma S, was deemed a good target for transcriptional engineering given it is responsible for controlling key genes for coping with environmental challenges (Weber *et al.*, 2005). Sigma S libraries were constructed in a similar way to those of sigma D (see Section 4.2.1 and (Alper & Stephanopoulos, 2007)). Briefly, the wild-type *rpoS* gene was amplified from genomic DNA and cloned in the pHACM plasmid, carrying the pSC101 origin of replication and an approximate copy-number of five per cell. The wild-type copy was amplified in an epPCR protocol using three mutation frequencies and cloned into the same vector. These libraries were used to transform different hosts, depending on the application.

4.3.1.1 Hyaluronic Acid Production

Simultaneously to screening the sigma D libraries for the improvement of hyaluronic acid (HA) production, similarly-constructed sigma S libraries transformed into Top10/

pMBAD-*sseABC* (Yu & Stephanopoulos, 2008) were also screened. A thorough description of the motivation for the study and the experimental protocols of the screen can be found in Section 4.2.1. A total of 78 sigma S mutants were isolated from the translucent colony identification step, and their HA titers were further confirmed using the alcian blue method. **Figure 4.2-1** in that Section shows the results and the performance of mutant S47, one of the few mutants that recovered the parental levels of production (most mutants exhibit a decrease in titer). Table 3 below summarizes the results from the shake-flask experiments carried out with the isolated mutants and controls.

Table 3. Hyaluronic acid production characteristics of isolated mutants

				Increase to C0 (%)		Increase to C1 (%)	
Strains	**DCW (g/L)**	**HA titer (mg/L)**	**C_{HA} = mg HA/g cell**[7]	**Titer**	**C_{HA}**	**Titer**	**C_{HA}**
C0[1]	2.06	404.8 ± 7.0	196.5 ± 3.4	—	—	—	—
C1[2]	2.21	509.8 ± 12.5	230.7 ± 5.6	—	—	—	—
D72[3]	2.12	561.4 ± 5.4	264.8 ± 2.5	38.7	34.8	10.1	14.8
S47[4]	2.11	479.0 ± 20.6	227.0 ± 9.8	18.3	15.5	-6	-1.6
D0[5]	2.11	425.0 ± 5.9	201.4 ± 2.8	5	2.5	-16.6	-12.7
S0[6]	2.91	695.6 ± 9.7	239.0 ± 3.3	71.8	21.6	36.4	3.6

[1] Parental strain Top10/pMBAD-sseABC with blank pHACM plasmid.

[2] Parental strain without additional plasmids.

[3] Sigma D mutant, see **Figure 4.2-1**.

[4] Sigma S mutant, see **Figure 4.2-1**.

[5] Parental strain with a wild-type rpoD gene in pHACM.

[6] Parental strain with a wild-type rpoS gene in pHACM.

[7] Specific productivity of HA in mg of product per mg of cell mass.

The data displayed in Table 3 shows a few interesting trends. For instance, the highest titers are achieved in a strain that overexpresses the wild-type version of the *rpoS* gene

(S0). Perhaps this overexpression, coupled with the overall stress conditions that are intrinsic to HA fermentations (see Section 4.2.1), favors competition of sigma S for free RNAP and results in cells that can better cope with high HA titers. Both mutants D72 and S47 show improvements compared to the C0 control, which is the HA producing host (C1) carrying a pHACM blank plasmid. Although these results show that the mutants improve the production compared to cells with similar plasmid burdens, this is not the case when contrasted with the pre-engineered strain, questioning the industrial relevance of these mutants. Regardless, this data supports the use of sigma S as an alternative target for transcriptional engineering, albeit in a qualitative fashion.

4.3.1.2 Carbon Dioxide Tolerance

One of the earliest projects of this thesis attempted to solve the apparent inhibitory effect of carbon dioxide in microorganism growth. Autogenous production of CO_2, summed to large hydrostatic pressures in bioreactors causes high local partial pressures of this gas, which leads to an overall decrease in growth and productivity (Onken & Liefke, 1989; Shang *et al.*, 2003). The most likely explanation is an end-product inhibition effect akin to that discussed for lactic acid in *L. plantarum* (see Section 4.2.2), coupled with a decrease in pH that accompanies the dissociation of carbonic acid in aqueous solutions (Lacoursiere *et al.*, 1986).

Preliminary studies done for developing carbon dioxide tolerance in *E. coli* suggested that selective pressures must be harsh enough if we intend to find outperforming members in the library. Available libraries were screened in the presence of high concentrations of carbon dioxide, as it was noted that the gas inhibits normal growth (stationary phase O.D. (600nm) of a wild-type strain was compared in flat and CO_2-rich medium and a ten-fold difference in cell density was observed). However, screening of sigma S (and, to a lesser extent, sigma D) libraries in the CO_2-rich medium was ineffective for isolating outperforming mutants. The ability of cells to grow in CO_2-rich medium improved with each round of sub-culturing, but it was not possible to isolate a clone that could still outgrow the control after re-transformation of the sigma factor-containing plasmid. Cells seemed to adapt to high CO_2 concentrations at least as fast as potentially improved variants grow, increasing the rate of false positives and negating the usefulness of the libraries. In fact, it has been found that CO_2 does not kill cells even when 100% CO_2 gas-flow is used (Lacoursiere *et al.*, 1986).

These results revealed an important fact: a cell with a mutated regulator must outgrow the rapidly-adapting background population by orders of magnitude, if the screening pressure does not kill the average and underperforming cells. The months of unfruitful efforts evidenced the known fact that purifying selection is a key time- and resource-intensive step in random searches for strain improvement (see Section 3.1.2). This was the first indication that not all phenotypes of interest can be easily improved.

4.3.2 *Use of sigma E and sigma H for Phenotypic Improvement in E. coli*

Early on, we recognized that ethanol fermentations represented an interesting application for our approaches. As will become apparent from the discussion below, sigma E and sigma H seemed ideal targets for improving environmental tolerance associated with ethanol fermentations.

In order to sample a larger sequence space, we considered building libraries in which both sigmas would be simultaneously mutated. As a first step towards this goal, we constructed an artificial operon expressing both *rpoE* and *rpoH* genes from a constitutive, synthetic promoter. The genes were amplified from genomic DNA and cloned using a triple ligation into a pZE vector downstream of the Q-variant of the $P_{LtetO-1}$ promoter (described in (Alper *et al.*, 2005)); this plasmid has a ColE1 origin of replication and contains the kanamycin-resistance gene (Lutz & Bujard, 1997). The correct wild-type sequences and operon structure was confirmed before amplification by epPCR was implemented as previously described. The mutant libraries were transformed into a DH10B *E. coli* cell line for further studies.

4.3.2.1 Heat and Ethanol Tolerance

In general, engineering environmental tolerance against stresses commonly encountered in ethanol and similar fermentations has been extensively explored inside and outside our laboratory, and for a variety of microbial systems (Abdel-Fattah *et al.*, 2000; Alper *et al.*, 2006; Cakar *et al.*, 2005; Demain *et al.*, 2005; Fiocco *et al.*, 2007; Fischer *et al.*, 2008; Graca da Silveira *et al.*, 2002; Ingram *et al.*, 1998; Ingram *et al.*, 1999; Klinke *et al.*, 2004; Lin & Tanaka, 2006; Ng *et al.*, 1981; Yomano *et al.*, 1998). The motivation for most studies is establishing an efficient system for biofuel production with high titers and productivities (a much more complete description of the problem and motivation can be found in Section 6.4.3). This ambitious goal is one that has driven enormous efforts in recent years and to which an equally large amount of funding has been directed.

To tackle the environmental tolerance problem, we wondered whether we could use transcriptional engineering with alternative sigma factors to further the achievements that started with the use of sigma D (see Section 3.3.3 and (Alper & Stephanopoulos, 2007)). There were at least two reasons for doing this. First, the stresses encountered in ethanol fermentations would favor transcription of regulons controlled by alternative sigma factors, and second, we could use the resulting mutants concomitantly with the previously-isolated sigma D mutant to try to magnify its effect.

There are several hypotheses for the toxicity of ethanol in bacterial systems. Similar to other alcohols, it is believed to act at the membrane, causing a general fluidization of the lipid bilayer and resulting in ion leakage through this barrier (Graca da Silveira *et al.*, 2002; Sikkema *et al.*, 1995). In turn, the leakage challenges the control of the intracellular pH, charge, and osmolarity, leading to energy dissipation and subsequent death. In addition, ethanol elicits a response similar to that of heat-shock, upregulating chaperones and similar protectants, suggesting that this alcohol may elicit or potentiate protein unfolding (Barbosa *et al.*, 1994; Yura *et al.*, 1993).

Another condition of interest for ethanol and other fermentations is high temperature. Cooling is a major energy input and a general hurdle of most bioprocesses; thus, running fermentations at higher temperature results in monetary savings and process simplification (Abdel-Fattah *et al.*, 2000). Ethanol production at high temperature has been recognized to have several advantages (Ng *et al.*, 1981), but it continues to be a challenge. A main reason is that heat is also known to cause membrane fluidization and protein denaturation, adding to the effects of ethanol (Missiakas & Raina, 1997; Shigapova *et al.*, 2005). Indeed, both ethanol and heat elicit similar responses in the cell, mediated by either or both sigma E and sigma H (Missiakas & Raina, 1997; Srivastava *et al.*, 2000; Yura *et al.*, 1993). Because these sigmas were likely to be bound to RNAP in high temperature and ethanol conditions, their mutant versions could conceivably produce the transcriptomic profile that could improve tolerance.

With this in mind, we screened the epPCR libraries in 50 g/L of ethanol at 42 °C. The challenge involved resuspending the cells in these conditions and plating them after 24 hr, and, therefore, it is termed a survivability assay. Other purifying selection techniques were tried, such as serial subculturing in 40 g/L and 42 °C, but we will not discuss these in detail.

Figure 4.3-1 shows representative data for 9 clones – 3 from each constructed library – and two controls – one carrying the empty pZE plasmid and one more bearing the wild-type operon in the same vector. The figure shows pre- and post-retransformation survivability after 24 hr in media with 50 g/L ethanol at 42 °C. As shown, the improvement is lost after retransformation, a trend that was observed in dozens of similar isolates. Furthermore, the data was consistently noisy so that it was hard to determine if the difference between pre- and post-retransformation survivability was due to genetic differences, to environmental conditions, or simply to stochastic effects. These two facts, added to the success we were encountering with the alpha subunit of the RNAP led to the abandonment of the sigma E and sigma H libraries for phenotypic improvement.

From the use of the three alternative sigma factors for transcriptional engineering we learned that not all targets are equally promising for improving a particular phenotype. This could be either because the choice of target is not adequate for engineering the phenotype that is being considered, which represents a theoretical limit for improvement, or because the quality of the libraries is not high enough. The latter implies that, even though variants of the target in question could improve phenotype, the library design is not such that these variants can be readily found given the fact that the search space is infinite. This issue was explained with an example in Section 3.2.1. The lack of information gained by this set of experiments (i.e. nothing more than the observation that the libraries at hand were not useful) was a key motivation for the experiments described in the next two Chapters.

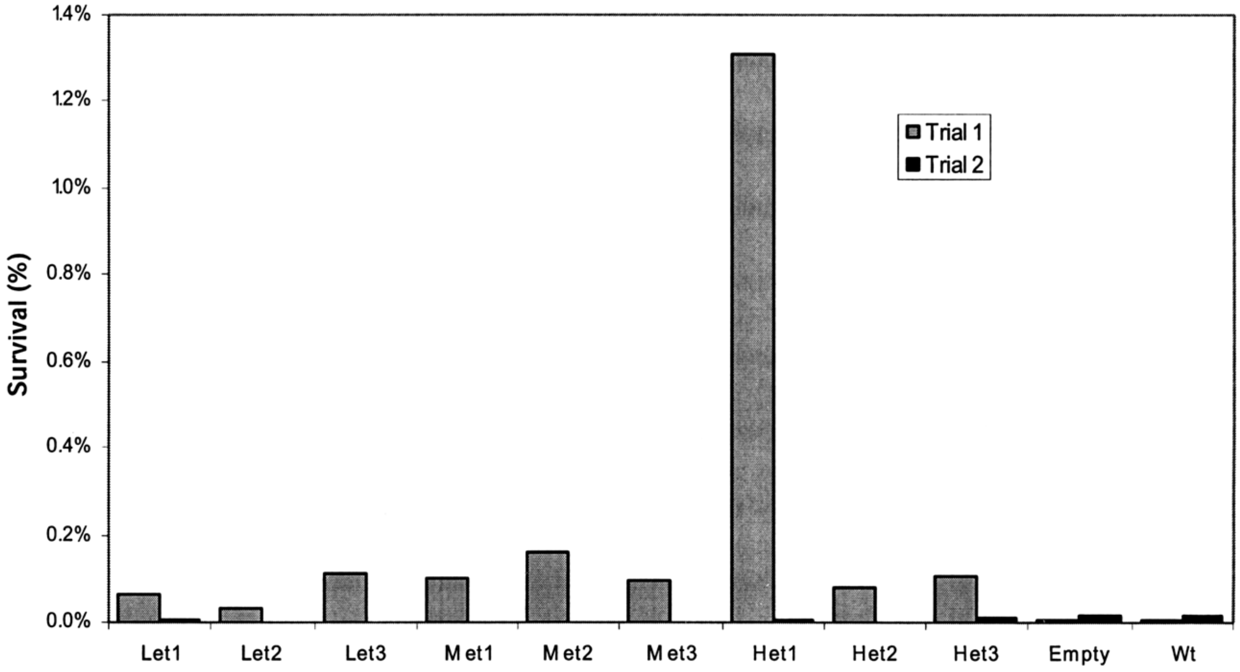

Figure 4.3-1. Survivability assay for sigma E -- sigma H mutants

The graph shows example survivability data (at 24 hr, in 50 g/L ethanol and at 42 °C) from high ethanol and high temperature survivability assays. The first letter in the mutant designation shows whether it was isolated from the library with low, medium, or high mutation frequency. Trial 1 was done before the mutant sigmas were retransformed into the host, and trial 2 was done after retransformation. Two controls (DH10B with an empty plasmid or one expressing the wild-type operon) are shown for comparison.

4.4 Alpha Subunit of the RNA Polymerase

As discussed in Section 4.1, the sigma factor is not the only subunit of the RNAP in charge of promoter recognition, although the evidence supporting alpha-promoter interactions came later and is thus less detailed. Simultaneously to our work with alternative sigma factors, we explored the use of the alpha subunit for transcriptional engineering in the spirit of the general goal of this thesis. Let us now turn to the reasons for why this protein seemed promising for phenotypic improvements.

Each RNAP complex contains two identical alpha subunits, each composed of two independently-folding domains (the NTD and the CTD, see Section 4.1) joined by a flexible linker. The NTD is in charge of alpha dimerization, the first step in RNAP assembly, and, subsequently, of recruiting the β and β' subunits to form the core enzyme (Kimura & Ishihama, 1995). Although the NTD has been commonly thought of as a structural domain, recent findings suggest that it also has regulatory functions.

An implicit regulatory function of the NTD stems from its role in RNAP assembly. Since the RNAP is in short supply in the cell, and this phenomenon determines the competition of the different sigmas for free core enzyme, the NTD has an indirect effect on the differential preference of the RNAP for the various promoters. A more overt form of regulation by NTD takes place at some Class II promoters. Niu *et al.* reported that a patch of negatively-charged amino acids (residues 162-165 in alpha) located in the NTD interact with the catabolite activator protein (CAP) at residues 19, 21, 96, and 101 (Niu *et al.*, 1996). The authors speculate that the same or nearby amino acids in NTD can have regulatory functions at other promoters or with other effectors.

These and similar studies suggest that NTD may be a promising mutagenesis target for transcriptional engineering. Most of the determinants of alpha in charge of transcriptional regulation, however, are found in the CTD. This domain is formed by a non-standard helix (NSH) followed by four α-helices, whose structure has been determined (Jeon *et al.*, 1995). The CTD interacts in at least three ways at the promoter, each of which will be discussed in turn.

The first category of CTD interactions is that with DNA. It has been have proposed that alpha binds and recognizes bases in the minor groove, at either one or two UP promoter elements situated between bases -59 and -38 with respect to the transcription initiation site (Gourse *et al.*, 2000). Structurally, the first helix and the loop between the third and fourth helices of alpha are the regions directly in charge of DNA recognition at the UP promoter site (Gaal *et al.*, 1996). Estrem *et al.* reported on the consensus UP sequence based on *in vitro* binding-selection experiments, and determined that it consisted of an A/T-rich stretch starting at base -38 (Estrem *et al.*, 1999). Furthermore, they showed that proximal and distal UP sub-sites were able to stimulate expression up to 170- and 16-fold, respectively. The effect of CTD-DNA interactions varies widely from promoter to promoter, and is not always stimulatory, but may be inhibitory in some cases (Ellinger *et al.*, 1994; Ross *et al.*, 1998). This diversity in interaction effects implies that the CTD is responsible to some degree for the differential expression across promoters. In fact, the CTD has been recently reported to interact at most, if not all, promoters (Ross & Gourse, 2005). Having these properties makes the alpha subunit a promising target for transcriptional engineering.

The second category of CTD interactions is that with activators or repressors that bind upstream of the -35 promoter hexamer, both at class I and class II promoters. At class I promoters the CTD is the main basis for activation, interacting with proteins that bind overlapping the UP promoter region (Browning & Busby, 2004). A similar arrangement of promoter elements in which the CTD interacts in an inhibitory, rather than a stimulatory, fashion has been found for repression mechanisms, such as that mediated by the *galR* repressor (Choy *et al.*, 1995; Choy *et al.*, 1997). At class II promoters the protein effector binds overlapping the -35 hexamer, and the CTD is responsible for one among many interactions (Busby & Ebright, 1999).

The third category of CTD contacts is that comprising interactions with sigma. Several lines of research have indicated that this interface, situated in amino acids 257-261 of alpha (Benoff *et al.*, 2002), has several regulatory functions. For example, a E261K mutation has pleiotropic effects in phenotype, ranging from inability to grow in minimal media, reduced growth in LB broth, cessation of growth at 42 °C in some media, distinct colony morphology, among others (Jafri *et al.*, 1996). Additional evidence was offered by a study showing that certain mutants of sigma (in particular, in region 4.2) and others in αCTD exhibit a marked decrease in transcription from some promoters (Ross *et al.*, 2003). This study proposes that alpha's D259 and E261 form a direct link with sigma D's R603. A somewhat recent paper (Gourse *et al.*, 2000) summarizes the results from other studies that suggest αCTD-sigma interactions.

As suggested by the multiple regulatory functions outlined here, the alpha subunit of RNAP appeared to be a good target for transcriptional engineering. Because the amino acid determinants for these functions appeared to be spread out, and for the same reasons tendered in the case of sigma D libraries, the initial approach for utilizing this target was to construct mutagenesis libraries encompassing the entire coding region of alpha.

4.4.1 *Use of the alpha Subunit for Phenotypic Improvement in E. coli*

As a first step towards implementing our transcriptional engineering approach with the alpha subunit of the RNAP, we cloned the wild-type version of the gene into the pHACM vector (see Section 4.2). The *rpoA* gene was amplified from genomic DNA and inserted downstream of the P_{lac} promoter. Although the promoter is inducible (e.g. with isopropyl β-D-1-thiogalactopyranoside, IPTG), observations in our laboratory suggested that the expression from this element was leaky and did not require induction, especially in rich or complex media.

The correct insertion and sequence were verified prior to amplification with epPCR with three mutation frequencies. The amplified mutagenized products were then cloned back into the pHACM vector, resulting in three libraries denoted rpoA*L, rpoA*M, and rpoA*H, indicating their low, medium, and high mutation frequencies. The plasmids bearing the mutagenized genes were isolated and used to transform different host strains, depending on the application. Initially, we considered tolerance to n-butanol, and production of hyaluronic acid and L-tyrosine as the complex phenotypes in which our newly-developed tool would be applied.

4.4.1.1 Tolerance to n-butanol

Butanol has gained substantial attention for its potential as a biofuel, because it possesses several advantages over the more widely-used compound, ethanol (Durre, 2007). Most notably, butanol has a higher energy content and is not water soluble, which greatly facilitates its storage and handling (Durre, 2007; Wackett, 2008). However, the hydrophobicity of butanol makes it especially toxic for the production organism. Recent studies report the production of this and similar molecules in *E. coli* and allude to strain improvement methods for increasing butanol tolerance in this strain (Atsumi *et al.*, 2008a; Hanai *et al.*, 2007).

The toxicity of butanol is thought to arise from the fluidization and disordering of membrane lipids and the consequent leakage of ions through it (Sikkema *et al.*, 1995). In particular, dissipation of the transmembrane pH gradient has an energy uncoupling effect (Borneman *et al.*, 2007; Graca da Silveira *et al.*, 2002), similar to that caused by ethanol (see Section 4.3.2) and weak acids at low external pH (see Section 4.2.2) (Pieterse *et al.*, 2005; Valli *et al.*, 2006).

We began by screening the *rpoA* libraries in butanol in a DH5α *E. coli* host. We used serial subculturing in high butanol media for our purifying selection step, similar to that described earlier for *L. plantarum*. After several mutants were isolated and retransformed into the parental strain, a single mutant (denoted L33) was selected for further characterization. Growth in the presence of butanol was linear in time (not exponential), and the mutant exhibited a significantly steeper rate of growth (slope was 0.08±0.01 for mutant vs. 0.05±0.01 for wild-type in 0.9% n-butanol, see **Figure 4.4-1**) and higher accumulated cell mass (**Figure 4.4-3**).

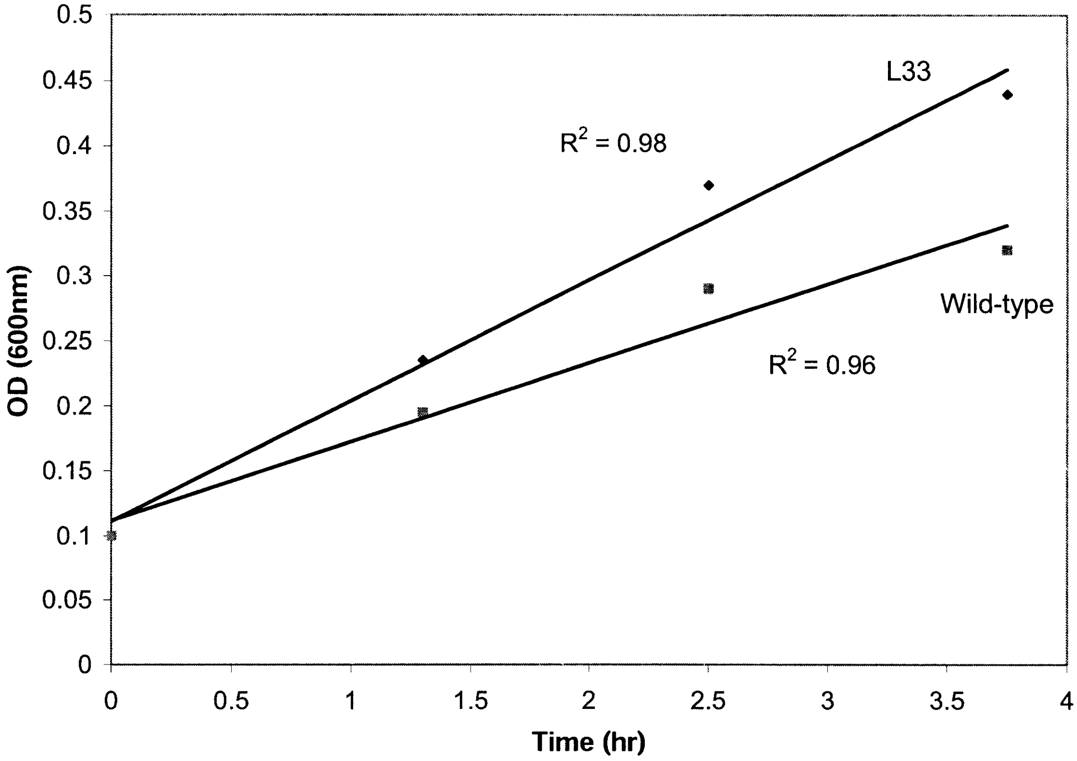

Figure 4.4-1. Growth assay in butanol

The plot shows growth curves of DH5α cells transformed with either the wild-type or the L33 mutant of *rpoA* in 0.9% butanol. The assay was done in duplicate and the slopes and their confidence intervals were calculated as given in the text.

A few points about **Figure 4.4-1** are in order since the graph shows a linear growth profile based on four data points. A more detailed growth curve (with more time points) could not be performed, due to butanol losses by evaporation associated with opening the tubes for sampling. This would increase the noise in the data and would make it hard to interpret. The linear profile refers to the fact that when the experiment with a few time points (usually 4) was repeated, a line was consistently a statistically better fit than an exponential curve (as judged by the sum of squared errors).

This implies that either the growth is slow and appears linear (i.e. in the vicinity of time t=0, the tangent line of an exponential curve with a small time constant approaches the curve itself), or that growth cannot be described as a first-order process because it is not limited by the number of individuals in the population. In the latter case, growth could be described as a zeroth-order process, as when limiting reactions for growth are occurring at saturation.

Upon analyzing the *rpoA* plasmid isolated from L33, we found a nucleotide substitution that changes the amino acid E244 to a stop codon, resulting in a truncated protein that lacks the αCTD (**Figure 4.4-7**). Interestingly, similar *rpoA* mutants have been isolated in the past and have been widely studied. A protein lacking the αCTD is capable of being assembled into the RNAP and carrying out transcription; however, it does not respond to signals in the DNA or to protein effectors (Igarashi & Ishihama, 1991; Igarashi *et al.*, 1991; Murakami *et al.*, 1996; Ross *et al.*, 1993; Savery *et al.*, 2002). For example, alpha subunits lacking the αCTD do not respond to the strong activating signals from the UP-region of *rrnB* promoters (Paul *et al.*, 2004; Ross *et al.*, 1993) or the cAMP-CRP complex (Igarashi & Ishihama, 1991).

The changes associated with these interactions would be significant even in the absence of other effects, considering that the products of *rrnB* promoters make up a major fraction of the total RNA in the cell (Paul *et al.*, 2004) and that more than 100 loci are activated by CRP (Botsford & Harman, 1992; Savery *et al.*, 2002). It is important to note that because the chromosomal copy of *rpoA* remains intact, the observed phenotype likely arises from the combined action of the truncated and non-truncated forms of the protein.

4.4.1.2 Tolerance to Other Solvents

The toxicity of alcohol solvents in bacteria is known to be related to their lipophilicity. The higher the solubility of the alcohol is in the membrane, the higher is its bioavailability, and the more pronounced its harmful effects are for the cell (Sikkema *et al.*, 1995). As a first step towards exploring the tolerance of mutant L33 to other solvents with potential biofuel uses, we tested the growth inhibition of several alcohols for our wild-type strain. As shown by **Figure 4.4-2**, there is a clear increase in toxicity with increases in hydrophobicity, as seen from the comparison between the linear and branched versions of propanol and butanol.

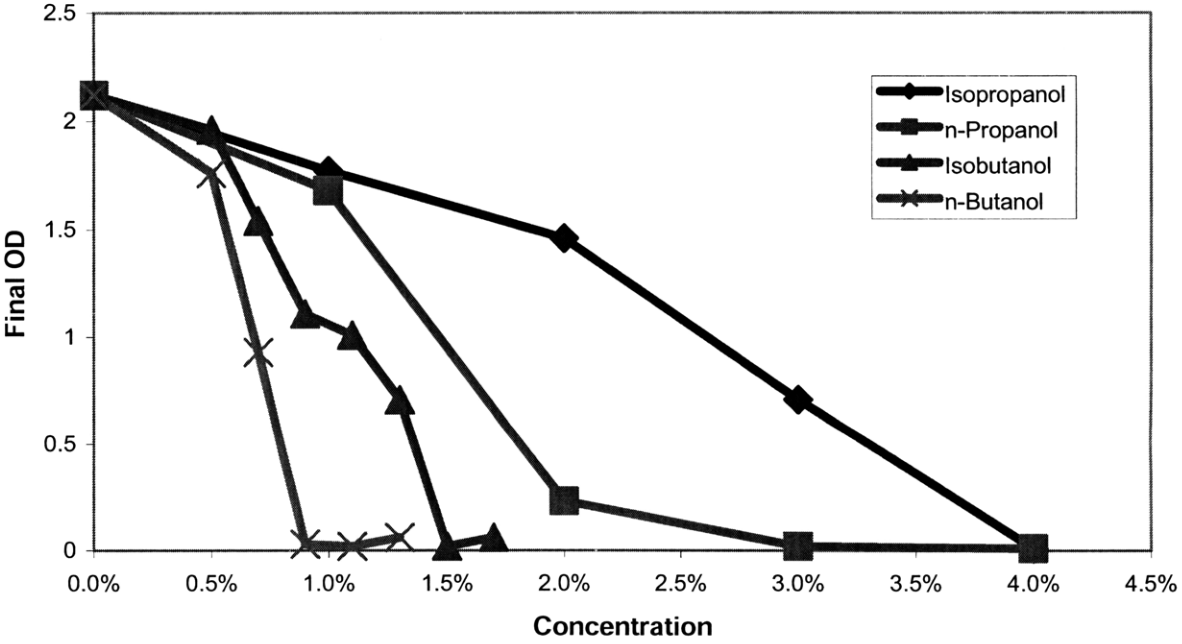

Figure 4.4-2. Test of toxicity of various alcohols

The graph shows the final optical density (600 nm) for DH5α *E. coli* cells grown in M9 minimal media and different concentrations of different short-chain alcohols.

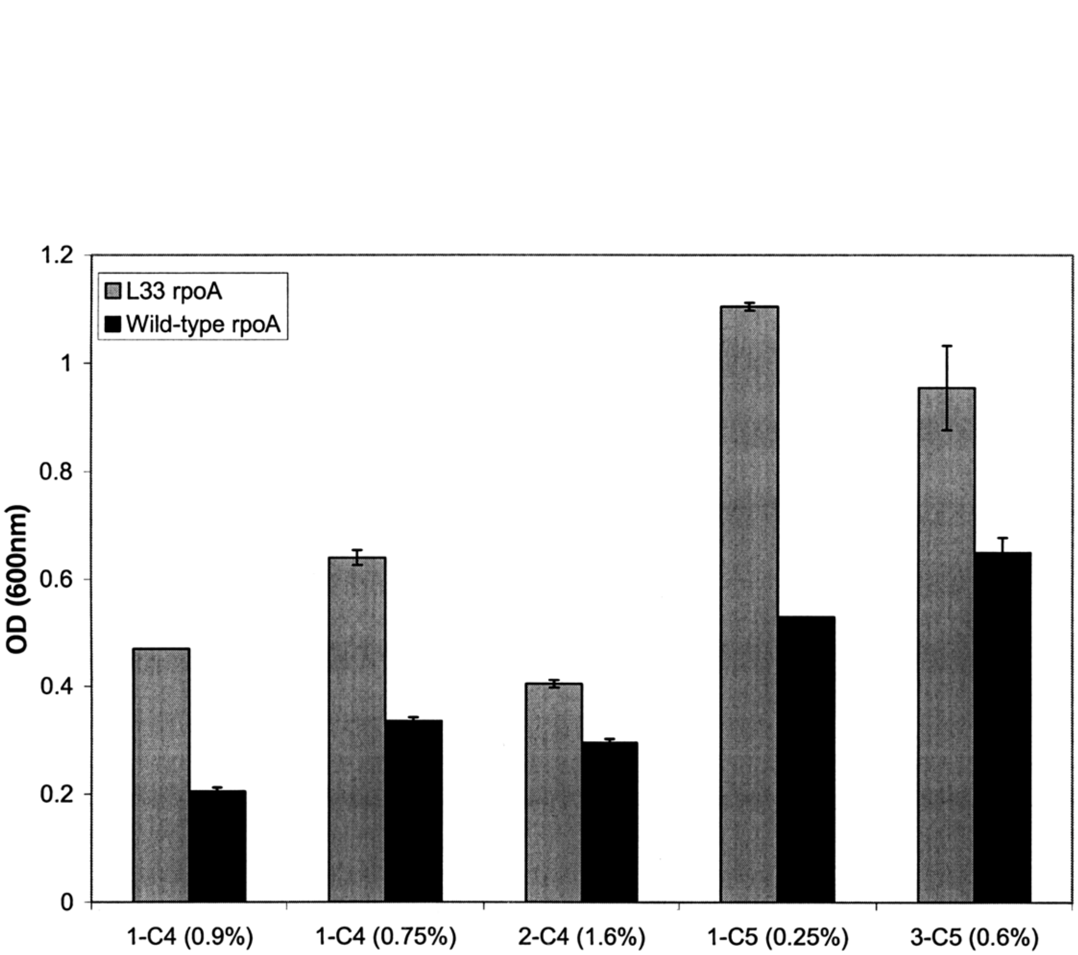

Figure 4.4-3. Growth assay for mutant and wild-type in various alcohol solvents

Overnight growth (after 24 hr) of DH5α cells transformed with either the wild-type or the L33 mutant of *rpoA* in different alcohol solvents. The alcohols are named first by the carbon atom where the hydroxyl is found and last by the number of total carbons in the molecule (e.g. 2-C4 is isobutanol). The concentration used is in parenthesis (v/v).

Next, we tested L33 in several alcohols; we chose n- and iso-butanol and n- and 3-pentanol since their high heating value makes them potential biofuels. We hypothesized that if the mutant *rpoA* negated the effects of butanol through a decrease in membrane fluidity or a related response, the mutant would also exhibit resistance to other solvents that are known to act by similar mechanisms. We tested the tolerance of the mutant to other alcohols and found that the strain performed better than the control for all cases, as measured by the accumulation of cell mass in the presence of the toxic solvent (**Figure 4.4-3**).

4.4.1.3 Tolerance to Ion Leakage of Butanol-Tolerant Mutant

In order to test the ability of the mutant to cope with ion leakage and energy uncoupling, we measured the intracellular pH (pH_i) in the presence of a weak acid at low extracellular pH (pH_e) with and without added butanol. Cells were acid-shocked by resuspending them in potassium phosphate buffer at pH_e of 4.7. By comparing the pH_i of wild-type and mutant cells in the presence of butanol to that of the control with no added butanol we tested the capability of the strains to maintain the pH_i when in contact with the alcohol.

The results are displayed in **Figure 4.4-4**. A negative pH difference implies that the pH_i of the strain in the presence of butanol is lower than that without butanol, which is the expected result for the wild-type (as shown). In contrast, L33 maintains a higher pH_i throughout the experiment compared to the wild-type with the same amount of butanol, and also compared to the control strain with no butanol (thus, the pH difference is positive). These observations suggest that the mutant L33 copes with the stress either by reducing the fluidity of the membrane or by increasing the rate at which protons are forced out of the cell, or by a combination of both.

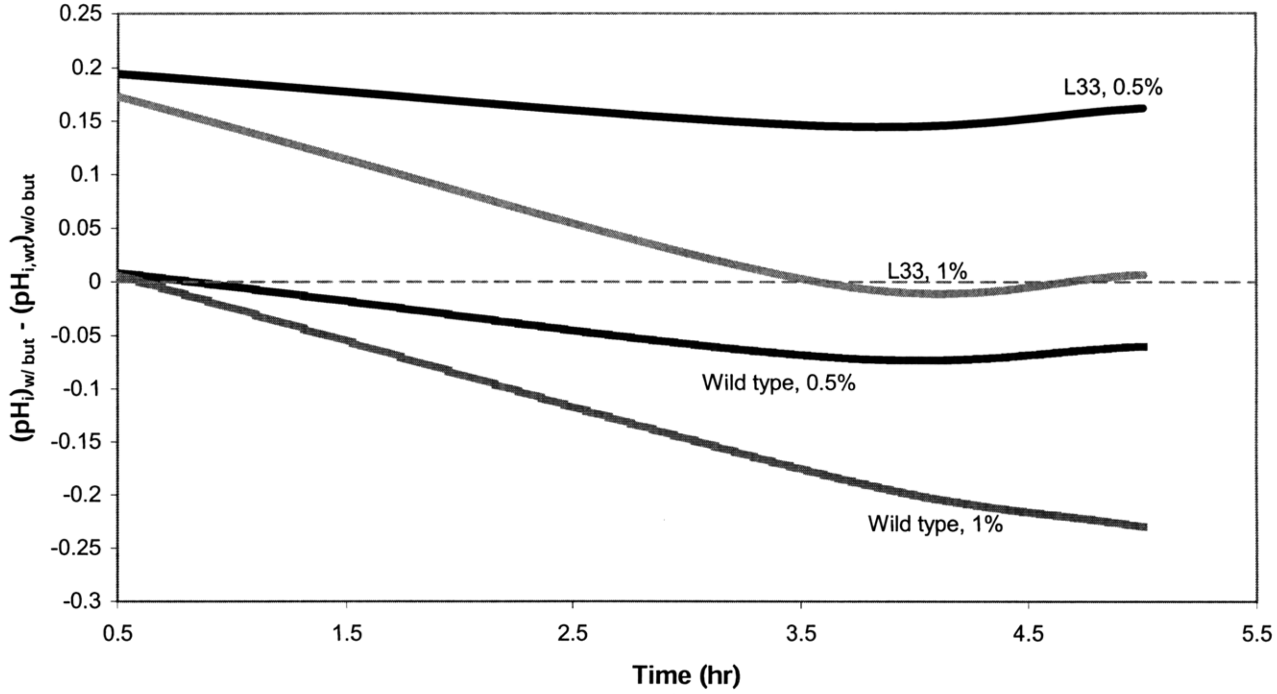

Figure 4.4-4. Ion leakage assay for mutant and wild-type

The graph illustrates the ability of the butanol-tolerant mutant (L33) and wild-type strains to maintain their intracellular pH (pH_i) in the presence of butanol. The y-axis shows the pH_i difference between cells resuspended in buffer (pH_e=4.7) with the indicated amount of butanol (v/v %) and wild-type cells resuspended in buffer with no butanol. Error bars are not shown, but the CV of the pH measurements was on average 0.4%.

4.4.1.4 Enhanced Production of L-Tyrosine

The aromatic amino acid L-tyrosine has several pharmaceutical and industrial applications, thus making it an interesting target for production in *E. coli* (Lutke-Eversloh & Stephanopoulos, 2007). For example, because this amino acid is the biochemical precursor for important neurotransmitters in the brain, its use in the production of treatments for cognitive ailments such as Parkinson's disease is being considered (Kim do *et al.*, 2007). L-tyrosine is also used as an essential dietary supplement in patients that suffer from phenylketonuria, a condition that interrupts the synthesis of tyrosine from phenylalanine (Lutke-Eversloh *et al.*, 2007).

Directed genetic modifications have been quite successful in increasing the titers of L-tyrosine for creating an industrial strain. For example, Lutke-Eversloh and Stephanopoulos reported that by introducing and overexpressing feedback-resistant versions of two controlling enzymes in the tyrosine pathway, *aroG* and *tyrA*, and by simultaneously deleting the *tyrR* transcriptional regulator, they were able to increase the production of the amino acid from non-detectable levels to 3.8 g/L (Lutke-Eversloh & Stephanopoulos, 2007). Additional overexpressions helped to increase the availability of its precursors, phosphoenolpyruvate and erythrose-4-phosphate, resulting in a mutant that reached 9.7 g/L.

Although this and comparable studies attest to the impact that traditional strain improvement approaches can have for enhancing tyrosine production, a random knockout search revealed that increases in titer may also arise from the modulation of seemingly unrelated or unknown factors (Santos & Stephanopoulos, 2008b). This early lead makes this system well-suited for transcriptional engineering studies.

We began by transforming the *rpoA* libraries into the pre-engineered L-tyrosine-producing strain that contains a few modifications in addition to those previously described: *E. coli* K12 Δ*pheA* *tyrR*::P$_{\text{LTET-O1}}$ *tyrA*$^{\text{fbr}}$*aroG*$^{\text{fbr}}$ *lacZ*:: P$_{\text{LTET-O1}}$ *tyrA*$^{\text{fbr}}$*aroG*$^{\text{fbr}}$. We then subjected it to a high-throughput screen based on the synthesis of the black pigment melanin, which exploits the fact that this compound is produced from L-tyrosine (Santos & Stephanopoulos, 2008b). The referenced study established that the titer and productivity of melanin correlates with L-tyrosine overproduction, and thus it is possible to infer the level of the compound of interest based on how dark do colonies appear and how fast does the color develop during the screen.

From an initial library size of 7.5x10^5 mutants, 30 of the darkest strains were selected by visual inspection and tested for L-tyrosine production in MOPS minimal medium. We were able to isolate a strain (denoted rpoA14) exhibiting a 91% increase in titer above the parental pre-engineered strain, with a final concentration of 798 mg/L L-tyrosine (**Figure 4.4-5**).

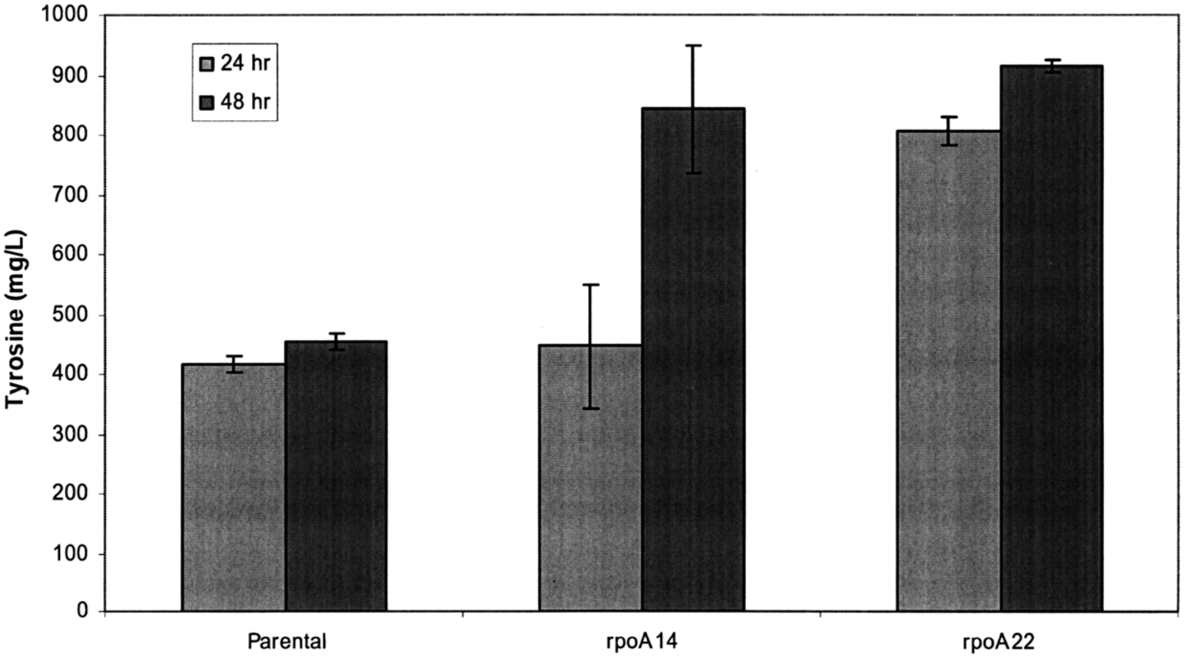

Figure 4.4-5. L-Tyrosine production of parental strain and mutants

The parental is an *E. coli* K12 Δ*pheA tyrR*::$P_{LTET\text{-}O1}$ *tyrA*^{fbr}*aroG*^{fbr} *lacZ*:: $P_{LTET\text{-}O1}$ *tyrA*^{fbr}*aroG*^{fbr} and the other two harbor the mutant *rpoA* genes. Mutant rpoA14 was isolated from the error-prone PCR libraries, and mutant rpoA22 from saturation mutagenesis libraries.

It is interesting to note that the observed L-tyrosine producing phenotype required *both* the background of the isolated strain and the presence of a mutant *rpoA* plasmid. Therefore, it is likely that this particular strain incurred additional mutations within the chromosome that act in concert with the mutant *rpoA* to enhance L-tyrosine production. Thus, *rpoA* mutagenesis can also act synergistically with natural variations introduced during normal replication processes.

Sequencing of the mutant *rpoA* gene revealed two amino acid changes, V257F and L281P (**Figure 4.4-7**), located in the αCTD near amino acids known to contact regulatory factors and the UP-element (Murakami *et al.*, 1996). The first change occurred in the NSH, while the other was located in one of the four α-helices of the αCTD (Gaal *et al.*, 1996). It is likely that both of these mutations alter the interaction of the αCTD with its target proteins or sequences through changes in the αCTD structure. In particular, the amino acid proline has been implicated in destabilizing α-helices in some conditions (Li *et al.*, 1996). A change in helix conformation could indirectly affect the positions of the amino acids responsible for making contacts with the promoter, thus altering the affinity of the RNAP for some of its targets.

In order to test whether other amino acid substitutions in positions V257 or L281 could improve the production of L-tyrosine further, we constructed a saturation mutagenesis library with mutations restricted to these residues. The resulting library was transformed into the parental strain and screened as before. After 40 of the darkest strains were selected and individually tested, one was chosen for further analysis. This mutant, denoted rpoA22, has a V257R mutation and no change in L281. As shown in **Figure 4.4-5**, the mutant produces similar final titers compared to rpoA14, but has a substantial increase in the rate of production. At 24 hr, when rpoA14 is still indistinguishable from the parental, mutant rpoA22 has already reached nearly 90% of its final titer. Therefore, this mutant offers an interesting opportunity for developing an industrial platform for continuous production.

4.4.1.5 Enhanced Production of Hyaluronic Acid

As an additional phenotype to improve using our new approach, we chose hyaluronic acid (HA) production, using the same high-throughput screening (HTS) method that we described in Section 4.2.1.1 (Yu *et al.*, 2008a). As a reminder, cells with increased titers are first recognized by their translucent appearance in solid medium, and second, the HA is quantified by precipitation of a dye, alcian blue.

Using this HTS platform, we isolated improved *rpoA* mutants as shown in **Figure 4.4-6**. The figure shows the HA concentration relative to the control, quantified using the alcian blue method previously reported (Yu *et al.*, 2008a). It is obvious from **Figure 4.4-6** that significant phenotypic diversity was introduced to the production strain via the *rpoA* mutant library, which is a qualitative indication that *rpoA* is a good target for transcriptional engineering (see next two Chapters). Mutant rpoA-HA, indicated with a hatched bar in **Figure 4.4-6**, was further examined. A single mutation (L254Q) located in the so-called 'non-standard helix' (NSH) of the αCTD (Gaal *et al.*, 1996) was revealed by sequencing (**Figure 4.4-7**).

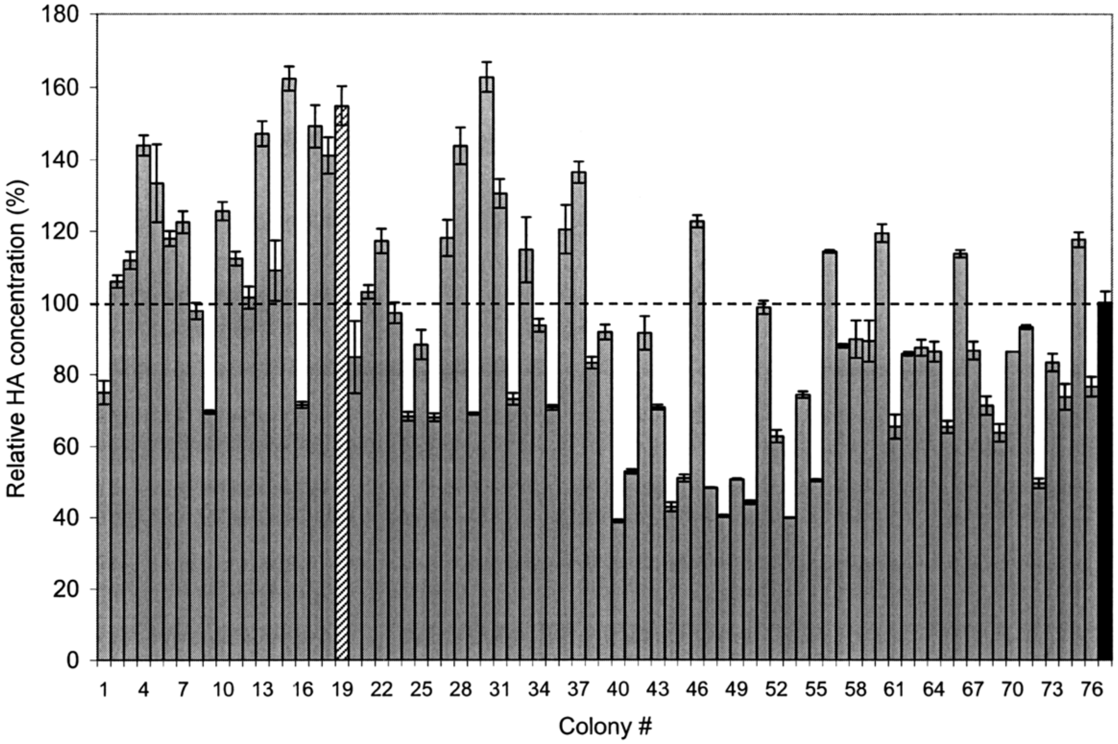

Figure 4.4-6. Alcian blue quantification of HA production by selected rpoA mutants

The host strain and the control (black) is *E. coli* Top10/(pMBAD-*sse*ABC), while the rpoA-HA mutant is indicated with a hatched bar. All samples were measured in duplicate.

4.4.1.6 Location of the *rpoA* Mutations

We have described how mutants of *rpoA* elicited improvements in three distinct phenotypes – butanol tolerance, L-tyrosine production, and HA accumulation – which opens the possibility that this technique could be used for improving a broad spectrum of other interesting traits. The evidence presented in the previous sections suggests that this capacity is related to the function of the αCTD, because all the mutations found were mapped to this region. This domain of the protein has been implicated in contacting promoter DNA and protein activators and repressors (Browning & Busby, 2004; Dangi *et* *al.*, 2004), suggesting that it is the regulatory function of the alpha subunit which gives rise to the pleiotropic alteration of the transcriptome that results in novel phenotypes. The αNTD has also been shown to be a target for transcription regulation (Niu *et al.*, 1996), but no mutations were found there in the present study. **Figure 4.4-7** summarizes schematically the location of the mutations in the alpha protein that correspond to the different phenotypes.

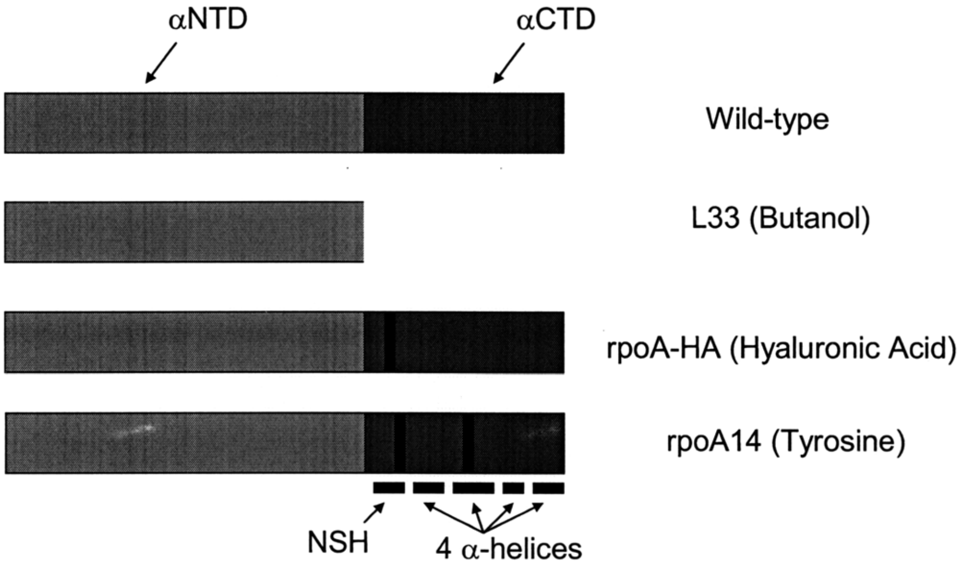

Figure 4.4-7. Schematic mapping of the mutations on the rpoA protein

Gross features, such as the N- and C- terminal domains (αNTD and αCTD), the non-standard helix (NSH), and the four α-helices of αCTD (Gaal *et al.*, 1996) are indicated.

4.4.2 *Phenotypic Specificity of the alpha Subunit Mutants and Comparison to Alternative Approaches for Strain Improvement*

The fact that all mutations found in this study were related to the function of the αCTD, but none is located in residues known to contact DNA or protein regulators, opens the possibility that the isolated mutants are not highly specific to the phenotypes for which they were selected. For example, rpoA14 may be a loss-of-function variant of the CTD, and therefore could confer better growth in butanol (recall that L33 is a loss-of-function of the CTD); the same could be true for rpoA-HA.

To test the specificity of the mutant genes, we transformed rpoA14 and rpoA-HA into DH5α and measured growth in 0.9% n-butanol as before. Neither shows an improvement similar to that of L33 (Table 4), which implies that these mutants have retained at least some CTD functionality.

To further test the specificity, we tried a similar experiment in which L33 and rpoA-HA were transferred into the tyrosine-producing parental strain (prepared by curing the rpoA14-containing plasmid). As shown in Table 4, all three plasmids confer similar increases in tyrosine production, which implies that small alterations of CTD function (probably through partial misfolding) can increase production in this background. In other words, the tyrosine-production phenotype is not highly specific to rpoA14, but the butanol-tolerance phenotype is specific to L33.

Previous studies have shown the usefulness of other transcriptional engineering targets for cellular engineering, so it is appropriate to compare the use of *rpoA* to other targets for improving the aforementioned phenotypes. For the case of butanol, *rpoD* libraries were tried in parallel with those of *rpoA*, but no *rpoD* variants were isolated that confer a growth advantage. We did not feel, however, that we have enough experimental evidence to suggest that *rpoD* could not be used for improving butanol tolerance. Perhaps such a mutant is theoretically achievable, but it was not found given the inherently finite size of the library. For the case of tyrosine, *rpoD* libraries have been screened and mutants were isolated with levels of tyrosine slightly higher than that of rpoA14. The data supporting this observation will be the focus of a paper that is, as of now, in preparation (Santos and Stephanopoulos). For the case of HA, *rpoA* mutants show an up to ~60% improvement in titer (**Figure 4.4-6**) compared to ~40% in the case of *rpoD* (see **Figure 4.2-1** in Section 4.2.1.1).

Although we do not reckon these comparisons to be a good basis for determining which target is best, we do believe that mutagenesis of the alpha subunit may complement the use of sigma factors for strain improvement. The alpha subunit is permanently bound to the RNAP holoenzyme and has been shown to interact with most, if not all promoters (Ross & Gourse, 2005), circumventing the fact that some transcriptional states may be hard to access by sigma D in certain conditions (Ishihama, 2000; Jishage *et al.*, 2002). The next Chapter focuses on developing the conceptual grounds and tools to make a fairer comparison between targets for transcriptional engineering, and between random search-based libraries in general.

Table 4. Phenotypic specificity of *rpoA* mutants

		Phenotype	
		Butanol[1]	Tyrosine[2]
Mutant Plasmid	rpoA14	14%	90%
	L33	129%	93%
	rpoA-HA	-3%	119%

1. Percent growth advantage (measured as the final OD_{600}) of DH5α containing mutant *rpoA* plasmids vs. wild-type rpoA plasmid. Experiments were conducted in 0.9% n-butanol.

2. Percent increase in L-tyrosine titer (48hr) compared to that of the parental strain. Mutant plasmids were introduced into an rpoA14 mutant background (generated by curing the original plasmid through serial subculturing).

Chapter 5

5. Phenotypic Diversity as a Metric for Evaluation of Random Strain Improvement Libraries

The discussion in Chapter 3 (Sections 3.2 and 3.3 in particular), in which we described several examples of random search-based library designs, insinuated that there are different routes one can take to elicit the same phenotype. Furthermore, each route can be applied with different experimental parameters (e.g. location and frequency of mutations, etc.). Because time and resources are limited, not all routes or parameters can be explored, neither simultaneously or sequentially.

The present thesis started at the time when strain improvement through engineering of the native transcriptional machinery was being tried for altering different phenotypes, and this technique became the focus of this work. Given the list of alternative tools for accomplishing the same goal, it was also relevant to study and compare our choice in the context of available options. Perhaps other library designs would work better or bring improved strains faster, and maybe other strategies for implementing the same library design were more efficient. The question of how to determine the potential of different random search-based approaches for phenotypic improvement became a central question of this thesis.

With these factors in mind, we began concretizing our thoughts about the value of and the means for evaluating different strain improvement approaches. In this Chapter, we cover the motivation behind evaluating the potential of different libraries, we describe the development of a metric for accomplishing this, and we delineate its implementation.

5.1 Motivation for Developing a Metric

Random search-based approaches for strain improvement are based on generating genotypic diversity and finding a variant of interest. As previously argued, purifying selection is the key time- and resource-intensive step (see Section 3.1.2); adding to the challenge is the fact that there is a long list of possible methods available for generating genotypic diversity. A main motivation for developing a metric for evaluating and comparing different library designs is thus the economization and simplification of strain improvement programs.

Evaluation of an experimental approach can be qualitative or quantitative; the latter necessitates the definition of a metric. There are at minimum two arguments for the usefulness of a metric for evaluating random search-based approaches for strain improvement: (i) it can provide the basis for comparing different experimental routes for building libraries; and (ii) it can aid in guiding and optimizing the construction of these libraries. Both hinge on the same principle, that of comparing different libraries and sequentially determining what design is most promising. Let us now explore these arguments in more detail. Throughout the discussion, one ought to keep in mind that evaluation of random search-based approaches may be a conceptually concrete objective but, experimentally, it must be done *indirectly* through evaluation of the libraries constructed with these approaches. Therefore, two libraries constructed based on the same design may differ in their quality because of the intrinsically imperfect reproducibility of experiments (see Section 3.2.1 for more on this argument).

5.1.1 *Comparing Different Experimental Approaches for Building Libraries*

As outlined in Sections 3.2 and 3.3, there are several ways in which genotypic diversity can be introduced into a population, resulting in different library designs. Even if one restricts the possibilities by only considering mutagenesis-based libraries, there is an infinite number of designs, depending on where does one choose to target the mutagenesis (e.g. a sigma factor, a short RNA molecule (siRNA or ribozymes), a ribosomal protein, a rate-limiting enzyme, etc.), the mutation frequency, the identity of mutations, etc. By changing these parameters, one can alter the characteristics of the libraries. Different library designs comprise different theoretical sequence spaces, and each of these spaces covers a different phenotypic space. In other words, depending on the library design, some phenotypes may be easily achievable in the resulting population and some may not.

Let us clarify the meaning of these ideas with an example. Imagine we are to improve the production of compound B based on the metabolic network of **Figure 5.1-1**, and consider two library designs. The first one, let us call it L1 (library 1), is a population of which the promoter of enzyme e_4 is randomized; the second, L2, is one in which the active site of enzyme e_{11} is targeted for mutagenesis. Because of the location and function of e_4 and e_{11} in the metabolic network, it is intuitive to assume that the sequence space of L1 will be mapped to a phenotypic space that is more likely to contain an improved variant with respect to the production of B. That is so because there are more mechanistic ways in which a mutation contained in L1 will translate into an improvement in the phenotype of interest compared to L2. Note that this comparison is set forth on probabilistic grounds: the fact that it is more likely to find the improvement in L1 than in L2 does not mean that there is no mutant in L2 that could potentially exhibit increased production of B (through an indirect effect).

Now, if the probable effect of e_4 and e_{11} on the production of B is not known and cannot be guessed, then it is difficult to know *a priori* which design would hold more promise for accomplishing our goal. A metric that would measure the potential of L1 and L2 for harboring a mutant with altered synthesis of B could allow deciding which library to screen in the absence of the mechanistic knowledge about the phenotype of interest (this information was provided by the metabolic map in the case of our example).

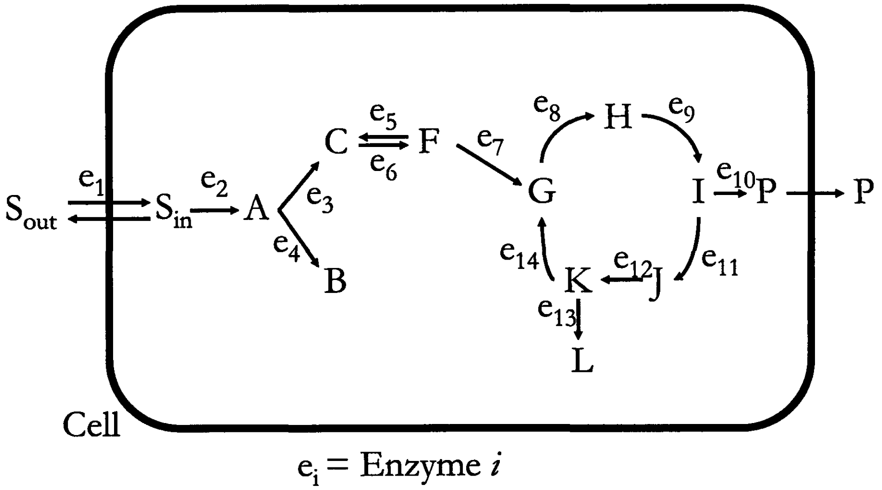

Figure 5.1-1. Example metabolic map

5.1.2 *Optimizing the Construction of Libraries*

If L1 fails to deliver a substantially-improved mutant, even when e_4 is known to be the rate-limiting enzyme for the production of B, one may consider optimizing the L1 library. In other words, the library design may be relevant in principle, but the parameters chosen to implement the design may be not. For example, perhaps the promoter randomization included an effector binding region, but it may be that for the case of e_4 this region is critical for having any expression at all. Mutations there would inactivate the enzyme, making it improbable to find improved variants in L1. In this case, a third library, L3, could have been constructed by restricting base pair changes to certain promoter regions – as in (Hammer *et al.*, 2006; Jensen & Hammer, 1998a; Jensen & Hammer, 1998b) – that do not include the effector binding sequence. Other designs could also be thought of, and, again, a metric that could allow a comparison of the options prior to screening would facilitate the effort of finding improved variants.

5.1.3 *Infiniteness of Sequence Space*

In the sense given by this example, different library designs result from choosing different targets for mutagenesis, whereas optimization results from modifying the parameters of the library design. In both cases, the goal is to build a sequence space with a higher likelihood of harboring a variant of interest. It is important to consider that even for small targets, theoretical sequence spaces are astronomically large, as discussed in more detail in Section 3.2.1.

In this context, optimization of a library design may simply involve reducing the search space so that it is easier to find a phenotype of interest. In other words, even when the first design of a library could in principle contain all variants present in a subsequent, optimized library, the latter may be better if there is a higher likelihood that the desired mutant will be found there. Mutations in the former design are said to be diluted in sequence space, so that particular variants are harder to be found.

Although we are considering mutagenesis-based designs, knockout and overexpression libraries, or any other ways of generating diversity are conceptually equivalent. In all cases, the infiniteness of the search space invites the development of a metric that aids in navigating it. We will now examine a concrete definition of such a metric, and explain how it can be used to guide the construction of libraries for strain improvement.

5.2 The Choice of Phenotypic Diversity as a Metric

The fact that, qualitatively, an optimal library design is one in which there is a high probability that a phenotype of interest will be found is intuitively correct, but insufficient for exploiting it in practice. In particular, as became apparent from the results discussed in Chapter 4 and the examples of Section 3.3, it is not *a priori* specified what traits are of interest, because the same libraries can be screened for improvement of different and even distant phenotypes. This is in contrast to cases where the property of interest is known from the start (as with most protein engineering searches), for which a better library can be pragmatically regarded as one that delivers a better trait. An alternative definition for a good library that is in better tune with the aforementioned requirements is, then, one that is phenotypically diverse. The aim becomes to design a sequence space that maps to a highly diverse phenotypic space, because in a vastly diverse population there is a higher *a priori* probability that any (yet unspecified) phenotype of interest can be found.

The metric for library evaluation must measure phenotypic diversity, as concluded from the previous paragraph, but the question remains of how to do this in practice. Because cellular physiology can be described as a highly interconnected network and we are manipulating central nodes of this system (Jeong *et al.*, 2000; Martinez-Antonio *et al.*, 2008), we do not need to evaluate the phenotype that we are immediately interested in. Instead, we can assume that if we measure diversity for a complex phenotype, dictated by the activity of many nodes in the network, we are indirectly sampling diversity for many distant phenotypes. This idea is supported by the observed pleiotropy of the mutations introduced in *rpoA* and *rpoD* (Chapter 4). Therefore, we hypothesized that we could use diversity in complex phenotypes to evaluate and compare strain improvement approaches.

5.3 Divergence: a Normalized Phenotypic Diversity Metric

We based our metric for capturing phenotypic diversity on the concept of speciation. Since the appearance of new species is not a discrete process, but one that results from the accumulation of traits, we envisioned that the engineering of new phenotypes can be regarded as small steps akin to those that arise during speciation. The individuals of a species are thought of as forming a tight cluster based on genotypic or phenotypic characteristics (Cummings *et al.*, 2008); therefore, a highly diverse population would be composed of many clusters, whereas a homogeneous population would form one or a few clusters.

In this case, we framed our effort for quantifying diversity in terms of a phenotypic cluster, and defined a metric based on phenotypic distance:

$$\begin{aligned} d &= \langle d_{i,j} \rangle \forall i,j \\ d_{i,j} &= |P_i - P_j| \end{aligned} \qquad \text{(Eq. 1)}$$

In Eq. 1, the phenotypic distance is calculated as the Euclidean distance between pairs of individual phenotype values (P_i and P_j), and the average phenotypic distance, d, is calculated as the average across all pairs. The value of d measures how different, on average, the members of a population are to each other with respect to a specific trait, quantified by *P*.

There are many mathematical substitutes for this formulation, the most obvious being the standard deviation, i.e.

$$d=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i-\langle P\rangle\right)^2} \qquad \text{(Eq. 2)}$$

The standard deviation can be thought of as a measure of the mean phenotypic distance between each member of the population and the average phenotype. Because the average phenotype is not necessarily a physically-meaningful value (as in the case of tight and distant clusters), we did not use the standard deviation in our definition of a metric. Nevertheless, a few informal calculations with our data showed a good correlation between the average phenotypic distances given by Eqs. 1 and 2.

It is well known that all populations, including those that are clonal, possess a certain inherent variability and biological noise (Elowitz *et al.*, 2002; Kaern *et al.*, 2005; Swain *et al.*, 2002). This means that the distribution of a phenotypic value is not uniform and, therefore, the corresponding phenotypic distance is non-zero. Hence, the distance value by itself is of little value unless it is properly normalized. The average phenotypic distance of a library population should then be compared to that of the unmutated control, yielding the "additional" phenotypic distance introduced in the library, which we called "divergence".

Because we consider a subsample of a population when measuring the average phenotypic distance (or else the experimental protocol and calculations would become intractable), the value of *d* has a distribution (i.e. the sampling distribution). Therefore, the normalization alluded to earlier must account for both the mean and the dispersion of the distribution of *d*; in general, this can be done using so-called statistical distance measures. We chose Bhattacharyya's equation (Eq. 3) to normalize the average phenotypic distance of a library by the distance of the control population (Hansen *et al.*, 2003):

$$BD = \frac{1}{8}(\mu_l - \mu_c)^T \left(\frac{\Sigma_l + \Sigma_c}{2} \right)^{-1} (\mu_l - \mu_c) + \frac{1}{2} \ln \left(\frac{\left| \frac{\Sigma_l + \Sigma_c}{2} \right|}{\sqrt{|\Sigma_l||\Sigma_c|}} \right) \qquad \text{(Eq. 3)}$$

Where Σ is the covariance matrix, μ is the mean vector, and the subscripts *l* and *c* are for the library and control populations, respectively. This form of the equation implies that the distribution of *d* is normal.

The divergence value given by Eq. 3 can be calculated with more than one phenotype. In that case, the distance is higher-dimensional, and so are the vectors and matrices in Eq. 3. The two-dimensional case (i.e. for two phenotype values) is exemplified by **Figure 5.3-1**, in which the distributions of *d* are graphed for a library and wild-type for two conditions (or phenotypes).

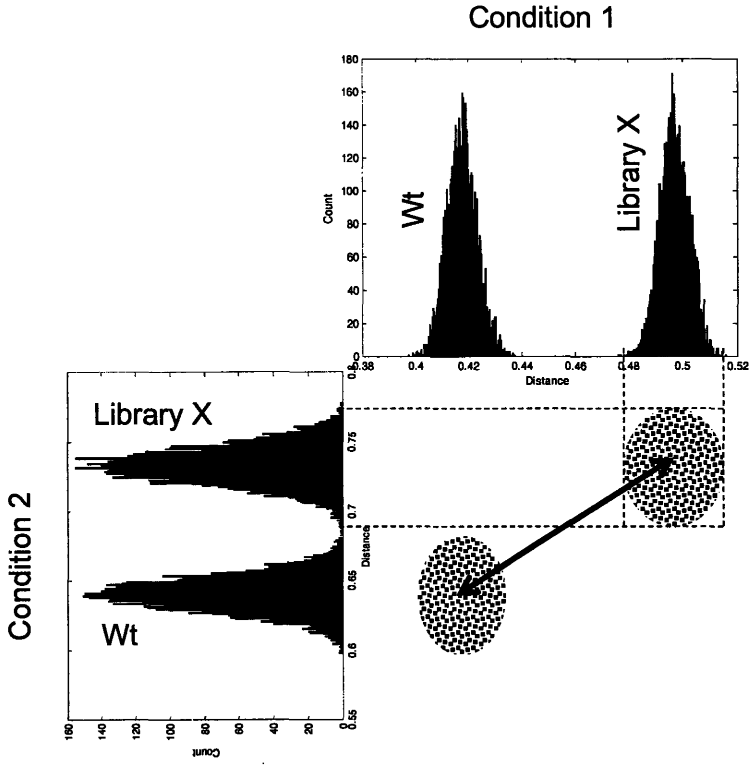

Figure 5.3-1. Schematic illustration of the divergence normalization method in 2-D

The distribution of average phenotypic distance, *d*, is shown in each axis for a generalized library (library X) and wild-type, clonal control (Wt). The dotted circles represent these histograms in 2-D. The "divergence" is the average phenotypic distance of a library population compared to that of the control. In the diagram, the divergence is represented by the double-headed arrow, except that the statistical measure used (Bhattacharyya distance) accounts also for the dispersion of the distributions (i.e. the "diameter" of the dotted circles).

To illustrate why it is important to account for the dispersion of the distribution when calculating the divergence, consider the case where the standard deviation of the two distributions in **Figure 5.3-1** increases while the mean is unchanged (i.e. the dotted circles become larger but remain centered around the same mean). As this happens, the difference between the distributions diminishes, yet the distance between the means does not. The Bhattacharyya distance between the library and control vectors is thus a statistically relevant measure of the divergence between these populations.

5.4 Quantification of Divergence Using Growth in Solid Media

The first system in which we chose to test the aforementioned ideas was in sigma D libraries of *L. plantarum*, described in Section 4.2.2. We chose growth rate as the phenotype for measuring diversity for two reasons. First, because it is a complex phenotype (i.e. dictated by many factors) of practical importance. Second, because it can be readily determined with high throughput by measuring colony area. This is possible because *L. plantarum* was shown to form round, very regular colonies when plated in MRS medium.

5.4.1 *Validation of Growth as a Phenotype for Diversity Quantification*

Before applying the experimental protocol to library evaluation, its efficacy in distinguishing colonies of various sizes was assessed to ensure that growth rate is a reliable phenotype for diversity quantification. This test was based, for the same reasons alluded to earlier, in the concept of clustering. We reasoned that if two clones that are different with respect to the phenotype to be measured can be reliably clustered together when mixed, then that phenotype would be a reasonable choice for diversity quantification.

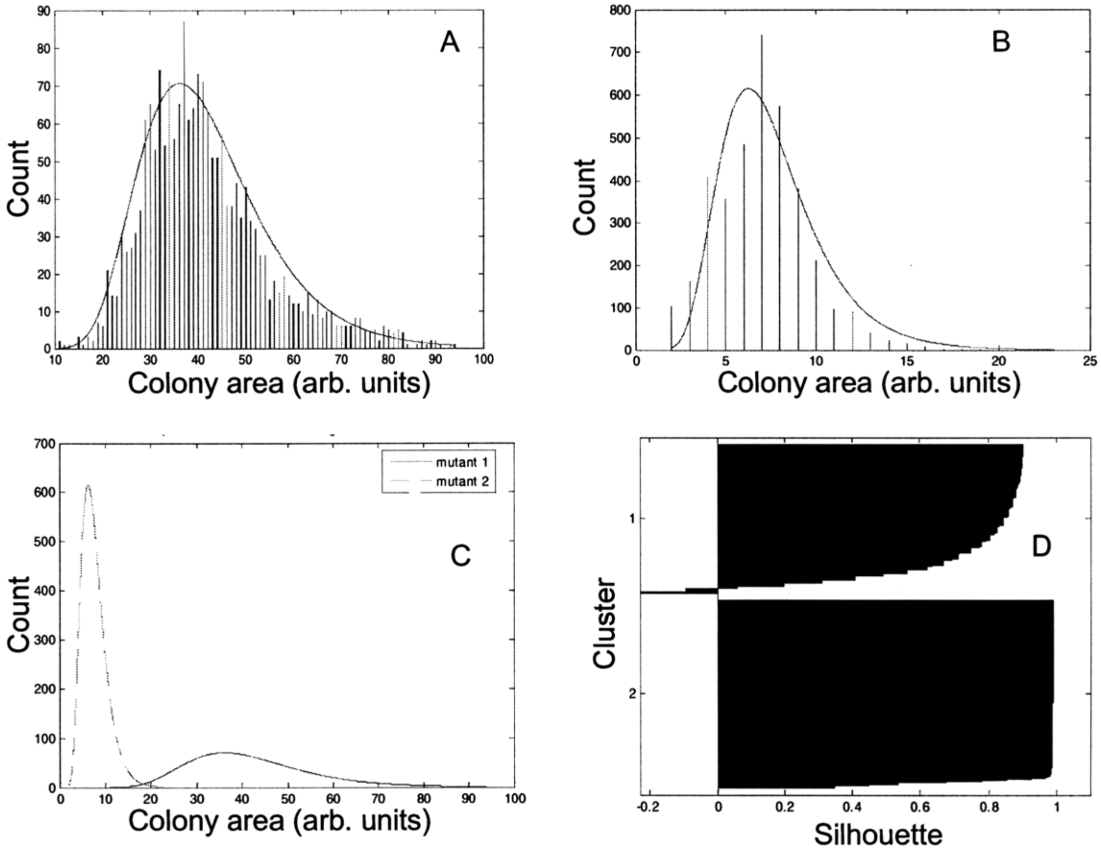

Figure 5.4-1. Validation experiments for growth as a phenotype for diversity quantification

(A,B) Colony size distributions of two mutants plated separately. The smooth line was constructed by fitting the histogram to a lognormal distribution. (C) Colony size distributions of the same two mutants of A and B plotted on the same graph for easier appreciation of the difference in colony size between them. (D) Silhouette (clustering) analysis for a mixture of mutants 1 and 2. The silhouette value (Eq. 4) for each colony is a measure of how similar (phenotypically) that colony is to colonies in its own cluster compared to colonies in other clusters, and ranges from -1 to +1. The plot shows that the populations may be clearly separated in two clusters with members of large silhouette values (most are larger than 0.5).

Two sigma factor mutants that formed colonies of different sizes were plated either separately or in a 50:50 mixture and the colonies were photographed and analyzed. As can be seen from **Figure 5.4-1** - A and - B, the areas of clonal populations of both mutants follow the same distribution (roughly lognormal), but their means are very distinct (**Figure 5.4-1** - C). The data generated by the mixed population was also analyzed to determine whether a clustering algorithm could separate the two clones based on colony size. **Figure 5.4-1** - D shows the silhouette value for the two clusters, a normalized measure of how similar is each individual colony to members of its own cluster compared to members of the other cluster; in general, this value ranges from -1 (for misplaced objects) to +1 (for accurately placed objects). Most silhouette values in our analysis are close to +1, indicating that the clusters are tight and that this method can be used to distinguish mutants based on colony size. The silhouette value was calculated for each object *i* using

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \qquad \text{(Eq. 4)}$$

Where b_i is the minimum of the average distances of object *i* to all objects in other clusters and a_i is the average distance of object *i* to objects in the same cluster (in this case, object *i* is the value P_i for colony *i* and a cluster is a group of colonies that have similar values of P_i).

5.4.2 *Determination of Divergence Using Growth*

The phenotypic diversity of four different populations was initially quantified (including the control). Three libraries of the principal sigma factor (*rpoD*) of *L. plantarum* were constructed by error prone PCR (epPCR) differing in their average mutation frequency as explained in Section 4.2.2. More than 4000 clones from each library were plated in different conditions and their colony areas were analyzed.

The phenotype used for quantification was actually the logarithm of the colony area, as its values were observed to be log-normally distributed. Then,

$$P_i = \ln A_i \qquad \text{(Eq. 5)}$$

The log-normal distribution could have at minimum two explanations: (i) the colonies grow exponentially, with a normally-distributed rate constant; or (ii) the colonies grow linearly, as when one assumes that the expansion is only due to growth of cells at the periphery (Panikov *et al.*, 2002), with a log-normally-distributed lag phase. Since the study of colony growth dynamics was not the focus of our work, we did not investigate this particular issue further, though we did find it interesting.

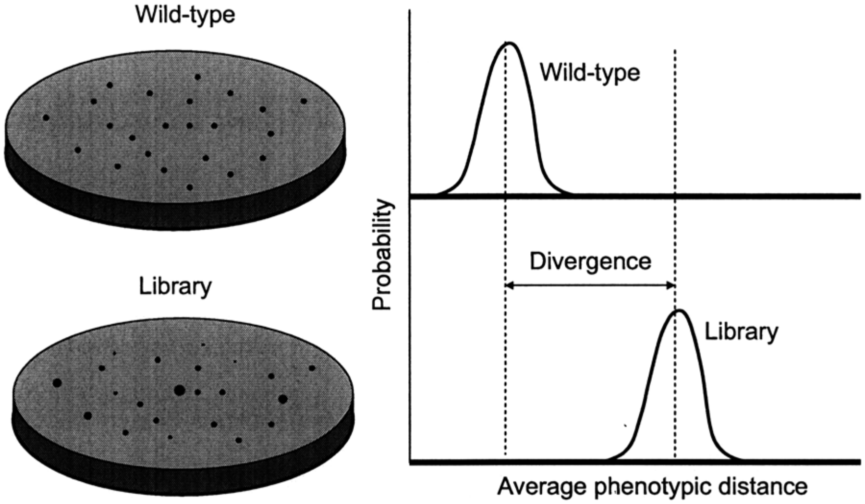

Figure 5.4-2. Experimental procedure for phenotypic diversity quantification

Populations (libraries or controls) are plated, photographed, and analyzed (left panel). Homogeneous populations have small average phenotypic distances, while diverse populations have larger ones (right panel). The illustration shows average phenotypic distances as distributions that result from the bootstrapping algorithm. The divergence is schematically shown for this one-dimensional case (one plating condition) as the separation between these distributions.

First, the average phenotypic distance was calculated for each population in the non-stressful condition; a larger distance implies a larger phenotypic dissimilarity among members of a population (schematically shown in **Figure 5.4-2**). The original size of each library (>10^5) is significantly larger than the number of clones that were analyzed, so the calculated average phenotypic distance (*d*, in Eq. 1) is in fact a sample average (as discussed in Section 5.3).

In order to assess the statistical significance of this metric, the value of the average phenotypic distance was bootstrapped to obtain its distribution (to input into Eq. 3). The bootstrap algorithm (Efron & Tibshirani, 1993) involves re-sampling (with replacement) the population and calculating the value of *d* for every such sub-population according to Eq. 1. Thus, the result is easily displayed in a histogram that reflects the probability that the "true" average phenotypic distance has a certain value (**Figure 5.4-2** and **Figure 5.4-3**).

Figure 5.4-3 shows the distributions obtained in this manner for the three libraries of varying mutation frequencies and wild-type control under non-stressful conditions. The graph shows that increasing the mutation frequency increases the average distance between the members of the populations. Furthermore, this increase of phenotypic diversity as a function of sequence diversity is statistically significant (there is little overlap between the distributions). The procedure was repeated for each of the three libraries (low, medium, high) and control, plated under each of three stress conditions (low pH (adjusted with HCl), osmotic/salt stress, and lactic acid) to determine whether this trend is also followed under stress. Because *L. plantarum* has different transcriptomic responses to these stresses (Pieterse *et al.*, 2005), plating in these conditions would contribute additional information to the diversity metric.

Then, we used Bhattacharyya's distance to calculate the divergence according to Eq. 3. Recall that because in each of the 4 conditions (no stress plus three stress conditions) the average phenotypic distance has a distribution (introduced by bootstrapping), this normalization must account for both the mean and the dispersion of the distributions. The results are shown in **Figure 5.4-4**.

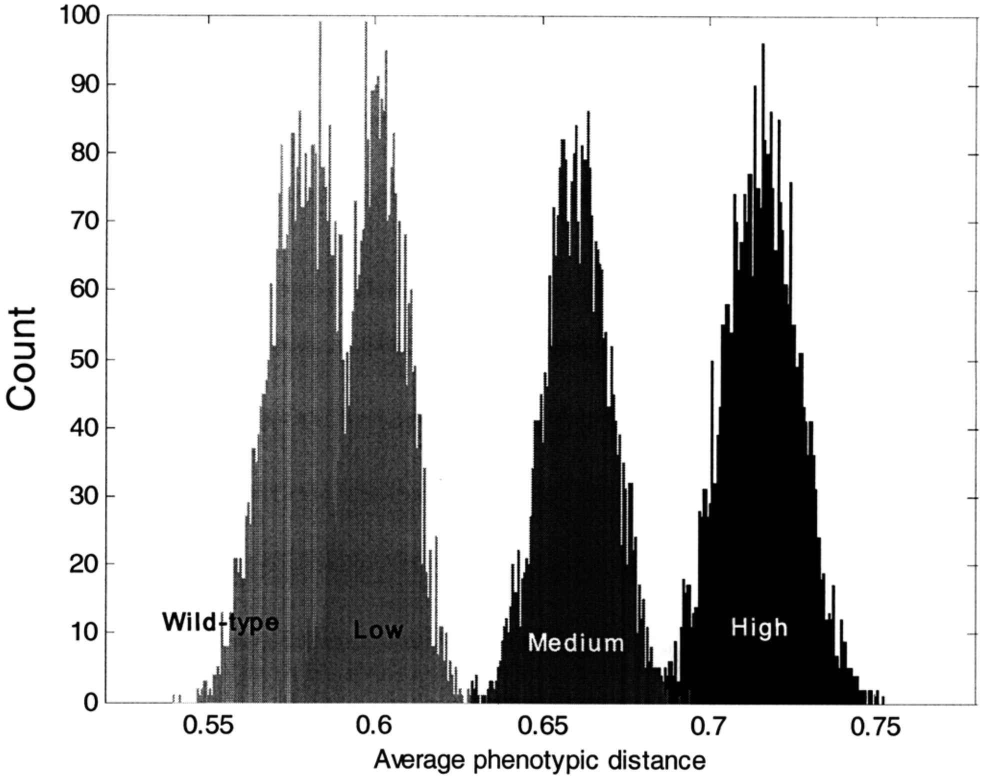

Figure 5.4-3. Distributions of the average phenotypic distances

The average (*d*, Eq. 1) of populations with different mutation frequencies (low, medium, high) is shown. These libraries and wild-type were plated under non-stressful conditions, photographed and analyzed to calculate their average phenotypic distance. The histograms were obtained by bootstrapping and reflect the probability that the "true" average phenotypic distance has a certain value.

5.4.3 *Comparison between Libraries Constructed via Transcriptional Engineering and Classical Strain Improvement*

Finally, we carried out the same analysis for a population that had been mutagenized using NTG (N-methyl-N'-nitro-N-nitrosoguanidine). This reagent is widely used in classical strain improvement to create cell diversity by introducing random mutations in the chromosome (see Section 3.1.1.1); it is regarded as one of the most effective chemical mutagens (Parekh *et al.*, 2000). One of the main goals of this part of the thesis was to compare the potential of libraries constructed by transcriptional engineering to those obtained by this well-established method.

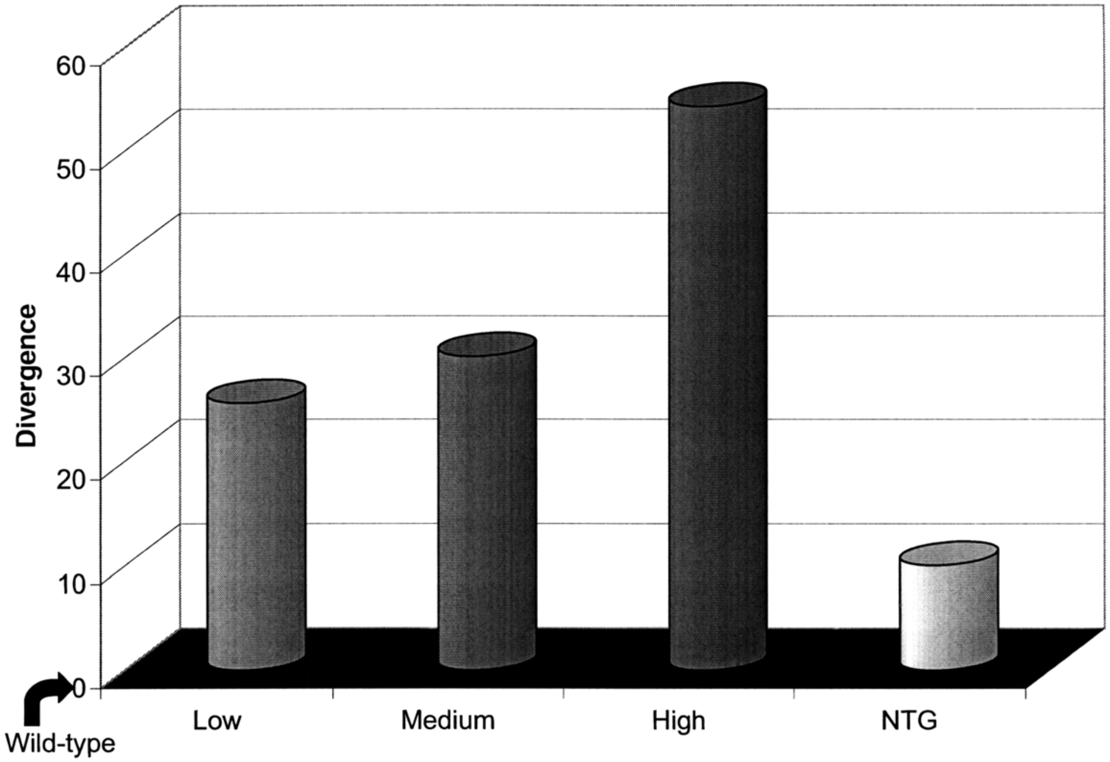

Figure 5.4-4. Divergence of the Lactobacillus libraries

The bars show the divergence from the unmutated control (which has zero divergence, as indicated), calculated using Bhattacharyya's equation (Eq. 4). This value is a statistical measure that describes the additional phenotypic distance of the libraries compared to that of the wild-type, and summarizes the information obtained from the four conditions considered in this study.

The results of the analysis are summarized in **Figure 5.4-4**, depicting the divergence given by Bhattacharyya's equation for each of the above four libraries. Because divergence is a normalized metric, the value for the unmutated control is zero, by definition. The graph shows that our NTG-mutagenesis library diverged less from the control than any of the sigma factor libraries. In other words, more phenotypic diversity was introduced by mutagenesis of the sigma factor than by genome-wide mutagenesis with NTG (at 40-50% killing).

5.5 Correlation of the Divergence Metric with the Probability of Finding an Improved Mutant

To investigate the predictive power of our diversity quantification method for improved phenotypes, we tested whether there was a correlation between our diversity metric and the probability of finding an improved mutant in a new stress. We screened each library a total of 192 times in 96-well plates under malic acid stress and calculated the probability that a screening event (i.e. a well) was successful (i.e. grew better than the control). We scored as "improved" the wells that exceeded the maximum OD found in the wells with wild-type cells (**Figure 5.5-1**). The results parallel the findings of the diversity metric indicating that improved diversity increases the probability of isolating mutants with improved phenotype.

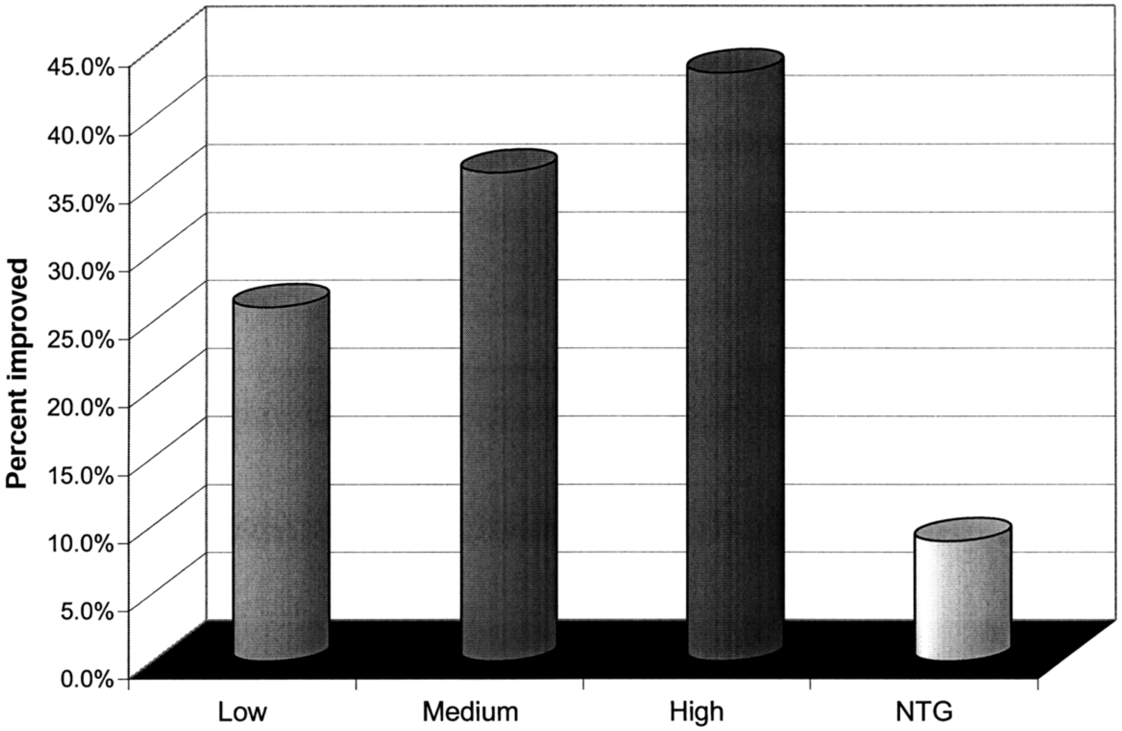

Figure 5.5-1. Estimate of the probability of finding novel phenotypes

The plot shows the percent improved wells with respect to the maximum wild-type OD reached under malic acid stress. Each library (and control) was screened under stress in a 96-well plate in two independent experiments, for a total of 192 times. The percentage represents the fraction of 'successful' screening events (wells) and is a measure of the probability of finding improved mutants in a population.

5.5.1 *Implications*

The results of the analysis support the following conclusions: (i) that mutations in the sigma factor allow introduction of phenotypic diversity; (ii) that this variability increases with mutation frequency; (iii) that localized mutagenesis of the sigma factor enhances diversity better than NTG mutagenesis of the entire genome (at 40-50% killing, following previously-established practices (Miller, 1972)); (iv) that the increased diversity is observed in different conditions (i.e. the mutations are pleiotropic); and (v) that diversity was correlated to the probability of finding improved mutants.

Conclusion (i) implies that the sigma factor is a good target for phenotypic alteration, and could have been derived, qualitatively, from the successful use of sigma factor engineering for isolating improved strains (see discussions throughout Chapter 4). Now, we confirmed it using a quantitative method.

These results are taken one step further by conclusion (ii). The increase in phenotypic diversity with *rpoD* sequence diversity argues that conclusion (i) is not accidental. Furthermore, these results echo previous studies in protein engineering (Drummond *et al.*, 2005), which establishes that it is possible to tackle strain improvement problems with protein engineering approaches, and suggests the use of other protein engineering techniques to optimize transcriptional engineering. Optimization would be especially useful when considering that as the mutation frequency of the library increases, so will the fraction of nonfunctional, misfolded mutants. As this fraction increases, the phenotypic diversity is expected to decrease, as more mutants will exhibit a phenotype that results from expressing misfolded proteins (i.e. upregulation of chaperones, proteases, etc. (Jurgen *et al.*, 2001)). An example of such a trend was observed later in the thesis, as discussed in Section 6.2.

Conclusions (iii) and (iv) have important practical consequences; the former because it establishes that targeting the global transcription machinery compares favorably to traditional techniques for evolving new strains, the latter because it evidences the versatility of this approach. Such versatility and the results that led to conclusion (v) imply that libraries of transcriptional regulators can be screened in multiple conditions.

The diversity quantification method presented can aid in prioritizing the screening efforts when libraries of other regulators are considered. Furthermore, the resulting metric can be used for optimizing and refining existing or novel strain improvement methods (see Chapter 6).

We found a good correlation between the diversity of the libraries and the probability of finding improved clones. In the condition tested, we found that it was about 5-fold more probable to find improved mutants in the sigma factor library with high mutation frequency than in the NTG-mutagenesis library, similar to the fold improvement in the divergence value. These results suggest that our diversity quantification method allows categorization of libraries according to a quantitative metric, which can be used for predicting the outcome of a strain improvement effort.

Even the highest probability of success was not extraordinary (~45%), but this is likely related to the strict definition used for scoring 'successful' screening events (one where the OD was higher than the maximum OD for the wild-type). 'Unsuccessful' screening events may also have improved mutants, but maybe in such low concentrations that they could not overtake the population in a single round of culturing.

The fact that mutagenesis at the chromosomal level was less effective than localized mutagenesis of the sigma factor may seem counterintuitive. After all, it may be argued, mutations in the sigma factor are a subset of the possible mutations in the chromosome. A similar argument was presented in Section 5.1.3, and then refuted based on the fact that favorable changes are diluted when the sequence space is too large (more on this below). Additionally, because the mutations are introduced in an additional copy of the sigma factor, we are effectively evolving an "alternative" sigma factor that confers the improved response. This mimics the process of gene duplication and function specialization that may have led to naturally-occurring alternative sigma factors (Errington, 1991).

The preceding discussion leads to the following question: is there an inherent advantage of targeting sigma (or other regulatory proteins) for phenotypic alteration? The results of the aforementioned experiments argue so, but previous findings are also closely in tune with our work. Transcription factor-binding variation has been recently found to supersede gene variation in closely related species, indicating that rapid phenotypic specialization is largely due to changes at the gene regulation level (Borneman *et al.*, 2007). Therefore, our comparison of the phenotypic diversity in sigma factor versus NTG-mutagenesis libraries seems to describe a natural mechanism of evolution. Most probably, mutations introduced by NTG are diluted in the large sequence space of the genome, so that it is uncommon that enough beneficial changes accumulate and cause a significant improvement in phenotype. On the other hand, mutations in the sigma factor may explore the "regulatory space" more efficiently.

Because we are targeting *trans*-acting regulatory mechanisms, a high degree of pleiotropy is expected. This was indeed observed, and it implies that small changes in these targets introduce profound phenotypic changes. As such, these preliminary results suggest that the principles of transcriptional engineering are fundamentally useful for whole-cell directed evolution, and that the diversity quantification method presented may allow optimization of this and similar approaches (see Chapter 6).

5.6 Other Complex Phenotypes for Quantification of Diversity

Growth, as measured by colony area, proved to be a reliable and uncomplicated phenotype to measure diversity, given the colony morphology displayed by *L. plantarum* in solid media. When we tried to adapt the method for *E. coli*, we observed a pronounced irregularity in colony sizes and shapes, which made the protocol for quantifying the colony area hard to replicate in this species. We concluded that other phenotypes would be needed. Ideal candidates should be complex, so that they can be used as a proxy for widespread transcriptomic changes, and should be easily measured in a high-throughput fashion. Flow-cytometry was deemed an ideal experimental platform, because of the wide variety of fluorescent dyes that are currently available and because single cells can be probed. Several complex phenotypes that can be studied with flow cytometry were considered.

For example, the dye 1,6-diphenyl-1,3,5-hexatriene (DPH) can be used to measure membrane fluidity, which changes in response to environmental cues such as osmolarity, temperature, the presence of solvents, and others (Gantet *et al.*, 1990; Muller *et al.*, 2000). The probe intercalates in lipid membranes and anisotropy measurements allow determining the relative ease with which DPH molecules can move when in the membrane (Bernal *et al.*, 2007; Bock *et al.*, 1989). Because membrane fluidity depends on the composition of the membrane (e.g. degree of lipid saturation, polar group in phospholipids, etc.) and in its adaptability when extracellular signals are present, DPH fluorescence can be amounted to a complex phenotype for diversity quantification. The method can be effortlessly adapted for use in flow-cytometry, so that individual cells are monitored and their distribution determined.

Another example, and one that was repeatedly used in the present thesis, is intracellular pH (pH_i). The pH_i can be used for quantification of diversity because it is affected by the relative levels of proteins and metabolites in the cell even when it is maintained in a narrow range (Kresnowati *et al.*, 2007). In addition, there are several probes for quantification of the pH_i, both in growing and non-growing conditions. In Section 4.4.1.3 we discussed how the pH_i was measured for assessing the increased tolerance of our solvent tolerant mutant, L33, to ion leakage through the membrane. These measurements were performed using a pH-sensitive variant of GFP (Miesenbock *et al.*, 1998). For quantification of phenotypic diversity based on pH_i, we chose a combination of carboxyfluorescein ester (CFSE) and 2’7’-bis-carboxyethyl-5,6-carboxyfluorescein ester (BCECF-AM) (Franck *et al.*, 1996; Spilimbergo *et al.*, 2005). The GFP variant was not used in an effort to minimize the burden on the metabolism of the cell when studying the different libraries. Both the small molecule dyes (CFSE and BCECF-AM) allow ratiometric quantification of the pH, so that the amount of probe per cell is normalized implicitly.

The divergence metric was calculated using the ratio of emissions (*E*) at the two different wavelengths (λ_1 and λ_2) as explained in Section 9.3.8. The phenotype for quantification is defined as

$$P = \frac{E_{\lambda 1}}{E_{\lambda 2}} \tag{Eq. 6}$$

5.6.1 *Validation of Intracellular pH for Phenotypic Diversity Quantification*

Before using pH_i for phenotypic diversity quantification, we tested our staining and probing protocol for distinguishing cells with changes in transcriptome. These changes were introduced by growing the cells in different media, as nutrients are common cues to reprogram metabolic pathways. We compared cells grown in M9 minimal media with glucose to cells grown in either rich media (LB) or M9 media with glycerol instead of glucose.

After cells were grown (to approximately the same OD), they were washed twice with PBS to make sure differences in pH_i were not due to extracellular effects. Cells were stained with BCECF-AM as explained in Section 9.3.8, and washed again with PBS. Finally, cells were resuspended in PBS with glucose as an energy source, to allow cells to equilibrate their pH_i. Fluorescence was measured in a spectrophotometer with the same parameters as before and the ratio calculated.

Cells grown on glucose could be distinguished from those grown in LB with a p-value of 0.03, and they could be distinguished from those grown in glycerol with a p-value of 0.05. The difference in transcriptome is expected to be much wider in the first comparison, and thus is not surprising that it corresponds to a smaller p-value.

5.7 The Relative Nature of the Divergence Metric

The divergence metric for a population is not an absolute number, even when the same phenotypes are used to compute it. The reason is that the phenotypic diversity of a population is sensitive to the experimental conditions, and therefore, it is essential that great care is taken for recording the specific parameters of a protocol. For example, for the case of using colony sizes of *L. plantarum* described in Section 5.4, the high salt condition resulted in average phenotypic distances that were significantly larger than for the lactic acid condition. This was true for all libraries tested, as well as for the control, and it can be regarded as inherent to our experiment; this difference in variability explains why normalizing the values of the libraries with that of a clonal, wild-type population is crucial. Normalization is also vital in order to account for imperceptible day-to-day fluctuation in experimental conditions, yet it cannot deal with all sources of variation. As a result, and in order to make statements about the evolutionary potential of different populations, one must always compare similarly-prepared libraries, and carry out the experiments for them in parallel.

The relativity of the divergence value is particularly pronounced when the phenotype used for diversity quantification is itself relative. For example, the fluorescence reading of a flow cytometer can be tuned with the dials of the instrument, and its values have no physical significance in themselves. In fact, the reading of a cytometer is given in arbitrary units. Because flow cytometry produces single-cell data with high-throughput and many complex phenotypes can be characterized with this technique, the limitation of a relative measurement is relevant to many situations of interest. Normalization allows foregoing a linear calibration (e.g. as in the case of pH_i), but the variability of the fluorescence value will be a function of the cytometer's parameters and protocol specifications. These changes in variability translate to changes in divergence, and the disparity in the divergence values that result from different experiments can be quite marked.

One must keep in mind, therefore, that it is the trends given by the divergence metric that are most informative in our efforts. These trends were found to be very reliable and constant even when experimental parameters were altered. When a result seemed to contradict previously-established trends, or when there was a nonsensical divergence value in our analysis, it was always possible to track the root to experimental errors. In that sense, it proved highly useful to check the quality of the data before using it to calculate divergence values, i.e., ensuring that the phenotypic distributions appeared well-behaved.

The intrinsic relativity of the divergence metric implies that it cannot be used to compare populations produced in different laboratories or at different times, which may be a limitation. However, as it will become obvious in the next Chapter, this must not present a problem for optimizing random strain improvement libraries. Let us now turn to the discussion of how the divergence metric developed in the present Chapter was used to optimize libraries of the targets found most useful in Chapter 4.

Chapter 6

6. Optimization of Transcriptional Engineering Libraries Using Divergence

Having developed a metric for comparing the evolutionary potential of different libraries, we are now in a position to answer more rigorously the question of whether transcriptional engineering can be improved. Recall that we are not interested in a particular phenotype, but we want to improve the tool in general. We have used, nonetheless, particular phenotypes throughout this thesis to evidence the practical significance of the tools we employed.

In this Chapter we focus on the two most promising targets for transcriptional engineering in *E. coli*: the principal sigma factor and the alpha subunit of the RNA polymerase (RNAP). Recall that the alpha subunit of the RNAP was established as another protein for performing phenotypic engineering in an effort to improve upon transcriptional engineering. We now turn to using the divergence metric to optimize this new target with the same goal in mind. Then, we apply the principles derived from this endeavor to the principal sigma factor.

Throughout this Chapter, one must be aware that, without the divergence metric we would not be able to make any statements about the quality of the libraries as we change their design, and therefore, it would not be possible to call the effort "optimization." Let us then turn to defining the term optimization in the context of strain improvement programs, and to describe the use of the divergence metric for guiding the construction of random libraries.

6.1 Optimization as a Reduction of the Search Space

Optimization, loosely defined, is the attempt of finding the best solution among all the feasible solutions of a particular problem. In order to undertake such a task, we must have a way of quantifying the progress we make towards the best solution. To solve an optimization problem, even partially (i.e. when the absolutely best solution is not required or attainable), we should therefore be able to evaluate the present and past states. Analogous to this formulation, we now aim at finding better strain improvement libraries, and thus, we need a method for evaluating our progress. The divergence metric can be used for this purpose, since it allows quantification of the potential of different populations and thus permits comparing the past and present states (as they apply to library design), albeit in relative terms. With these arguments in place, optimization of random search-based libraries becomes an effort of sequentially designing libraries, evaluating them using the divergence metric, and altering the design in the hope that the divergence of the new library increases.

Given the infiniteness of sequence space and the difficulty in finding interesting or improved variants, we can safely say that a large fraction of the members of a library have uninteresting phenotypes. This amounts to our earlier argument that useful mutations are essentially diluted in the vastness of the search space, so that improving traits of interest becomes an experimentally intensive task. We must therefore focus our search space to regions that can potentially increase the divergence. In different terms, optimization can be understood as sequentially delimiting the search space by ignoring genetic determinants that when altered result in phenotypically redundant variants, but keeping those that result in new phenotypes.

Another reason for the redundancy in phenotypes even in the presence of genotypic diversity is robustness. At a systems level, the cell must maintain key functions in spite of mutations or changes in the extra- or intracellular environment (Kitano, 2004). This implies that phenotypic diversity will tend to be damped by adjustments in the system as a whole that aim at keeping the healthy metabolic state of the cell. Since this tuning will show up in at least some changes to the intracellular environment, and since these same adjustments also allow evolution (Kitano, 2004), the general mapping between diversity and evolutionary potential should still hold as a general principle.

Now that we understand optimization as it applies to the improvement of random search-based libraries for strain improvement, we can illustrate this concept with an example. As we will momentarily see, the divergence metric does not only allow evaluating the present and past states with respect to evolutionary potential, but it also provides useful information that suggests how future modifications to a library design should be effected in order to increase the quality of the library.

6.2 Optimization of alpha Subunit Libraries for Improving Butyrate Tolerance

In Section 4.4, we proved the usefulness of the alpha subunit of the RNAP for improving complex phenotypes, such as butanol tolerance and the production of L-tyrosine and hyaluronic acid. At that point, we used three libraries in which the entire coding sequence of *rpoA* was mutated with three mutation frequencies, resulting in libraries rpoA*L, rpoA*M, and rpoA*H with low, medium, and high frequencies, respectively.

We were also interested in a butyrate-tolerant mutant, because this compound can be used to produce butanol (in a two-step fermentation (Tashiro *et al.*, 2004) or catalytic reduction (Gustafson *et al.*, 1989)) and propane (Fischer & Peterson, 2008), both of interest as renewable fuels. The toxicity of butyrate is thought to arise from dissipation of the pH transmembrane gradient, similar to other weak acids, although limited research has been conducted in this regard (Zigova & Sturdik, 2000).

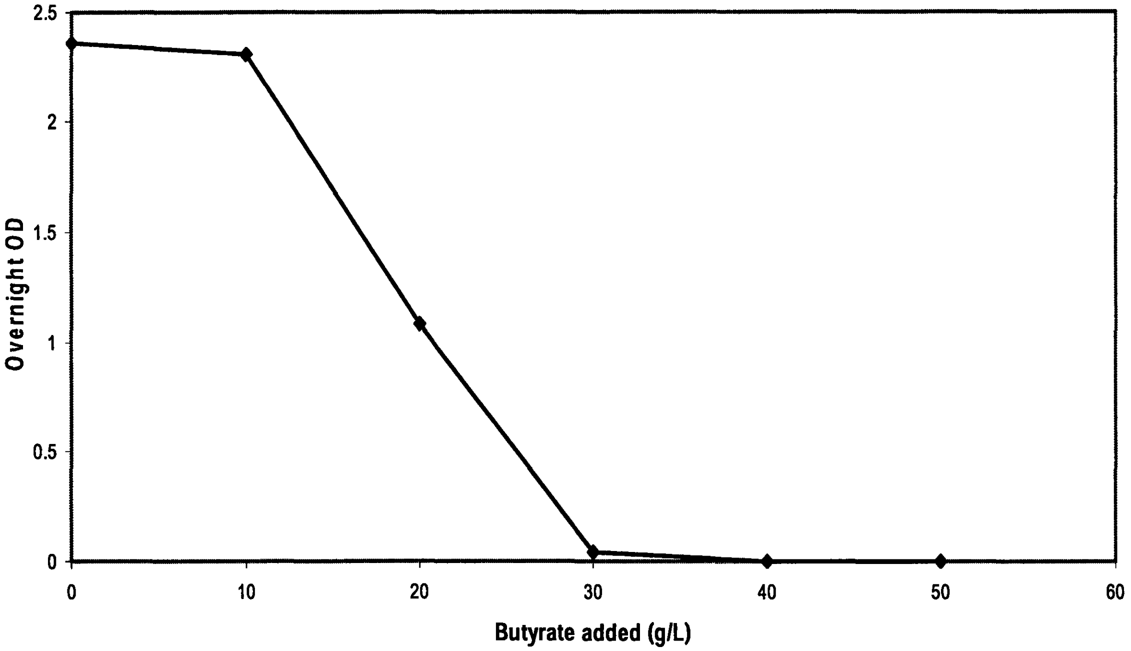

Figure 6.2-1. Test of toxicity of butyrate

The graph shows the growth inhibition of different concentrations of butyrate for *E. coli* in rich medium. The initial pH was close to neutral and an initial OD = 0.05 was used from a stationary-phase culture. The media was prepared from stock solutions of 2X LB and 150 g/L sodium butyrate.

In order to assess the toxicity of butyrate to *E. coli*, preliminary studies were performed in rich media (LB broth). As shown in **Figure 6.2-1**, there is a clear dependence between final growth and the initial amount of butyrate in the media. We also wanted to investigate whether butyrate has a bacteriostatic effect (i.e. it prevents growth) or a bactericidal effect (i.e. it kills the cells). **Figure 6.2-2** shows a killing curve of *E. coli* in minimal media supplemented with 30 g/L of butyrate and glucose as a carbon and energy source. For this study, the pH was adjusted to 5, close to its pK_a, which has a value of 4.8.

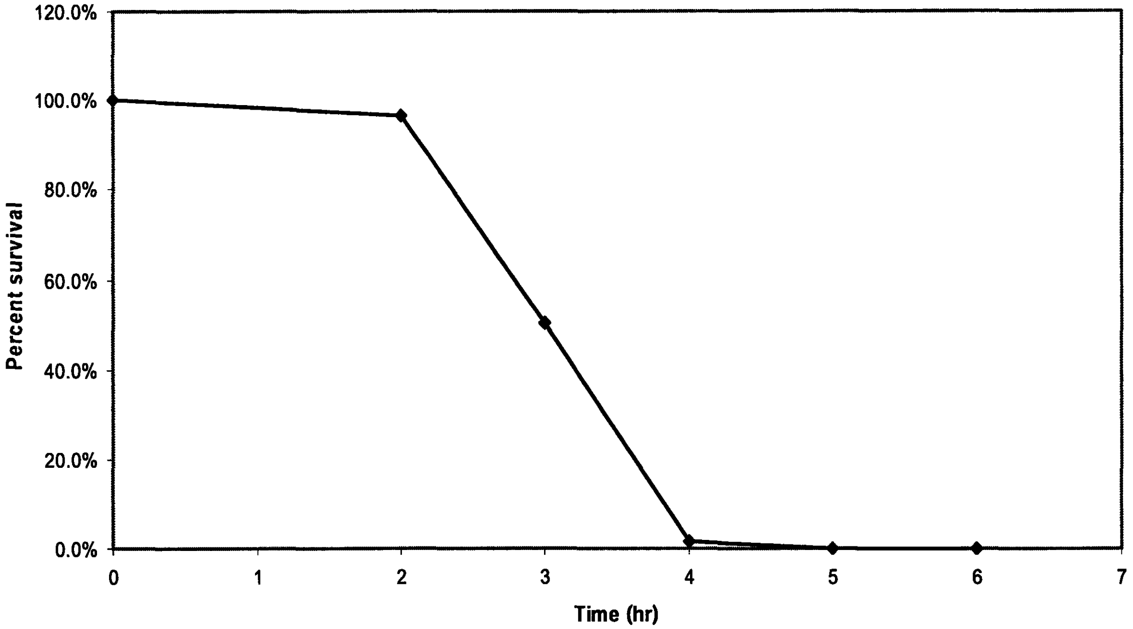

Figure 6.2-2. Bactericidal effects of butyrate

The graph shows the killing curve of *E. coli* with 30 g/L sodium butyrate, with an initial pH of 5. The assay was done in M9 minimal media, with glucose used as a carbon source. The killing rate was calculated by plating culture dilutions and counting surviving colonies per unit volume at different time points.

6.2.1 *Optimization of Random Search-Based Strain Improvement Programs*

The need for optimization of strain improvement libraries was not obvious in Section 4.4, in which we reported the successful use of the alpha subunit of the RNAP for finding traits of interest. In fact, undergoing an optimization program would make most sense when screening existing libraries fails to produce variants of interest. Recall that in the majority of library designs, we cannot cover the search space experimentally, which becomes a particularly relevant problem when screening for phenotypes of interest fails to deliver improved variants. In this case, the result of one experiment rarely suggests ensuing experiments, because it is difficult to ascribe the failure to particular steps of the random search protocol. This changes if we can evaluate and improve the libraries themselves.

When we screened the rpoA*L, rpoA*M, and rpoA*H libraries in butyrate, we were unable to isolate improved mutants, even when several screening conditions were tried (see next Section). With the divergence metric in place, we proceeded to formalize the idea of optimization as it refers to a strain improvement program. We can regard a "random approach for finding an improved mutant" as an iteration of two steps: building a library and screening it. Because screening is the resource- and labor-intensive step (Demain *et al.*, 1999; Kittell *et al.*, 2005), it makes sense to carry it out only if the expected outcome of the experiment is better than that of constructing a new library, that is, if the *a priori* probability of finding a good mutant is greater than it was in the previous iteration. This process can continue until constructing new libraries becomes expensive (e.g. for fully-synthetic libraries) or no obvious way of improving the library is available (e.g. by changing the mutation frequency, the localization of mutations, etc.).

6.2.2 *Utilization of Previously Constructed Libraries*

We began our quest for a butyrate-tolerant mutant by screening the three *rpoA* libraries with different mutation frequencies throughout their coding region that had been fruitfully used to isolate mutants (see Section 4.4). Several screening techniques were used, such as survivability in butyrate (based on the results of **Figure 6.2-2**) and serial subculturing.

We also altered the screening conditions, as these are known to influence the phenotype that is enriched. For instance, Warnecke and coworkers recently reported that a strategy that consisted in decreasing stress in serial batch cultures increased the selectivity and sensitivity of their selection (Warnecke *et al.*, 2008). Another example of a screening parameter that can be changed, and one that is rather intuitive, is culture volume. Reducing the volume per batch, and concomitantly increase the number of batches to keep the number of individuals to be screened constant, makes it easier to distinguish a culture with a mutant that grows faster than the background population. Taking these facts into account, we ran several selection experiments with different stress gradients (increase or decrease concentration of butyrate) and volumes per batch. We also altered other environmental parameters that can affect butyrate toxicity such as amino acid supplementation, pH, choice of buffer (M9 or MOPS), and level of oxygenation (aerobic vs. microaerobic cultures).

Figure 6.2-3 shows the results from four of the above listed conditions for each of the three libraries that had been previously screened for butanol tolerance and L-tyrosine and hyaluronic acid production. The graph shows the maximum recorded growth advantage of the libraries vs. the control populations, which gives a rough sense of the concentration of faster-growing variants during our selection experiment (i.e. the theoretical enrichment). Even though positive enrichments were recorded, not a single mutant was found that grew significantly faster than the wild-type when tested individually in the same conditions used for selection. Although a few improved mutants were sporadically found in the initial test experiments, all proved to be either due to background chromosomal mutations or adaptation, since the phenotypes were lost upon re-transformation.

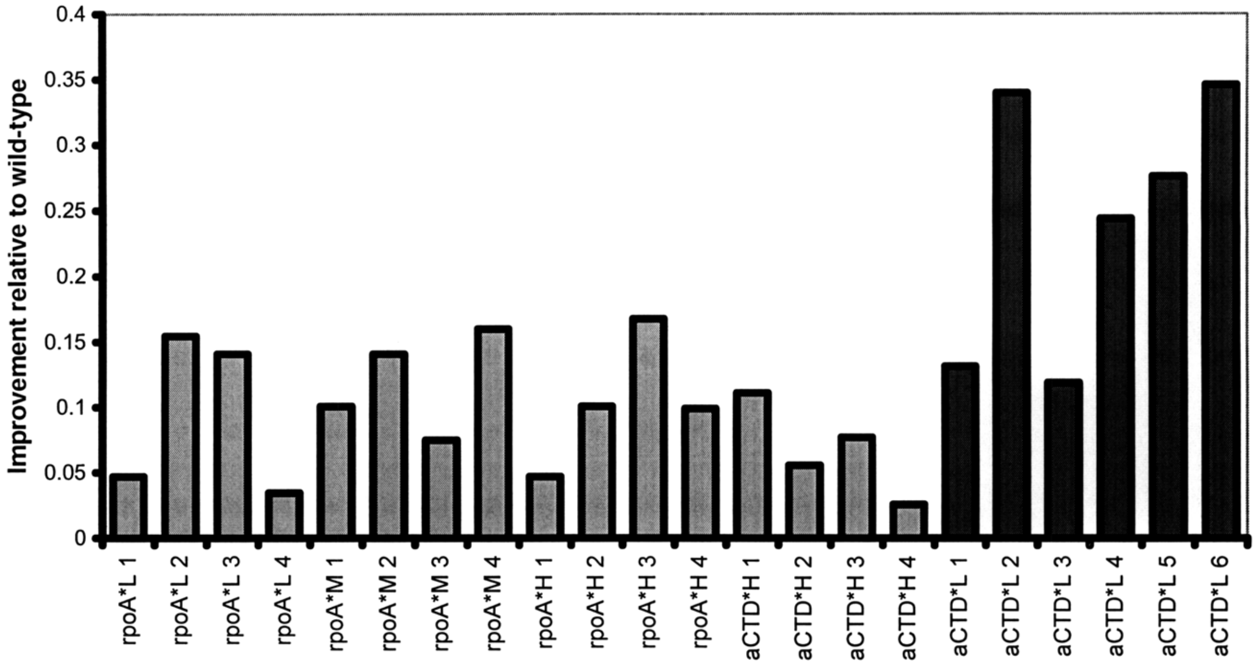

Figure 6.2-3. Selection experiments in butyrate

The graph shows the maximum recorded advantage in OD (600 nm) of cultures of the libraries relative to the control in different screening conditions, that is, the theoretical enrichment of improved clones. The conditions are: 1. M9 medium, 15 g/L butyrate throughout screening; 2. MOPS medium supplemented with amino acids (5%), decreasing butyrate concentration (18, 15, 12 g/L); 3. MOPS medium, 15 g/L butyrate throughout screening; 4. MOPS medium supplemented with amino acids, 15 g/L butyrate throughout screening. For αCTD*L, two repeats of the last set of conditions are given by runs αCTD*L 5 and 6. For rpoA*L, rpoA*M, and rpoA*H, some conditions were tried more than once (not shown) to rule out experimental error as the reason for not obtaining improved mutants (see discussion in the text). Even though a positive theoretical enrichment is shown for all cases, no improved mutant was isolated in any library except the αCTD*L, suggesting that transient advantages in OD of up to ~15% can be considered noise.

We hypothesized that either (i) the alpha subunit of the RNAP was not a good target for improving butyrate tolerance; (ii) our existing libraries did not contain mutants that showed such improvement even when this was theoretically possible; or (iii) the improved mutants were present in such low concentrations that they could not be sufficiently enriched by the selection experiments. If the true cause was either of the latter two options, an optimization program could increase the probability of finding a mutant.

6.2.3 *Improvement of Libraries by Reducing the Search Space to the C-terminal Domain*

We began by focusing on the existing libraries. Using pH_i, we quantified the diversity in the rpoA*L and rpoA*H libraries, which we had extensively screened in butyrate, albeit with no results (see Section 5.6). As shown in **Figure 6.2-4**, there is an increase in divergence when sequence diversity in *rpoA* is increased, but our inability to find improved mutants suggested that a new, more phenotypically diverse library was needed. Our previous study on the alpha subunit resulted in three improved mutants, all of which had nucleotide changes in the αCTD (see Section 4.4). Therefore, we hypothesized that diversity could be increased by directing mutagenesis to this region of the protein. We constructed a library in which this region was mutagenized with high frequency, after observing that highest phenotypic diversity is accomplished with extensive mutagenesis (**Figure 5.4-4**, **Figure 6.2-4**, and Sections below).

Quantifying the phenotypic diversity of the new library (denoted αCTD*H) contradicted our expectations (**Figure 6.2-4**). Not only did the diversity not increase by focusing the mutations to the αCTD, but it actually decreased. Although the prospect of finding an improved mutant in this library was low, we screened in butyrate to test our strategy. This screening step could have been eliminated if time was of essence or if the protocol was too costly. Four independent selection experiments confirmed our expectations; we were unable to isolate improved mutants, thus a new library was needed.

We thought of two possible explanations for the decrease in diversity in αCTD*H compared to rpoA*H: (i) that by focusing the mutations to this domain we lost diversity because mutations in the N-terminal domain (αNTD) also confer novel phenotypes (e.g. by modulating the assembly of RNAP complexes (Kimura & Ishihama, 1995) or by transcriptional regulation at class II promoters (Niu *et al.*, 1996)); or (ii) that the mutation frequency was too high, and that the diversity was lost because when a useful mutation was obtained, its effect vanished due to subsequent mutations. In other words, high mutation frequencies may reduce the diversity in our library because many clones display the same phenotype: that of expressing an alpha subunit with a non-functional CTD (a similar argument was briefly described in Section 5.5.1).

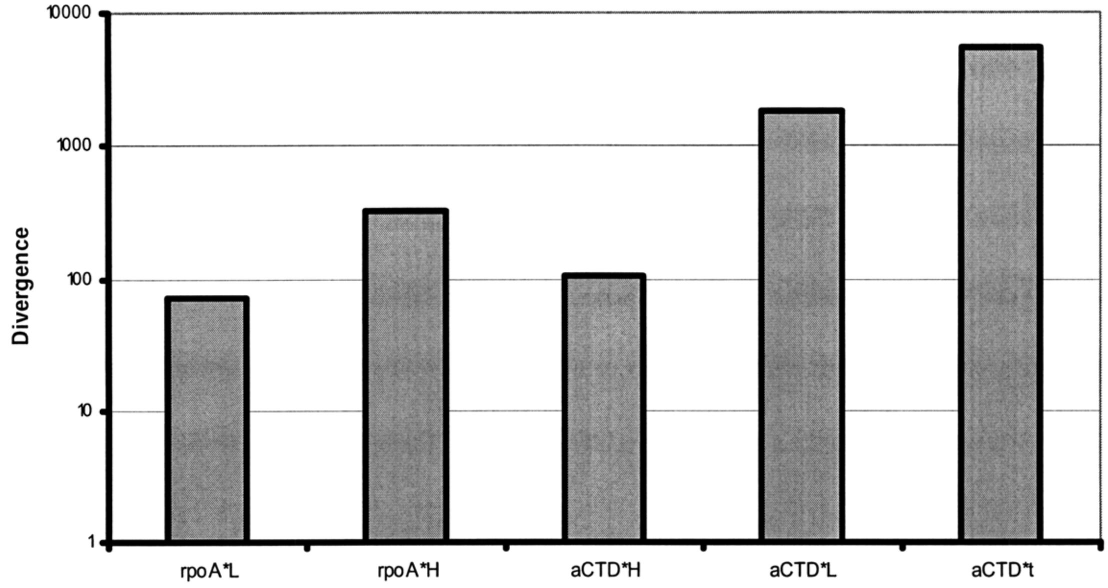

Figure 6.2-4. Divergence of rpoA libraries during optimization

The values shown here are based on intracellular pH as the phenotype for quantification both in growing and non-growing cells. Note that the divergence value is a relative measure and that it is used only for comparing different populations; thus, all the values shown were experimentally determined at once. Nomenclature: rpoA*L and *H are epPCR libraries of the entire coding region of the alpha subunit with low and high mutation frequencies; aCTD*L and *H are epPCR libraries of the CTD of alpha with low and high mutation frequencies; aCTD*t is a library in which amino acid changes are restricted to a few surface residues located in the CTD.

To test these hypotheses, we constructed a library in which the mutagenesis is focused on the αCTD, but with lower mutagenesis rate (denoted αCTD*L). Quantifying the diversity of this library favored the second hypothesis (**Figure 6.2-4**). This library has in fact higher diversity than that of the *rpoA* library with high mutation frequency throughout the coding region (rpoA*H). The mutation frequency in the CTD of rpoA*H is comparable to that of αCTD*L, but the latter has markedly more diversity. Thus, the most likely explanation for the diversity in rpoA*H is that it arises from changes in the αCTD in the context of an αNTD with transcriptional functions that are either not as important as those of the CTD or are sparsely found in sequence space.

6.2.4 *Isolation and study of Butyrate-Tolerant Mutants*

When we screened the αCTD*L library, now the library with the highest divergence, in the same conditions that were previously tried, we observed higher enrichments in some conditions (**Figure 6.2-3**). Following a similar experimental protocol than as with the other libraries, we selected a few dozen colonies for further characterization, isolation of the mutant plasmids, and re-transformation into a fresh background. From the resulting pool, two strains were selected. The mutants show a 23% and 40% improvement in growth rate in the presence of 15g/L butyrate (**Figure 6.2-5**).

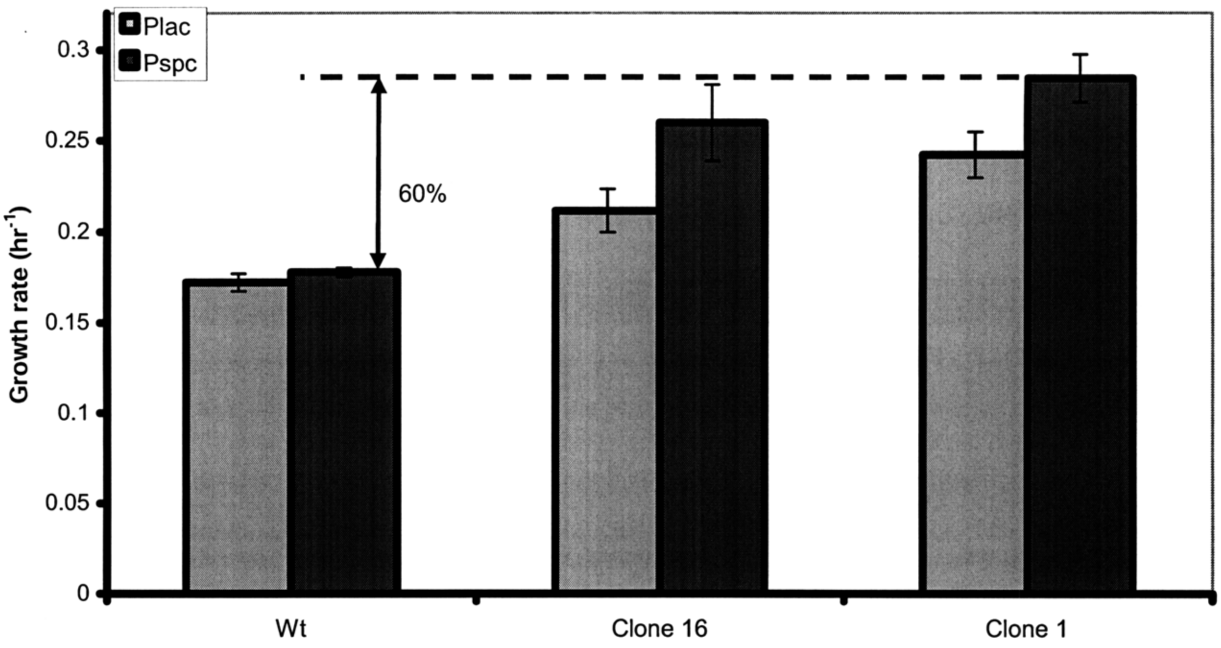

Figure 6.2-5. Growth in butyrate of wild-type and mutants

Growth rate of K12 *recA*⁻ transformed with wild-type or mutant versions of *rpoA* under the control of two promoters (*lac* and *spc*). Mutants #16 and #1 have the same amino acid sequence, but an additional synonymous mutation in #16 changes a common codon for glycine to a more uncommon one. As shown, increasing the expression level of the mutant (using P_{spc}) increases the growth advantage over the wild-type by up to 60%.

6.2.4.1 Sequence Analysis

The two plasmids isolated from these strains were sent for sequencing. Not coincidentally, the two mutants have the same amino acid sequence and only one amino acid change with respect to the wild-type (S299T), consistent with the diversity assessment suggesting that small changes in sequence in the αCTD result in large changes in phenotype. Amino acid S299 is directly involved in interacting with UP promoter elements (Gaal *et al.*, 1996); therefore, the mutation should alter the affinity of the RNAP for several targets, resulting in the novel phenotype.

The mutant with lower improvement (23%) differs from the mutant with higher improvement (40%) in a synonymous substitution that changes a codon that is frequently used in *E. coli* (GGT for glycine) with an unusual codon (GGA). In other words, a likely explanation for the change in growth rate given that the two mutants have the same amino acid sequence is that the difference in improvement stems from differences in protein level, with the one having the common codon (and the highest growth advantage) being expressed better.

6.2.4.2 Promoter Replacement

Since all of the mutants in our library are merodiploids for the *rpoA* gene, the new phenotype should arise from the interplay between wild-type and mutant alpha subunits, perhaps even in the same RNAP complex (recall that each complex has two alphas). Therefore, the relative level of the mutant vs. wild-type alpha protein should be a parameter that influences the observed traits. This was suggested by the codon usage difference between the two isolated mutants, but we wanted to test whether we could take advantage on this fact to increase butyrate tolerance further.

With this in mind, we placed the mutant and wild-type genes under a stronger promoter (P_{spc}) to see if we could improve the growth rate further. This promoter is a constitutive promoter, and it has been suggested to be the promoter that initiates most of the native *rpoA* transcription (Cerretti *et al.*, 1983; Post *et al.*, 1978). With this modification, we obtained an up to 60% improvement in growth rate (**Figure 6.2-5**). This advantage is substantial, considering that productivity of a metabolite in a continuous reactor is related to growth rate.

Our success with increasing the tolerance when increasing the promoter strength invited the use of a yet stronger promoter. No further increase was observed, however, when expression was placed under the P_{N25} promoter (Brunner & Bujard, 1987). Actually, a slight decrease in overall growth rate was seen, suggesting that the cell may be experiencing a burden when expressing the protein from the stronger promoter. Further improvements could have been achieved by manipulating the relative levels of mutant and wild-type alphas without changing the total amount of this protein. However, this would have entailed a lengthy effort, and we deemed it to be unnecessary to advance the main arguments of this thesis.

6.3 Relationship between Divergence and Posterior Probability of Finding Improved Mutants

The mutants we isolated were theoretically present in all the libraries that we had constructed before the aCTD*L library, given the parameters used to construct these populations (i.e. targeted regions, mutagenesis rate, identity of mutations, etc.). In other words, all library designs could have delivered the improved variants. This fact raises the question of whether the divergence metric actually reflects a probabilistic difference for finding the mutants in the different populations. Ideally, the metric would point in the direction of the population most likely to deliver the butyrate tolerant mutants. However, this generalized correlation is only strictly true when analyzing the results across many selections for a variety of traits, since the divergence metric is a measure of phenotypic diversity and not of the probability that a particular phenotype will be found in a population.

With this in mind, we analyzed the probability of finding the S299T mutant (the posterior probability) in the different libraries that we constructed, using information about the length of the fragment that was subjected to mutagenesis, the average mutation frequency of each library, and assuming that the mutations follow a Poisson distribution (Firth & Patrick, 2005). We assumed this distribution for simplicity, although more recent studies have reported that a modification to this formula is more accurate in certain cases (Drummond *et al.*, 2005). The assumption proved correct, as no significant discrepancies were found when the more intricate algorithm was used.

Table 5 shows that the S99T mutant could be found most frequently in the αCTD*L library, more than an order of magnitude more frequently than in any other library tested. It is important to note that this is the frequency of amplified PCR products at the DNA level, not the frequency in the cell library (see discussion in Section 3.2.1). The distinction is vital since different variants will be amplified with respect to others depending on their effect on growth rate in the steps prior to purifying selection; only then would the improved variants exhibit an advantage.

The table shows that the population with the highest phenotypic diversity had the highest probability for the improved mutant to be found, which implies that we did not find the mutant in the αCTD*L library accidentally. An even more compelling argument for the information contained in the divergence metric is the fact that all mutants that have been isolated up to date have 1 or 2 mutations in the αCTD. This argument is more in line with the definition of the divergence metric as reflecting the *a priori* probability, and not the posterior probability as calculated here.

Table 5. Posterior probability calculation for finding the improved mutant (S299T) in different libraries

	Bases subject to mutagenesis	Probability of having 1 mutation occurring	Probability of having the mutation in the right base	Probability of the change being the one required	Frequency of mutant (one in:)
rpoA*L	1300	7.33E-02	7.69E-04	0.33	5.32E+04
rpoA*M	1300	6.38E-03	7.69E-04	0.33	6.11E+05
rpoA*H	1300	1.11E-03	7.69E-04	0.33	3.51E+06
aCTD*L	250	3.58E-01	4.00E-03	0.33	2.09E+03
aCTD*H	250	1.49E-02	4.00E-03	0.33	5.04E+04

6.4 Exploiting Information Derived from Divergence for Constructing an Optimized alpha Subunit Library

6.4.1 *Design and Construction of the αCTD*t Library*

Although the goal of isolating an improved mutant had been achieved, we had gathered enough information to optimize our libraries further. Given that the diversity of αCTD*L is higher than that of αCTD*H, we hypothesized that this domain of the protein is very sensitive to mutations. Non-specific amino acid changes may prevent the αCTD from folding properly so that it cannot attain the conformation necessary for interacting with promoters. This suggested the construction of a library in which mutations were restricted to surface amino acids of this domain, thereby introducing diversity and at the same time preventing the formation of many non-functional, unfolded variants.

There was an obvious argument against such a design. Mutations in amino acids that function in key interactions may affect the phenotype to the point of lethality. This was in fact a problem with a similarly-designed library of the principal sigma factor in *E. coli*, in which mutations were located in DNA-binding regions. This design had resulted in small library sizes with little sequence diversity (both of which cause reduced phenotypic diversity; see Section 6.6). In other words, localizing mutations to vital regions of central regulators may be counterproductive (a similar argument was offered for leaving key conserved amino acids of γ-humulene synthase intact during protein engineering of this enzyme (Yoshikuni *et al.*, 2006a)); indirect changes in transcription factor function may be better at altering the transcriptome while preventing lethal effects.

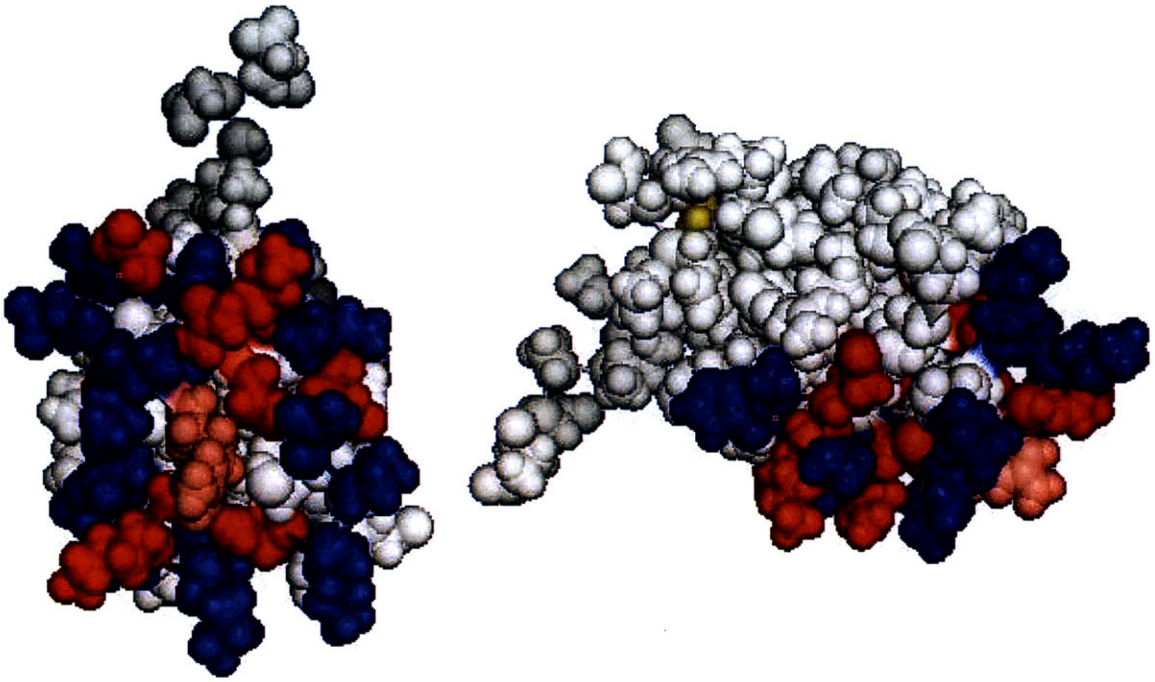

Figure 6.4-1. Design of the aCTD*t library

The C-terminal domain of the alpha subunit (from (Jeon *et al.*, 1995)) was modeled using PyMOL (DeLano Scientific LLC), and the surface amino acids (blue and red) surrounding amino acid R265 (in pink) were considered further. This amino acid is known to contact DNA (Murakami *et al.*, 1996), so it was chosen as a point of reference. From these, the red were selected as targets for mutagenesis.

We first constructed a library in which the chosen surface amino acids (12 in total) were mutated with fully-degenerate codons (**Figure 6.4-1**). The choice of amino acids was suggested by structural information (Jeon *et al.*, 1995) and previous studies (Murakami *et al.*, 1996) (see Methods for *E. coli*, in Section 9.3, for the list of mutated residues), but our selection is most probably sub-optimal. The experiment was considered proof-of-concept, and the exact design was deemed less important than the principles behind it.

We observed very low transformation efficiencies, probably because the average number of mutations per sequence was too high. **Figure 6.4-2** shows the approximate probability of a sequence in our design having a certain number of base pair changes (assuming each mutation is a Bernoulli trial). The number of amino acid changes (x-axis) is calculated assuming that a base pair change will be non-synonymous 70% of the time (Drummond *et al.*, 2005) (this may be a rough approximation, yet serves the purposes of the present argument). One curve shows the probabilities for the design in which the 12 amino acids are fully-degenerate. This design allows substitution of any amino acid with any other one, but results in a very high average number of mutations (17-21 amino acid changes are most frequent). The very low transformation efficiency probably reflected the fact that the presence of several mutations in these central residues is lethal, as mentioned earlier.

To address this concern, we investigated a different library design, in which bases at the chosen positions are spiked with non-wild-type bases with a certain probability. At first, we selected 6% as the probability of substitution with each of the non-wild-type bases, resulting in the second curve shown in **Figure 6.4-2**. This design yielded a reasonable library size (~40,000 clones), and reflected that lower mutation frequencies indeed produced viable variants. We quantified the divergence of this library, and obtained a marked increase in diversity (**Figure 6.2-4**). This result suggests (i) that surface amino acids in the CTD are indeed a good target for altering phenotype, and (ii) that the new design, with lower mutagenesis rate, still exhibited a useful degree of diversity.

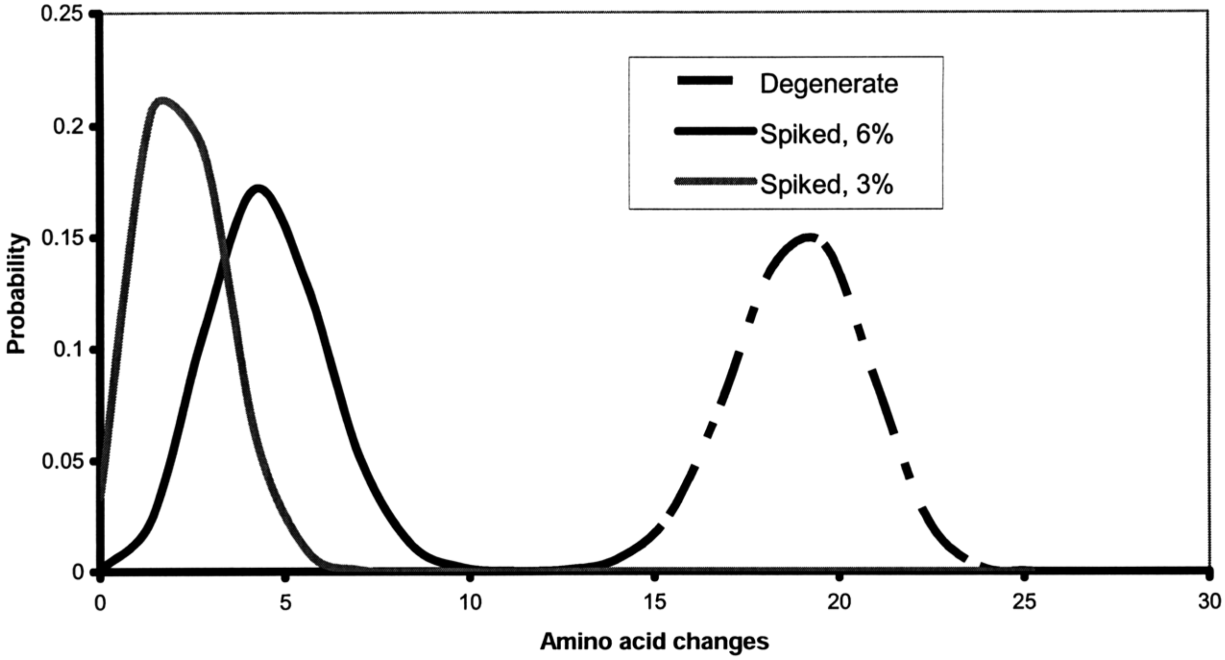

Figure 6.4-2. Distribution of amino acid changes in different aCTD*t library designs

The graph shows the probability to find a variant with a particular number of mutations, depending on the design of the αCTD*t library. The library has amino acid changes restricted to 12 locations, and base pair mutations are allowed either with degenerate bases (i.e. each base is found with probability 25% at each of the corresponding locations in the DNA coding sequence), or spiked (6% and 3% refer to the probability that each non-wild-type base will substitute the wild-type).

6.4.2 *Effect of Promoter Strength, Mutagenesis Rate, and Library Size in Divergence*

One main product of our optimization study was the αCTD*t library with 6% spiked non-wild-type bases (in the following discussion, we will refer to this library simply as αCTD*t), in which several surface amino acids in the C-terminal domain were targeted for mutagenesis. We hypothesized that further modifications to the library could increase the diversity it conferred. From Section 6.2.4 recall that, at least in some instances, changing the promoter strength that controls the expression of the mutant alpha subunit has an effect in the phenotype of interest. This was the case of butyrate tolerance, in which increasing the level of expression of the mutant *rpoA*s (by substituting the *lac* promoter with the *spc* sequence (Post *et al.*, 1978)) resulted in a concomitant increase in growth rate in the presence of the toxic compound. No further increase was observed, however, when expression was placed under the yet stronger P_{N25} promoter (Brunner & Bujard, 1987).

These observations raised the possibility that increasing the promoter strength in a library could have an effect in diversity. Thus, we constructed a new library under the control of the stronger *spc* promoter, quantified the diversity, and analyzed the aforementioned effect. (For these experiments, only the pH_i in the growing condition was used for diversity quantification). As shown in **Figure 6.4-3**, increasing the promoter strength does increase the diversity, suggesting that at the lower level of expression, some mutants do not have a perceivable effect in phenotype, as determined by pH_i.

Incidentally, during optimization of the experimental protocol, two libraries with different sizes were constructed, and thus the effect of library size could also be studied. As expected, a larger library has more diversity. The reason is that, even though two libraries with the same design have the same theoretical search space, any actual, physical library encompasses a subset of this space (unless in the rare cases where the search space is significantly smaller than the obtained physical library). A larger physical library contains a larger subset than a smaller one, resulting in a higher divergence value, as observed.

We also tested the effect of mutation frequency on diversity for the same library design by lowering the fraction of non-wild-type bases in the spike mixture to 3% (instead of 6%, see **Figure 6.4-3**). This experiment was suggested by the success of lowering the mutagenesis rate from our initial trial, one which used degenerate bases. An additional lesson from the optimization effort was that for the case of αCTD*t, a lower mutation frequency lowers the diversity, an outcome that has been repeatedly observed in libraries constructed via error-prone PCR (e.g. see Chapter 5 and previous Sections in this Chapter).

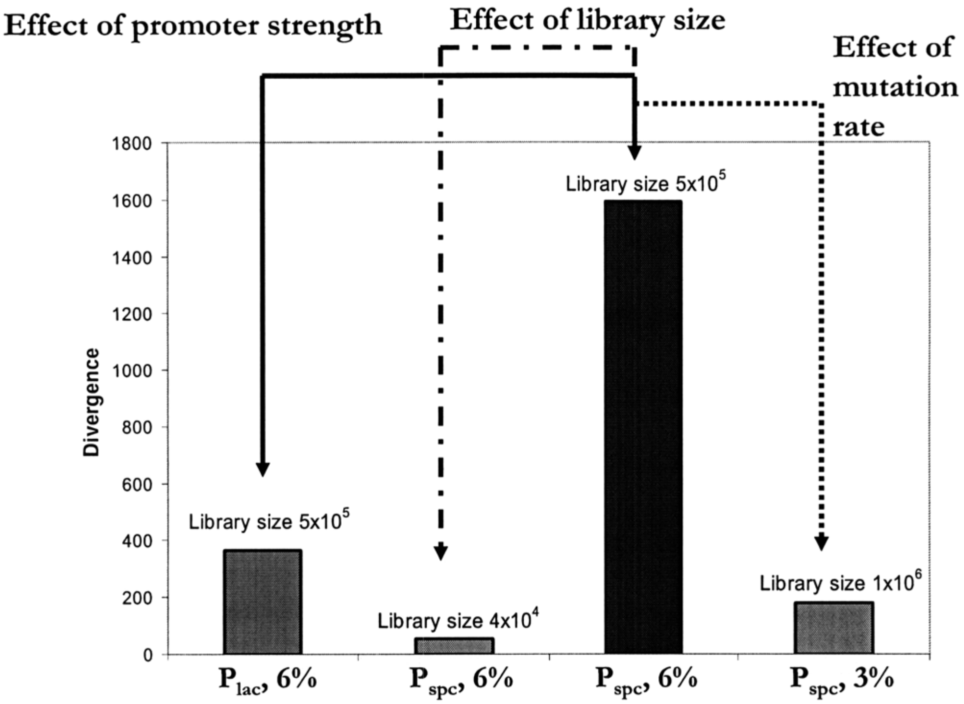

Figure 6.4-3. Simple variations to a library design

The graph shows the divergence value of different αCTD*t libraries, all of which have amino acid changes restricted to a few surface amino acids in the C-terminal domain of the alpha subunit. The x-axis shows the promoter (*lac* or *spc*) and the mutation frequency per base change (i.e. 6% means that each of the 36 selected base pairs has 6% chance to mutate to a non-wild-type base). Library sizes are indicated.

6.4.3 *Use of aCTD*t Library for Improving Ethanol Productivity in Ethanologenic KO11 E. coli strain*

6.4.3.1 Cellulosic Ethanol as a Biofuel

Ethanol, today's predominant biofuel, is currently manufactured using feedstocks such as cane-derived sucrose and corn-derived starch. In Brazil, sugar cane juice and sugar cane molasses are used as sources of sucrose. Using *Saccharomyces* yeast strain as the host organism, industrial ethanol yields on sucrose are up to 93% of the stoichiometric maximum. Both continuous and batch production is used, with residence times in the fermentors being in the order of 6-10 hrs (da Silva et al., 2005). Production is more than 15 billion gallons per year (bgpy).

In the US, the predominant feedstock is maize. Studies have reported industrial yields in the range of 2.65 to 2.71 gallons per bushel of maize, which amounts to approximately 93.6 to 95.8% of the stoichiometric maximum (McAloon *et al.*, 2000; McLeod *et al.*, 2002). Ethanol titer post-fermentation is around 9% by weight (McAloon et al., 2000).

The process for production of ethanol from these sources is highly efficient (from the standpoint of conversion), although several arguments impede a wider or longer-term acceptance of this route. In particular, the utilization of food sources for fuel has been generally unpopular, but other concerns such as net energy ratio (especially for corn-derived ethanol) have emerged.

Cellulosic biomass, on the other hand, enjoys a much more massive resource base than what is available from maize or sugarcane (Perlach et al., 2005). It is expected to play a dominant role in biofuels production in the near future. However, cellulosic biomass is more difficult to convert into fermentable sugars than is corn or sugar cane, because (i) five-carbon sugars, mainly xylose, account for 10 – 25% of the total carbohydrates, which cannot be utilized by the native yeast; (ii) the presence of lignin, a highly recalcitrant network polymer of aromatic alcohols that accounts for 17 – 25% of common cellulosic biomasses (van Maris et al., 2006), makes cellulose much more resistant to hydrolysis than starches and simple oligosaccharides; and (iii) when hydrolyzed, the resulting broth contains a variety of toxic compounds, which adds to the harmful effects of ethanol (see, also, Section 4.3.2).

The first obstacle can be overcome through the selection and/or engineering of fermentative microbes capable of anaerobic fermentation of xylose and other five-carbon sugars to ethanol. The third can be tackled with the methods described in the present thesis.

6.4.3.2 A Traditional Metabolic Engineering Approach for Developing an Ethanologenic *E. coli* Strain

The laboratory of Prof. L.O. Ingram developed an *E.coli* strain that is able to ferment sugars to ethanol much more efficiently than the wild-type. This strain was constructed by plasmid-borne overexpression of pyruvate decarboxylase (coded by the gene *pdc*) and alcohol dehydrogenase (coded by *adhB*) from *Zymomonas mobilis* (Ingram & Conway, 1988; Ingram *et al.*, 1987). The first enzyme catalyzes the conversion of pyruvate to acetaldehyde and carbon dioxide, and the second enzyme reduces the acetaldehyde to ethanol in a NADH-dependent fashion. The recombinant strain benefited from the high glycolytic flux of *E. coli* and the ease of its manipulation, while at the same time exploiting the ability of this species to utilize both 5- and 6-carbon sugars.

Although this initial approach was successful in producing an ethanologenic *E. coli* strain, two additional modifications were needed to construct an industrial platform for ethanol production. First, because the production of organic acids from pyruvate was still significant, additional modifications were needed to increase the yield from sugar.

Second, because the two key genes for ethanol production, *pdc* and *adhB*, were expressed from a plasmid, the phenotype was unduly unstable.

A marked increase in yield was achieved by elimination of carbon sinks that competed with ethanol. For example, a deletion in pyruvate formate lyase (coded by *pfl*), which drains the pyruvate pool to produce formate and acetyl-CoA, decreased the production of acetate by 70% (Ohta *et al.*, 1991). The same study reported a deletion in fumarate reductase (coded by *frd*), which almost eliminated succinic acid production.

The challenge of plasmid instability was tackled through chromosomal integration of the *pdc-adhB* cassette; the operon also contained a chloramphenicol resistance gene (i.e. to act as a selective marker for integration), resulting in the widely-studied strain KO11 (Ohta *et al.*, 1991). The decrease in copy number due to integration was compensated by an overexpression of the operon achieved by selecting in high chloramphenicol concentrations (up to 600 μg/mL). A spontaneous mutant that overproduced chloramphenicol acetyltransferase (coded by *cat*) also showed overexpression of *pdc* and *adhB*, with a concomitant increase in ethanol production. Our laboratory received a derivative of KO11 of which the *cat* gene had been excised. Our control was this derivative strain transformed with a plasmid-borne wild-type copy of *rpoA*.

6.4.3.3 The Challenge of Fermenting Biomass Hydrolysates

The challenge of cellulosic biomass utilization has been undertaken in different ways. For example, so-called consolidated bioprocessing (CBP) uses highly cellulolytic and ethanologenic organisms like the thermophilic *Clostridium thermocellum* either exclusively or in co-culture with other thermophilic, higher-producing sugar fermenters (Lynd *et al.*, 2002; Ng *et al.*, 1981). These organisms permit cellulase production, cellulose hydrolysis, and fermentation to occur anaerobically in the same process vessel. The difficulty in genetically modifying *Clostridia* and large energetic demand for anaerobic cellulase production (Lynd et al., 2002), have invited efforts with organisms that ferment sugars in solution. The development of KO11 strain described in the previous section is one such approach.

When the production strain cannot degrade lignocellulosic material directly, the substrate for fermentation may contain toxic compounds. This has been widely discussed in the biofuels literature for feedstocks derived from lignocellulosic hydrolysates, as pre-treatment by acid hydrolysis produces a mixture of oligosaccharides, organic acids, phenolic derivatives, and furans (Sakai et al., 2007), all but the first of which are inhibitors of growth for many microorganisms. Let us discuss a few from the variety of efforts pursued for coping with the challenge of environmental challenge in order to appreciate the significance of our work.

Lignocellulosic derivatives that have received most attention are furfural, hydroxymethyl furfural (HMF), acetic acid, and phenolic compounds. The amount and identity of the inhibitors after detoxification of hydrolysates depends on the method used (overliming, laccase treatment, charcoal, etc.) (Klinke et al., 2004). Although toxicity and detoxification issues have been mostly explored for ethanol, some studies for other fermentations such as butanol exist (Ezeji *et al.*, 2007).

Because different compounds exert toxicity through different mechanisms and their effects appear to be coupled (Ezeji *et al.*, 2007; Klinke *et al.*, 2004; Taherzadeh *et al.*, 1997b), solving the problem of environmental tolerance to the substrate cocktail by rational approaches and simple process modifications seems unlikely. However, partial successes from this front have been reported. For example, overexpression of ADH6 in *S. cerevisiae*, an 5-HMF reducing enzyme, has enhanced conversion of the inhibitor which could be probably used in detoxification, but no increase in ethanol productivity was reported (Petersson et al., 2006). A similar effort with ZWF1, encoding a glucose-6-phosphate dehydrogenase, resulted in higher furfural tolerance (Gorsich et al., 2006). Simultaneous overexpression of the genes would probably be synergistic, as the reduction of the inhibitor by ADH6 is NADPH-dependent, and ZWF1 is hypothesized to help this step by committing its substrate to the pentose-phosphate pathway, which produces the reduced cofactor. Overexpression of the enzyme phenylacrylic acid decarboxylase (coded by PAD1) resulted in *Saccharomyces* strains improved in ethanol productivity in the presence of ferulic and cinnamic acids (Larsson et al., 2001).

Manipulation of the fermentation pH has been used for alleviating tolerance to acetic acid, as toxicity is mainly effected by the protonated species (Lawford & Rousseau, 1998; Taherzadeh *et al.*, 1997a). However, pH control is undesirable because of the additional cost associated with it, and because low pH reduces the risk of contamination as discussed above. Transferring a gene of acid-resistant *Oenococcus oeni* that responds to different stresses resulted in an *E. coli* strain with improved low pH tolerance (Morel et al., 2001). These or similarly constructed hosts may be better suited for fermentations at low pH.

Improvements from random approaches have also been reported. Genome shuffling of ethanologenic *Candida krusei* has delivered acetic acid-resistant mutants that perform better than the parent in ethanol fermentations in the presence of the inhibitor (Wei et al., 2007). The usefulness of classical strain improvement (i.e. mutagenesis and selection) methods for improving tolerance of yeasts to lignocellulosic hydrolysate components has also been reported (Liu et al., 2005; Sonderegger et al., 2004). Similarly, studies describe that *Pichia stipitis* long-term adapted to increasing concentrations of hardwood hydrolysate partially neutralized or alkalinized with lime had higher ethanol productivity and titer (Nigam, 2001a; Nigam, 2001b).

We discussed in Section 4.3.2 our attempt of using sigma H and sigma E libraries for improving ethanol tolerance in *E. coli*, although at that point we did not use an ethanologenic strain, nor we challenged our cells in biomass hydrolysate medium. In an effort to construct a better platform strain for ethanol production from biomass, we focused on the development of a KO11-derived strain that is also able to withstand the compounds present in the fermentation of cellulosic hydrolysates. The background offered in this Section serves also as a reference for the experiments described later in Section 6.6.

6.4.3.4 Selection and Isolation of Improved Mutants

We began by transforming the αCTD*t library with a mutagenesis rate of 6% expressed from the *spc* promoter (see Section 6.4.2) into the ethanologenic KO11 derivative (which we will call the 'parental strain'). Two selection routes were chosen, either based in liquid or solid media (**Figure 6.4-4**). The mutants we describe in this section were isolated by subjecting the cells to 40 g/L of ethanol, and plating in non-stressful solid rich media after 48 hr.

Table 6. Survivability of different mutants in 40 g/L ethanol at 48 hr

Colony	CFU/mL	Colony	CFU/mL
1	4.38E+06	11	2.10E+06
2	7.53E+06	12	1.28E+06
3	1.90E+06	13	2.52E+06
4	5.35E+06	14	6.11E+06
5	6.42E+06	15	5.28E+06
6	3.31E+06	16	2.69E+06
7	8.65E+06	17	2.05E+06
8	4.52E+06	18	>8.65E+06
9	8.23E+06	19	3.39E+06
10	4.71E+06	Control	3.10E+04

Table 6 shows the results from 19 colonies tested individually after selection, in the same challenging conditions. These were re-transformed into the parental strain and re-tested for ethanol tolerance. Three improved strains were obtained, shown in **Figure 6.4-5**.

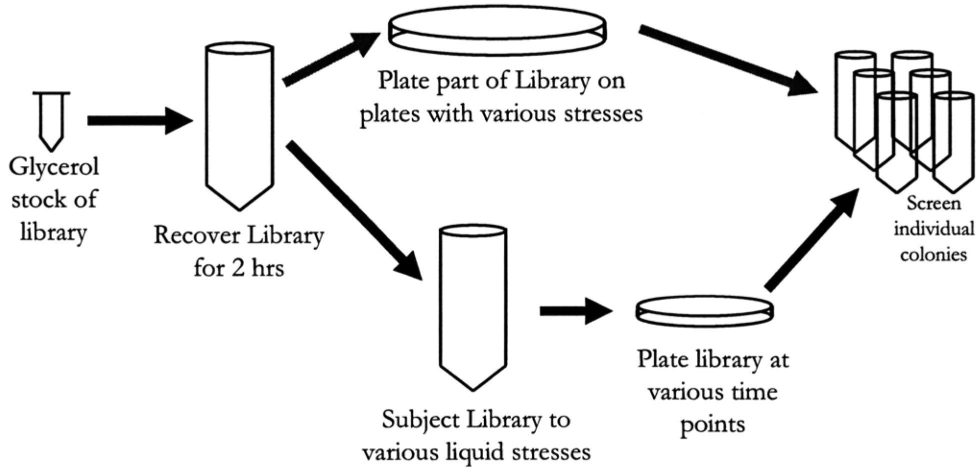
Glycerol stock of library
Recover Library for 2 hrs
Plate part of Library on plates with various stresses
Subject Library to various liquid stresses
Plate library at various time points
Screen individual colonies

Figure 6.4-4. Routes for isolating an ethanol- and hydrolysate-tolerant mutant

The route used to select for the mutants described in this Section is in black; an alternative path used in other experiments is shown in grey.

The fermentations shown in **Figure 6.4-5** were obtained by inoculating overlimed bagasse hydrolysate, with a total sugar concentration of 10 wt-% (this concentration was adjusted with a concentrated xylose solution and ensured by HPLC). The media was enriched by 5% corn-steep liquor, which provides a mix of nitrogen sources. At 36 hr, a fed-batch strategy was adopted in which additional xylose and corn-steep liquor were introduced to maintain ethanol production.

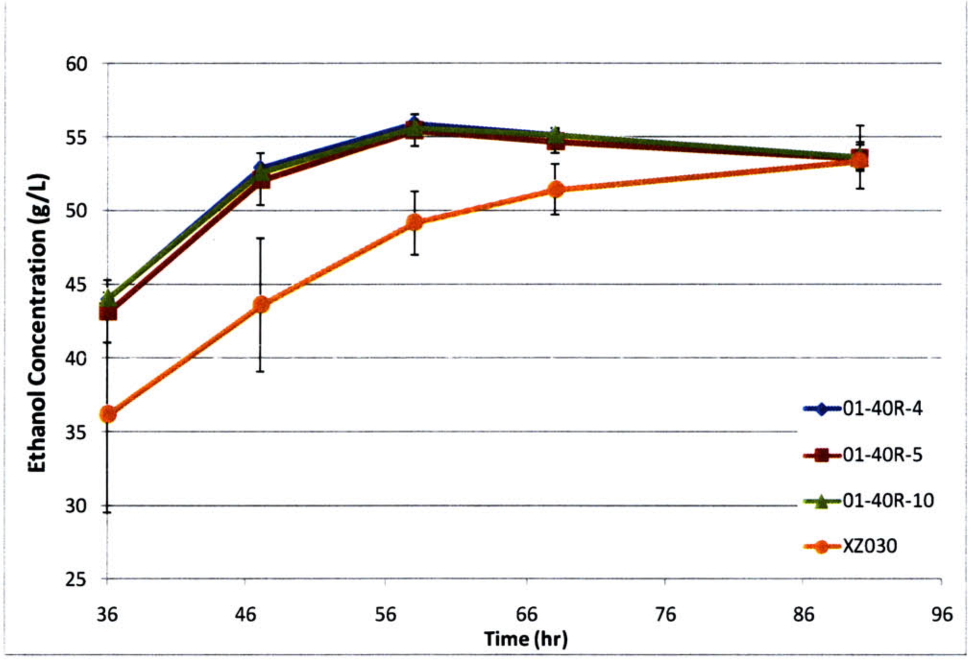

Figure 6.4-5. Fed-batch hydrolysate fermentation of three improved mutants and control

The three mutants are symbolized by the diamond, square, and triangle, while the control, denoted XZ030, is symbolized by a circle.

As shown by **Figure 6.4-5**, the ethanol productivity of the mutants is significantly improved (in the statistical sense) by ~10-15%. Concurrent measurement of sugar consumption was used to determine the yield for ethanol production, and we observed that both mutants and control show a similar value of 0.45 g/g, or about 88% of the theoretical maximum.

6.5 Divergence in a Single-Locus Phenotype

In Chapter 5, we described the use of complex phenotypes as proxies to quantify the impact of a particular library design on the global intracellular environment. We argued that, since we are interested in the *a priori* probability that a new phenotype will be found, we can measure this diversity using a measurable phenotype that is impacted by many nodes in the physiological network.

The bacterial RNAP, being the only enzyme complex in charge of transcription, must itself integrate all the signals that result in disparate expression of the different promoters in the cell (see Section 4.1). Not all promoters are regulated in the same way: some are simple and depend only in their DNA sequence, while others rely on a combination of activation, lack of repression, local structure of the chromosome, etc.

For the case of a simple promoter, mutations in the promoter region are known to cause wide changes in expression (Alper *et al.*, 2005; Hammer *et al.*, 2006; Jensen & Hammer, 1998a; Jensen & Hammer, 1998b). We hypothesized that the same could be observed if we mutated the discriminatory determinants of the RNAP, instead of those at the promoter. Diversity in the affinity of RNAP for a promoter would show as divergence in a simple phenotype, for example, the expression of a single gene.

In order to test this, we constructed a strain that expressed green fluorescent protein (GFP) in a very low-copy number plasmid (Jones & Keasling, 1998; Jones *et al.*, 2000), and from a constitutive promoter (we induced the P_{trc} promoter present in pKLJ03 with IPTG). We then transformed three of our *rpoA* libraries or a wild-type copy into this strain, and measured their fluorescence, which was then used to calculate the divergence as determined by GFP expression. All the libraries express the *rpoA* from the weaker P_{lac} promoter (i.e. this αCTD*t library is the same shown in **Figure 6.2-4**).

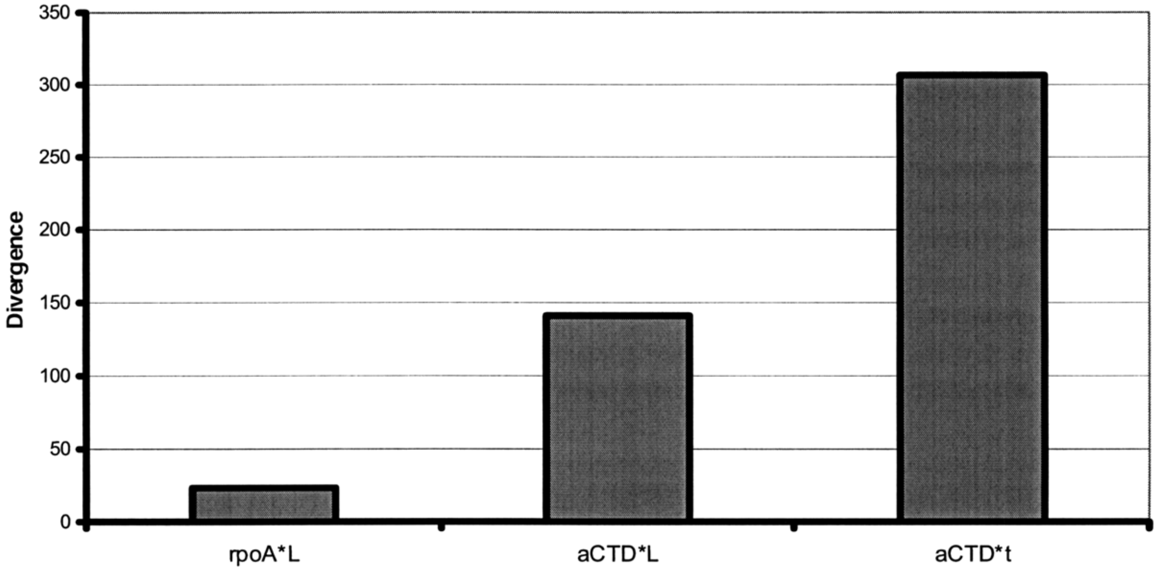

Figure 6.5-1. Divergence in a simple phenotype

The divergence was calculated using fluorescence of GFP as a single-locus phenotype for diversity quantification.

As shown in **Figure 6.5-1**, the observed trends are in accord with those established using a complex phenotype (pH_i). This fact suggests that some of the diversity that we observed using the complex phenotype had to be due to an alteration of the affinity of RNAP for promoters, at least at the DNA level (i.e. without considering protein effectors). Moreover, it suggests that the promoter discrimination function of the alpha subunit is affected following the same trend as for complex phenotypes, with the αCTD*t having the highest fraction of variants with altered interactions. From this fact, it follows, though not strictly, that the DNA determinants of a sizable number of promoters dictate a significant fraction of regulation (as opposed to protein effectors or other signals), which makes intuitive sense.

6.6 Optimization of sigma D Libraries

During our optimization efforts of the alpha subunit libraries we learned that it was possible to reduce the search space to regions that are important for transcriptional regulation, and we concluded that such a reduction led to a marked increase in phenotypic diversity. As a second test case for proving the usefulness of the divergence metric to optimize libraries, we chose the principal sigma factor, sigma D, which had already proven successful for improving traits of interest in *L. plantarum* and *E. coli* (Chapter 4).

A key advantage of doing this was that the structure and function of the *E. coli* sigma D protein has been studied in great detail (Campbell *et al.*, 2002; Dombroski *et al.*, 1992; Dove *et al.*, 2003; Gardella *et al.*, 1989; Nickels *et al.*, 2002; Ross *et al.*, 2003; Siegele *et al.*, 1988; Siegele *et al.*, 1989; Waldburger *et al.*, 1990). These reports, and especially the one by Campbell (2002), allowed us to apply some of the lessons that led to the development of the αCTD*t library on optimizing libraries of sigma D.

6.6.1 *Targeted Mutations in Promoter-Contacting Residues*

Based on the fact that targeting a few amino acid changes in the surface of the αCTD responsible for some of alpha's regulatory functions was the most successful approach for increasing divergence, we aimed at emulating this design. Recall from Section 6.4.1 that the design in which the chosen codons were substituted by degenerate triplets suffered from lethal effects, which suggested that a similar effect was possible for the case of sigma D. Therefore, spiked oligonucleotides were preferred for imparting localized mutagenesis.

We chose surface and DNA-contacting residues based on two criteria: structural information, obtained mostly on the results from Campbell (Campbell *et al.*, 2002), and sequence information, based on an alignment of the six sigma factors in *E. coli* that belong to the sigma 70 family (**Figure 6.6-1**). The first criterion follows directly from the evidence put forth by the *rpoA* divergence trends for complex and simple phenotypes. The second was provided by the hypothesis that, because all promoters in the cell must be recognized by one of the sigma factors, mutating amino acids with conserved physicochemical properties across more than one factor (but not in all of them) may change the pattern of recognition of sigma D to that of a different sigma.

With these facts in mind, we designed two libraries with surface residues hypothesized or known to be important for transcription targeted for mutagenesis: one library was constructed with amino acids that contact the -10 promoter region, and a different library with those that contact the -35 promoter region. The mutation frequency of each library was adjusted so that the total number of amino acids to be mutated was, on average, the same as that for the αCTD*t library with 6% base-exchange rate. A detailed design is shown by the highlighted residues in **Figure 6.6-1**.

-10-binding region

```
RPOD ECOLI      FLDLIQEGNIGLMKAVDKFEYRRGYKFSTYATWWIRQAITRSIADQARTI 450
RPOS_ECOLI      LLDLIEEGNLGLIRAVEKFDPERGFRFSTYATWWIRQTIERAIMNQTRTI 165
RP32_ECOLI      QADLIQEGNIGLMKAVRRFNPEVGVRLVSFAVHWIKAEIHEYVLRNWRIV 124
FLIA_ECOLI      LDDLLQAGGIGLLNAVERYDALQGTAFTTYAVQRIRGAMLDELRSRD--- 87
RPOE_ECOLI      LTDQV------LVERVQKGDQKAFNLLVVRYQHKVASLVSRYVPSGD--- 45
FECI_ECOLI      -----------MSDRATTTASLTFESLYGTHHGWLKSWLTRKLQSAF--- 36
```

-35-binding region

```
RPOD_ECOLI      THDVLAGLTAREAKVLRMRFGIDMNTDYTLEEVGKQFDVTRERIRQIEAK 593
RPOS_ECOLI      IVKWLFELNAKQREVLARRFGLLGYEAATLEDVGREIGLTRERVRQIQVE 308
RP32_ECOLI      LTDAMQGLDERSQDIIRARW-LDEDNKSTLQELADRYGVSAERVRQLEKN 273
FLIA_ECOLI      VMEAIETLPEREKLVLTLYY----QEELNLKEIGAVLEVGESRVSQLHSQ 227
RPOE_ECOLI      VFRTIESLPEDLRMAITLRELDGLSYEEIAAIMDCPVGTVRSRIFRAREA 180
FECI_ECOLI      LDSMLDGLNGKTREAFLLSQLDGLTYSEIAHKLGVSISSVKKYVAKAVEH 163
```

Figure 6.6-1. Sequence alignment of different sigma factors in E. coli

All the featured factors belong to the sigma 70 family of sigma factors. The left-hand side before the sequence shows the name of the gene corresponding to the following (from top to bottom): sigma D, sigma S, sigma H, sigma F, sigma E, and sigma fecI (see Section 4.3). The numbers at the right-hand side correspond to the amino acid position given by the last letter of the sequence. The amino acids chosen for mutagenesis are highlighted.

Two libraries were constructed because, in contrast to the alpha subunit of RNAP, sigma D has two clusters of amino acids that contact the promoter, instead of one. Experimentally, this means that the two regions are hard to target simultaneously, because the spiked triplets are introduced with the aid of mutated oligonucleotides. Having two libraries instead of one also opened the possibility of comparing the two designs with the divergence metric and learn which region of sigma D can impart more diversity. In other words, the structural (and functional) separation of the two sigma regions provided a natural division into two search spaces. This would later be exploited for reducing the space in the course of optimization.

When we performed the experiments for constructing these two libraries, we observed very low transformation efficiencies, even after extensive troubleshooting. **Figure 6.6-2** shows the transformants per mL of different such experiments; although we focused on perfecting the -10-binding region library, we initially observed equally poor efficiencies for the case of the -35-binding region.

We hypothesized that, if mutations in the chosen residues were lethal, we could calculate the expected number of transformants by calculating the probability that a variant in DNA library had no mutations. Assuming a Poisson distribution for the number of mutations (as before), and assuming a transformation efficiency that would result in ~10^4 CFU/mL, we estimated that about 400 colonies would survive if only unmutated plasmids resulted in non-lethal clones. This value is in the observed order of magnitude, suggesting that our hypothesis could be valid. In order to confirm this, we sequenced a few clones that were produced in one of the transformations. Indeed, all clones were genotypically identical, which would result in a population with low or null divergence.

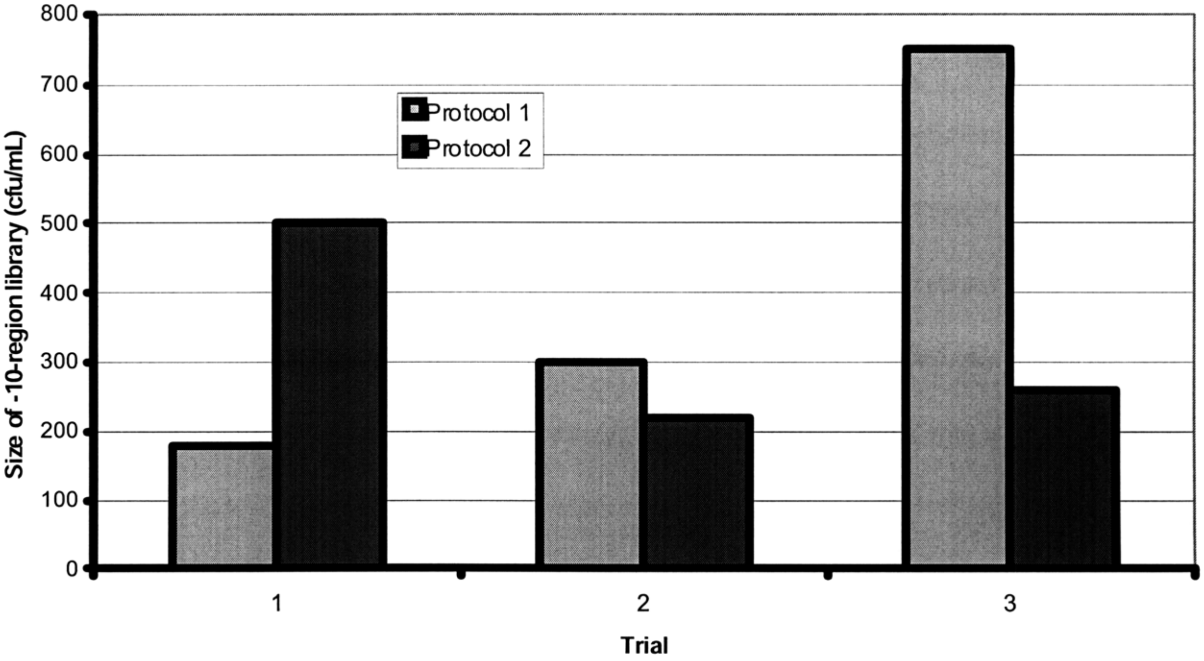

Figure 6.6-2. Initial size of -10-binding region libraries

The graph shows the number of transformants obtained in three different trials (in which different enzyme stock solutions were used) undertook to construct the library with targeted mutations to the region of sigma that contacts the -10 promoter hexamer. The two bar colors refer to experiments in which gel electrophoresis was either performed (protocol 1) or not performed (protocol 2) prior to ligation.

Although the lethality of mutants in the chosen amino acids continues to be our working hypothesis, other possibilities were later imagined. For example, poor primer quality could result in a large fraction of variants containing deletions or insertions instead of spiked base changes; in that case, lethal effects would be caused by gross changes in sigma D structure, not the more subtle effects of nonsynonymous substitutions. It is also possible that the toxic effects of the chosen substitutions were restricted to the -10-binding region library, and that we could have obtained larger library sizes for the -35-binding region library if we had extensively troubleshot it as we did for the former. In any case, the experiments described here suggested the need for libraries with a modified design.

6.6.2 *Targeted Mutagenesis of Promoter-Binding Regions*

Targeting mutations to particular DNA-binding amino acids in sigma D proved ineffective, but we still believed that reducing the sequence space as a means of optimizing the *rpoD* libraries was possible. One way of attaining this goal would have been choosing a different set of amino acids to target for mutagenesis. Because it was not obvious which residues to target for a new design, we instead chose to select the regions that contained the surface amino acids that were previously selected (spanning a couple of hundred base pairs each).

The libraries of the region of sigma D that binds the -10 promoter hexamer, which we denoted -10*L and -10*H according to their mutation frequency, were constructed by epPCR of a base pair sequence between amino acids 422 and 456; this region includes all of region 2.4 and parts of regions 2.3 and 3.0 (see supplementary material of (Campbell *et al.*, 2002)). The libraries of the region of sigma D that binds the -35 promoter hexamer, denoted -35*L and -35*H, were similarly constructed by targeting the region from amino acid 546 and until the stop codon; this region contains most of region 4.1 and all of 4.2 (same source as above). **Figure 6.6-3** illustrates the design of these four libraries schematically.

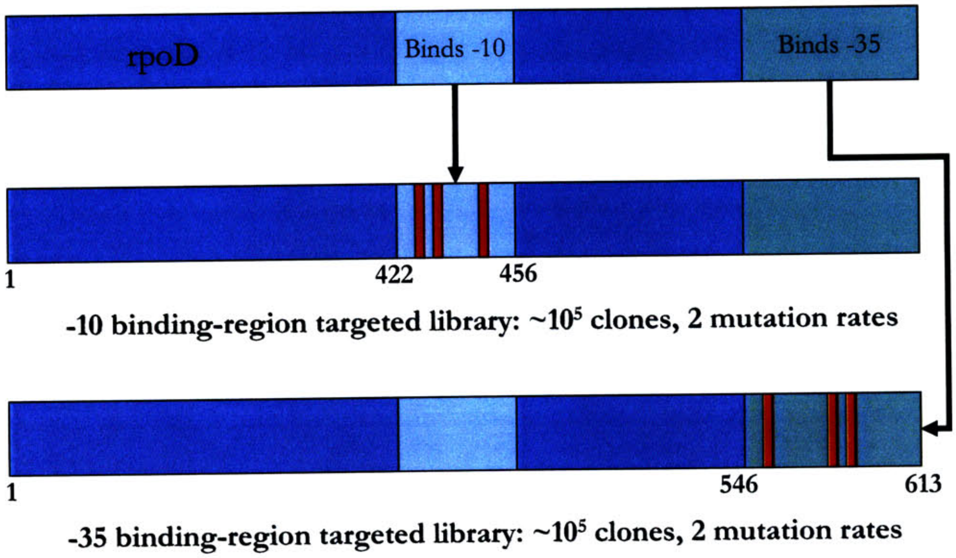
rpoD
Binds -10
Binds -35
1
422
456
-10 binding-region targeted library: ~10⁵ clones, 2 mutation rates
1
546
613
-35 binding-region targeted library: ~10⁵ clones, 2 mutation rates

Figure 6.6-3. Design of the libraries targeted to the 2.4 and 4.2 regions of sigma D

Generalized mutations are indicated in red for schematic purposes only; they are not intended to show the actual location or frequency of mutations in our libraries.

After constructing the libraries, we used the protocol for quantifying the divergence using pH_i, except that only growing conditions were used as phenotypes for measuring diversity (recall that we had used both growing and non-growing conditions to produce the data of **Figure 6.2-4** in Section 6.2.3, but we later simplified the protocol further). We also quantified the divergence of a library of *rpoD* with high mutation frequency (rpoD*H) in order to compare the new libraries to the best sigma factor library available then. In addition, we included the αCTD*t in our analysis to judge the new libraries in light of the library that showed the highest level of divergence so far. The results are shown in **Figure 6.6-4**.

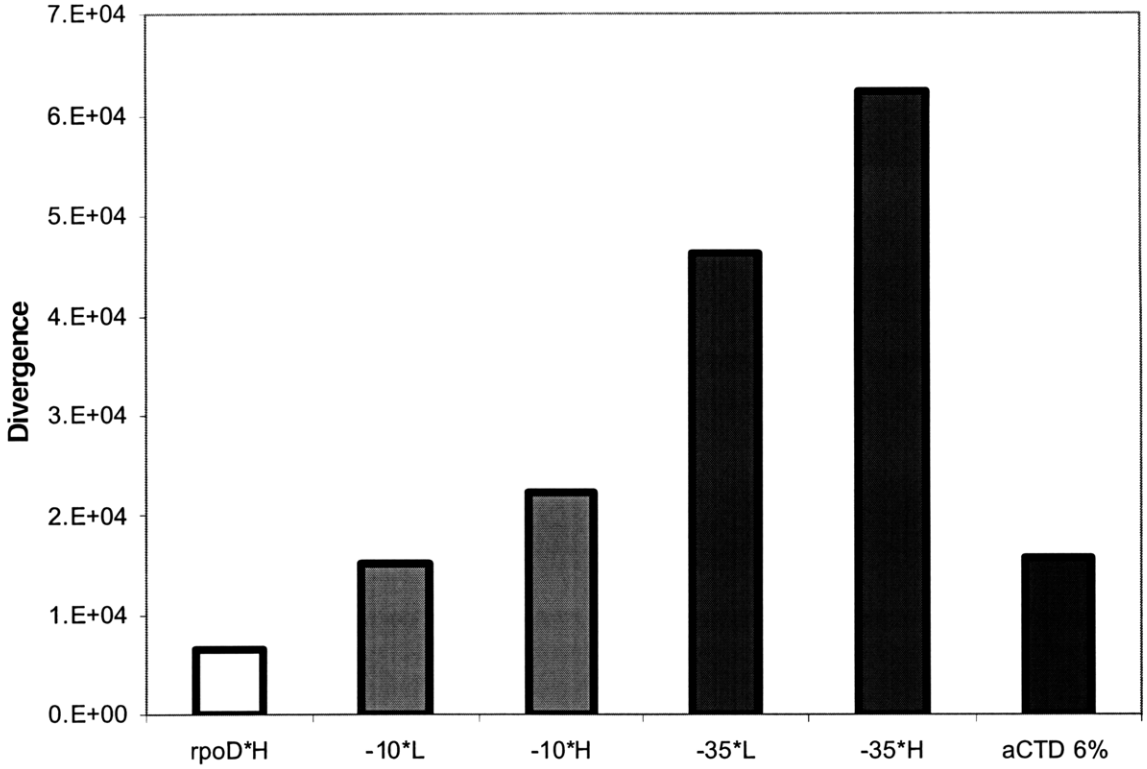

Figure 6.6-4. Divergence of the libraries targeted to the 2.4 and 4.2 regions of sigma D

The divergence was calculated using pH_i in growing conditions, and contrasted to that of a rpoD library with high mutation frequency throughout the coding region (rpoD*H) and to that of the αCTD*t library that had showed highest divergence so far.

From **Figure 6.6-4**, we can observe that (i) localizing mutagenesis to regions in charge of DNA-binding is effective for imparting phenotypic diversity; (ii) such diversity is significantly greater than that of non-targeted mutagenesis to rpoD; (iii) such diversity compares favorably with that of previously-optimized *rpoA* libraries; (iv) mutagenesis of the region that binds the -35 promoter hexamer can introduce higher phenotypic diversity than that of the region that binds the -10 promoter hexamer; and (v) higher sequence diversity translates into higher phenotypic diversity for the case of the new targeted libraries.

Observations (i) and (ii) imply that an optimization effort based on reducing the search space from ~2 kbp (as for rpoD*H) to 200-300 bp is an effective means of improving the library design. They also reinforce our hypothesis that the divergence metric can help in finding regions rich in phenotype-altering potential located in the midst of uninteresting regions.

Observation (iv) implies that it is easier to modify the specificity of sigma D for the promoter by changing its interactions at or near the -35 promoter hexamer than at or near the -10 promoter hexamer. The most likely explanation for this is that region 4 of sigma has regulatory functions in addition to contacting DNA, as it is known to interact with protein effectors and with the C-terminal domain of alpha (Dove *et al.*, 2003; Nickels *et al.*, 2002; Ross *et al.*, 2003). Recall from Section 4.2.2.3 that one of the *L. plantarum* mutants (S6) had a mutation in this region, which could have been taken as an early lead to the potential of the -35* libraries for phenotypic improvement.

On the other hand, the regions that were chosen for mutagenesis in the -10* libraries are also known to be responsible for DNA melting during transcription initiation (see Section 4.1). Mechanistically, it is easy to imagine how mutations in these regions could not only change the affinity of sigma for the -10 promoter hexamer, but at the same time interfere with transcription initiation. If that is the case, it is possible that some variants of the -10* libraries would not be able to function in promoter melting, reducing the effective size of the library and decreasing the divergence. This possibility adds to the explanation for observation (iv) given above.

Finally, observation (v) is in tune with the majority of the evidence obtained from the divergence metric: that increased sequence diversity leads to increased phenotypic diversity. Notable exceptions are the trend observed for the αCTD*L vs. the αCTD*H libraries (see Section 6.2.3), and the case of the αCTD*t libraries with fully-degenerate codons vs. that with 6% spiked non-wild-type bases (see Section 6.4.1). Before concluding our description on optimization of transcriptional engineering libraries, let us turn to a final instance that illustrates how the divergence metric can be employed for guiding the use of libraries to quicken selection experiments.

6.6.3 *Use of sigma D Optimized Libraries for Improving Survivability in High-Ethanol Overlimed Hydrolysate*

Even though the main goal of the thesis was to develop an understanding and the necessary methods for optimizing random search-based approaches for phenotypic improvement, we screened our libraries for traits of interest to illustrate the practicality of our approaches. So far, we have explained the use of transcriptional engineering libraries to improve several phenotypes in *L. plantarum* and *E. coli.* Since we had also explored our optimization method on libraries of sigma D in *E. coli,* we felt compelled to prove their usefulness with regard to a phenotype of interest.

For the same reasons given in Section 6.4.3, we sought mutants that were tolerant to conditions present in ethanol fermentations of biomass hydrolysate. We transformed our previously optimized libraries into the ethanologenic industrial strain (derived from KO11) and selected for survivors in 50 g/L ethanol and overlimed bagasse hydrolysate. Since the -10* and -35* libraries may act through different mechanisms, we not only performed the selection experiments in the -35*H library – the one with highest divergence –, but we tested both -10*H and -35*H libraries.

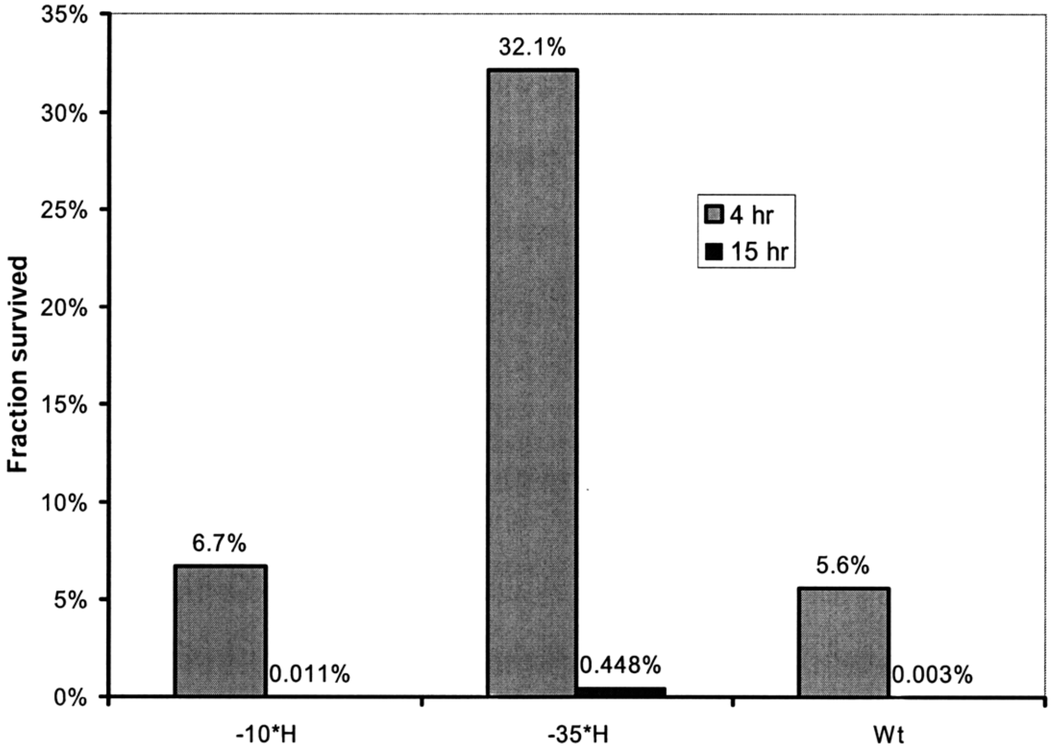

Figure 6.6-5. Library selection in ethanol and overlimed hydrolysate

The bars show the fraction of cells surviving at 4 and 15 hr of the two tested libraries and the control (Wt; host strain transformed with a plasmid-borne copy of *rpoD*). The data was calculated by counting colonies initially and at the different time points. Since the trends are hard to distinguish at 15 hr, the value represented by each bar is printed as well.

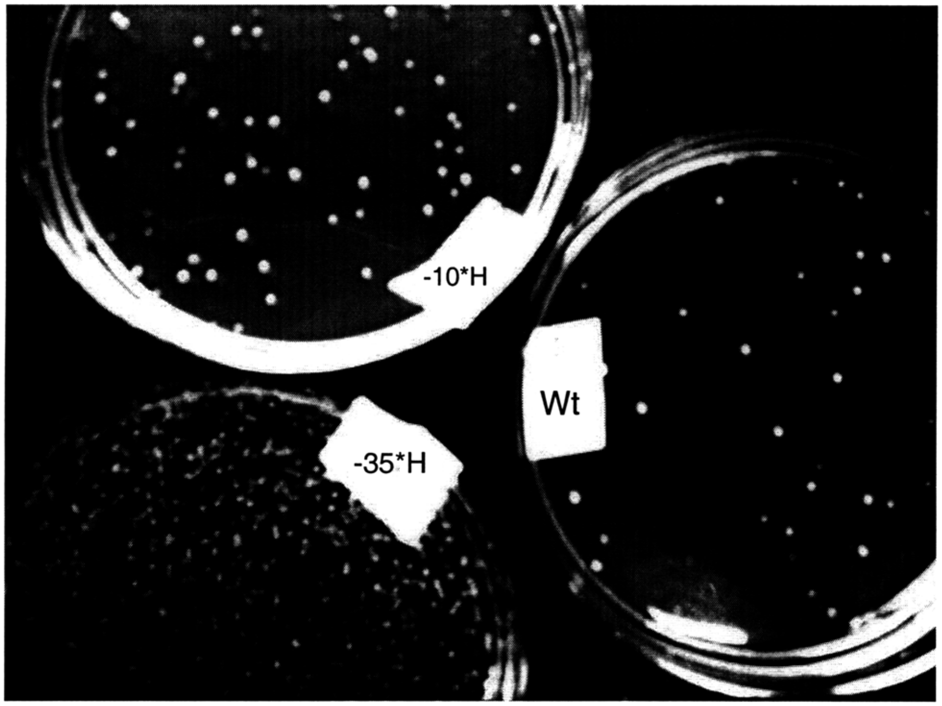

Figure 6.6-6. Qualitative illustration of library selection in ethanol and overlimed hydrolysate

The same volume of media was plated from each library (at 15 hr). The data shown in the previous figure was calculated using both this dilution and a dilution plated at time zero.

The results of the selection experiment are shown, quantitatively and qualitatively, by **Figure 6.6-5** and **Figure 6.6-6**, respectively. As shown, the trends shown by **Figure 6.6-5** are in good agreement with those extracted from the divergence metric, although this needed not be the case. For example, it is possible that the -35*H library has more diversity than the -10*H library, but that a smaller fraction exhibit the phenotype selected for by this protocol. The reason is that this experiment has a “digital” outcome: either a variant survives or it does not. A more relevant indication that the divergence of the -35*H library confers useful information is that the mutant found to be improved was obtained from this library.

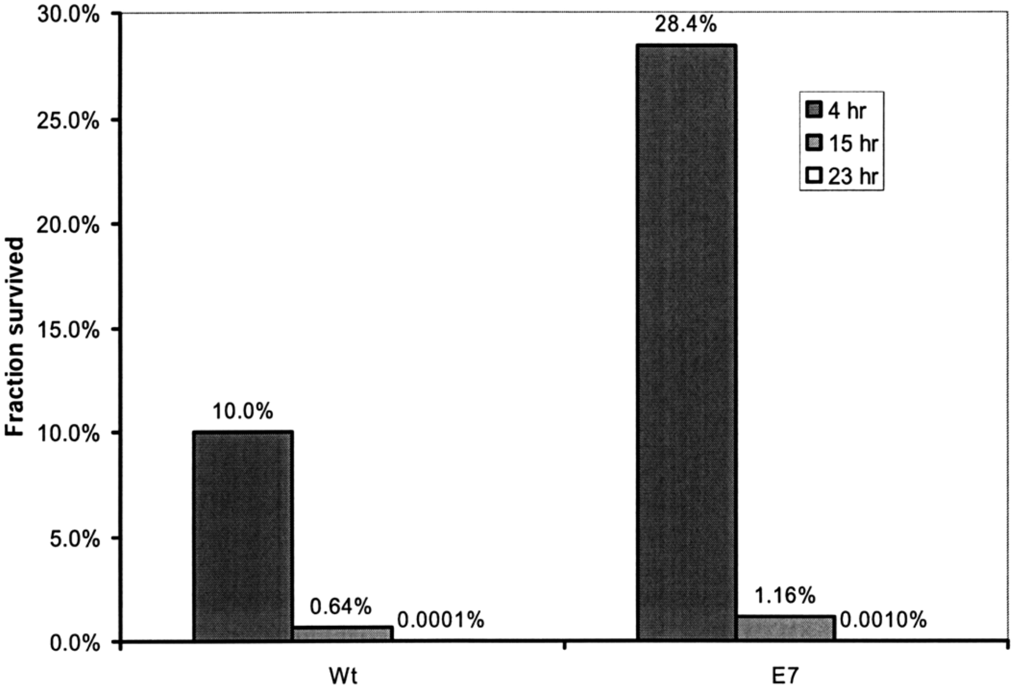

Figure 6.6-7. Improved mutant from our selection in ethanol and overlimed hydrolysate

The experiment was carried out similarly to that of selection. Mutant E7 was found in the -35*H library.

After testing several mutants from both libraries, the best individual (denoted E7) was chosen for further characterization; this mutant was indeed isolated from the population with highest divergence, the -35*H library. We isolated the mutant plasmid and re-transformed it into a fresh background to test the ability of the mutant *rpoD* to cause the improved phenotype. As shown by **Figure 6.6-7**, mutant E7 has a higher survival rate compared to the control at all time points tested. Upon sequencing the mutant, we observed two mutations A549V and R560H located in the region that was targeted for mutagenesis. It is interesting to note that, as shown by **Figure 6.6-8**, these mutations are adjacent to amino acids previously identified as critical for sigma D interactions (see Section 6.6.1). In addition, both cause amino acid substitutions with similar physicochemical properties to the native residues. This suggests that subtle changes located near important regions can be effective for altering phenotype.

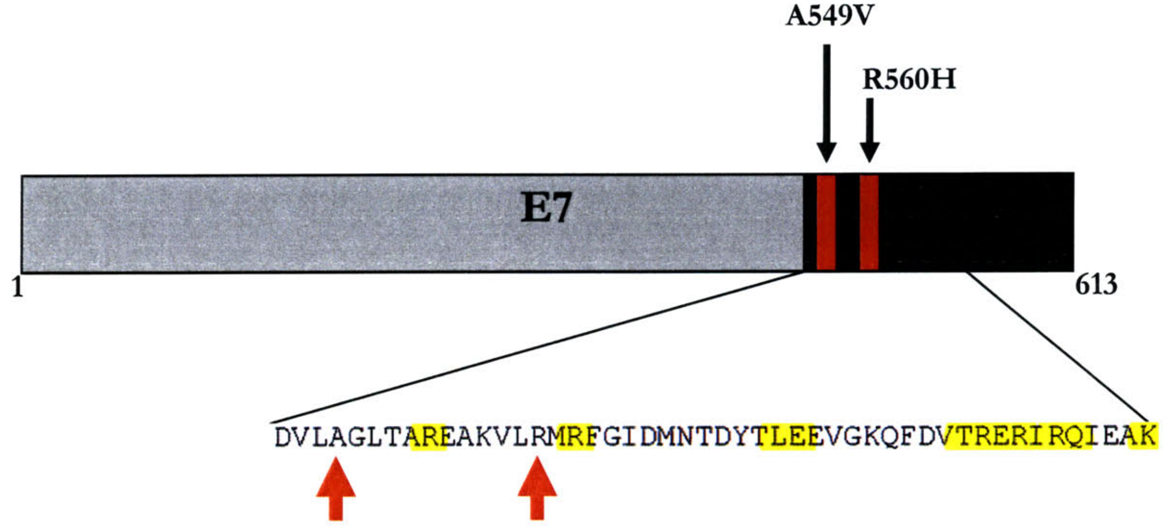

Figure 6.6-8. Sequence analysis of ethanol and hydrolysate tolerant mutant

The location of the mutations is indicated with red arrows. Amino acids that were chosen for localized mutagenesis, described in Section 6.6.1, are highlighted in yellow.

When we calculate the posterior probability of finding mutant E7 in all the previously-constructed rpoD libraries, it is naturally greatest in the -35* libraries. In fact, it is slightly higher in the -35*L library than in the -35*H library. For the same reasons given by the opening paragraph in Section 6.3, this does not contradict the information given by the divergence metric. The fact that we were able to find a valuable mutant most easily in the library with the most diversity, in contrast, does reflect the usefulness of the divergence metric for directing strain improvement efforts.

In order to complete our analysis, we tested mutant E7 for its ability to ferment overlimed hydrolysate, similarly to the mutants isolated from the αCTD*t library (Section 6.4.3). Our experiments showed no improvement in ethanol productivity or titer when comparing the mutant and control strains. This was in spite of the survivability advantage of E7 compared to the same control strain. There are many explanations for why survivability in the stress conditions does not translate to increases in productivity or titer, but, practically, it suffices to say that the selection conditions used did not enrich for the phenotype of interest. "You get what you screen for", the classic adage of protein engineering, has an equivalent in cellular engineering which must be kept in mind when designing selection experiments.

Chapter 7

7. The Divergence Metric as a Tool for Studying Interactions that Alter Phenotype

In Chapter 6, we described the use of the divergence metric for guiding and optimizing random search-based libraries. Although we stressed the practical aspect of the metric, we suggested, at times implicitly and at others explicitly yet succinctly, that the information provided by the divergence metric can be used to learn something about the protein targeted for mutagenesis. Concretely, we can study which domains, regions, or amino acids can affect the phenotype of the cell globally. Because we are working with transcriptional regulators, we can study which determinants of the chosen proteins can alter the transcriptome most effectively.

From the results of the previous Chapters, we have also seen that, when targeted appropriately, a reduction of the search space leads to an increase in diversity. But, what are the structures or functionalities that are most important for the diversity observed in the reduced space? What is the smallest sequence that can be regarded as the basis for diversity? How does one find these regions experimentally? In this Chapter, we aim at answering all these questions, by studying the alpha subunit in yet more detail. We expect that the methods presented here could be applied to targets other than alpha in the future.

7.1 Determinants of Diversity

While optimizing the transcriptional engineering libraries, we considered proteins, their domains, regions or sub-regions within those domains, and even groups of amino acids as targets for mutagenesis. Because libraries of these targets were shown to alter phenotype, they can be regarded as "units" of diversity, or, in the terminology of this thesis, as determinants of diversity. Conceptually, a determinant of diversity could be decomposed in sequences, structures, or functionalities; such decomposition may help explaining and exploiting the underlying mechanisms that give rise to the observed effect in phenotypic alteration.

For example, in Section 6.2 we showed that the C-terminal domain of alpha (αCTD) could be regarded as a determinant of diversity within the protein that concentrated useful functions for phenotypic alteration. We also deduced that the αCTD had more potential than that of the N-terminal domain (or NTD). The basis for selecting the CTD as a determinant for diversity was founded mostly in structural arguments, given that this is an independently-folding region of the protein, and it harbors several functions (see Section 4.4).

The potential of the determinants, quantified by the divergence metric, depends on the ability of the targets to affect key regulatory functions effectively when mutated. We have concluded that this efficacy is a balance between centrality – so that the chosen determinants can alter the phenotype globally and through different mechanisms – and sequence plasticity – so that the chosen determinants can be mutated without disrupting the cellular physiology to the point of lethality. We will refer to the combination of these qualities as "regulatory flexibility".

In this way, we can rephrase some of the conclusions of the preceding Chapters in light of the new terminology. For example, one can say that the group of amino acids chosen for position-specific mutagenesis in the surface of regions 2.4 of *rpoD* (see Section 6.6.1) has little regulatory flexibility, but the group located in the surface of the αCTD has significantly more. Note that this property, which is a quality of the group of residue positions at hand and not necessarily the positions individually, can be studied by comparing the divergence of different library designs. Framing the potential of different determinants as sources of regulatory flexibility can aid in obtaining information while optimizing a library or when comparing targets for mutagenesis

7.2 Single Residues as Determinants of Diversity

The smallest determinant of diversity considered so far has been groups of amino acids. The αCTD*t library, which has the highest phenotypic diversity of the *rpoA* libraries, was constructed by allowing mutations in a few amino acids located in the surface of the αCTD. The selection of those amino acids was to some extent arbitrary, and therefore, the αCTD*t library may be sub-optimal. Amino acids not considered in this design could be important, and some of those included could be dispensable.

These remarks raise the question of whether one can experimentally consider single amino acids as determinants of phenotypic diversity. If so, regions and sub-regions of proteins could be probed for their capacity to alter the intracellular environment with any resolution. This is because an amino acid is the smallest physicochemical determinant of a protein coded by DNA, and because it can be substituted independently to all other similarly-defined determinants.

Considering single amino acids as determinants for diversity would allow determining which structures or their associated functionalities are central to regulation and are worth exploring in more detail. Also, it would open the possibility to reduce the search space to the point where it can be comprehensively screened, which necessitates that only a small number of amino acids are targeted for mutagenesis.

Similar methods for directed evolution based on structural information or other rational arguments have been described (Reetz & Carballeira, 2007), but given that the proteins of interest regulate transcription through several mechanisms and that the libraries will be screened for improving many, possibly unrelated phenotypes, these cannot be readily used. Moreover, residues chosen because they have central roles in transcription are not indisputably desired, as mutations in these amino acids can be lethal or under strong selection pressures, which leads to small regulatory flexibility. The diversity quantification method can again serve to evaluate potential target residues.

7.3 Amino Acids in the alpha Subunit C-terminal Domain (αCTD) as a Test Case

7.3.1 *The Difference between Residues with Low and High Regulatory Flexibility*

As proof-of-concept, we constructed saturation mutagenesis libraries of two amino acids in the αCTD, based on their function and degree of conservation across species. The positive control (one expected to confer significant diversity) was R265, which has been described as a side-chain essential for αCTD-DNA interactions (Benoff *et al.*, 2002; Jeon *et al.*, 1995; Savery *et al.*, 2002). On the other hand, V242 was chosen as a negative control, since it is located in the flexible linker between NTD and CTD and is poorly conserved (Murakami *et al.*, 1996); therefore, a library at this position is expected to result in low diversity.

The quantification protocol was modified to deal with the fact that different amino acids are likely to play a more important role in some growth conditions than in others (Benoff *et al.*, 2002; Fritsch *et al.*, 2000; Lochowska *et al.*, 2004). Therefore, we measured the pH_i in different media that varied on the choice of carbon source, on the absence or presence of amino acids, and availability of complex nutrients. The initial analysis of these two libraries indicated that the R265 library had an order of magnitude larger divergence compared to that of V242 (see **Figure 7.3-1** in next Section).

7.3.2 *A More Complete Account of Function-Diversity Relationships*

To complete the study, several amino acids in the αCTD were chosen based on structural and functional information. In particular, residues in three determinants responsible for transcriptional regulation were selected: (i) D259, L262, K271, and E273 on or near the "261 determinant" (some have proposed the existence of an overlapping "273 determinant" but we do not for simplicity); (ii) R265, N268, and C269 on the "265 determinant"; and (iii) E286 and L290 on the "287 determinant" (Benoff *et al.*, 2002; Fritsch *et al.*, 2000; Kedzierska *et al.*, 2007; Luscombe & Thornton, 2002; Savery *et al.*, 2002). These residues were chosen because their function has been described in the literature and because they were all randomized in the αCTD*t library; therefore, their combined effect on diversity had already been studied.

Figure 7.3-1 shows the results of quantifying the diversity of the saturation mutagenesis libraries of the chosen amino acids, and **Figure 7.3-2** shows the data averaged by determinant. As shown by **Figure 7.3-1**, the V242 library has the lowest diversity, as expected, while the D259 library has the highest. The latter has been implicated in a direct intersubunit interaction with amino acid R603 located in region 4.2 of sigma D, in the vicinity of the contact of the sigma factor with the -35 promoter box (Ross *et al.*, 2003).

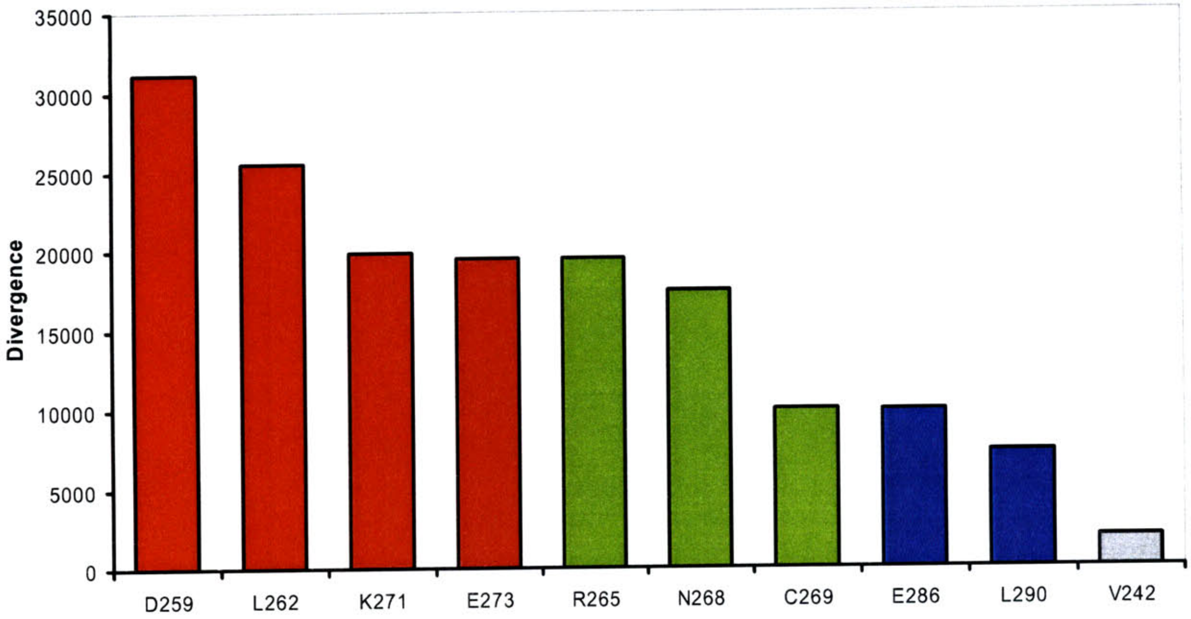

Figure 7.3-1. Divergence of individual amino acids in the αCTD

The x-axis denotes the identity and the location of the amino acids, while the color of the bar denotes the determinant to which the amino acid belongs (red, green, and blue for the 261, 265, and 287 determinants respectively). The negative control, V242, is shown in gray for comparison.

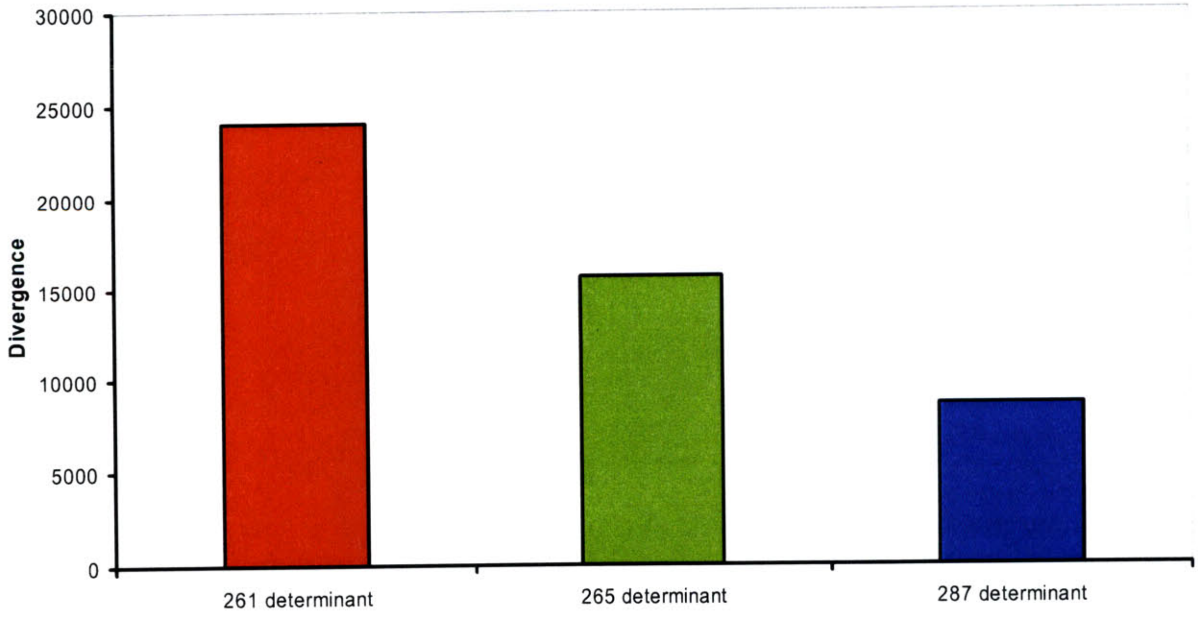

Figure 7.3-2. Divergence of the αCTD averaged by determinant

In addition, all residues tested in the "261 determinant" produce high diversity, suggesting that this face of the protein has the largest potential for transcriptional engineering through *rpoA*. In contrast, members of the "287 determinant", located in the opposite face of the CTD and mainly implied in protein-protein activation at class-I and class-II promoters (Savery *et al.*, 2002), showed the lowest diversity, suggesting that this activity of alpha has a lower potential to alter global phenotypes (i.e. a lower regulatory flexibility). Residues in the "265 determinant", which are responsible for alpha-DNA contacts at the UP promoter site, produced an intermediate level of diversity (**Figure 7.3-3**).

It is possible to conclude that saturation mutagenesis libraries can be used to study how diversity-conferring potential varies with position along the sequence of the regulator, giving leeway for probing regions of proteins with high precision. The structure/location-function correlations used to explain the diversity trends are a good indication that the methods presented give meaningful results, which could be exploited when testing regulators that have been studied with less detail. However, some knowledge regarding where influential residues are likely to be found is desirable, if not compulsory, in order to keep the number of libraries and evaluation experiments tractable. This may complicate the use of the presented protocol in proteins that have been barely explored.

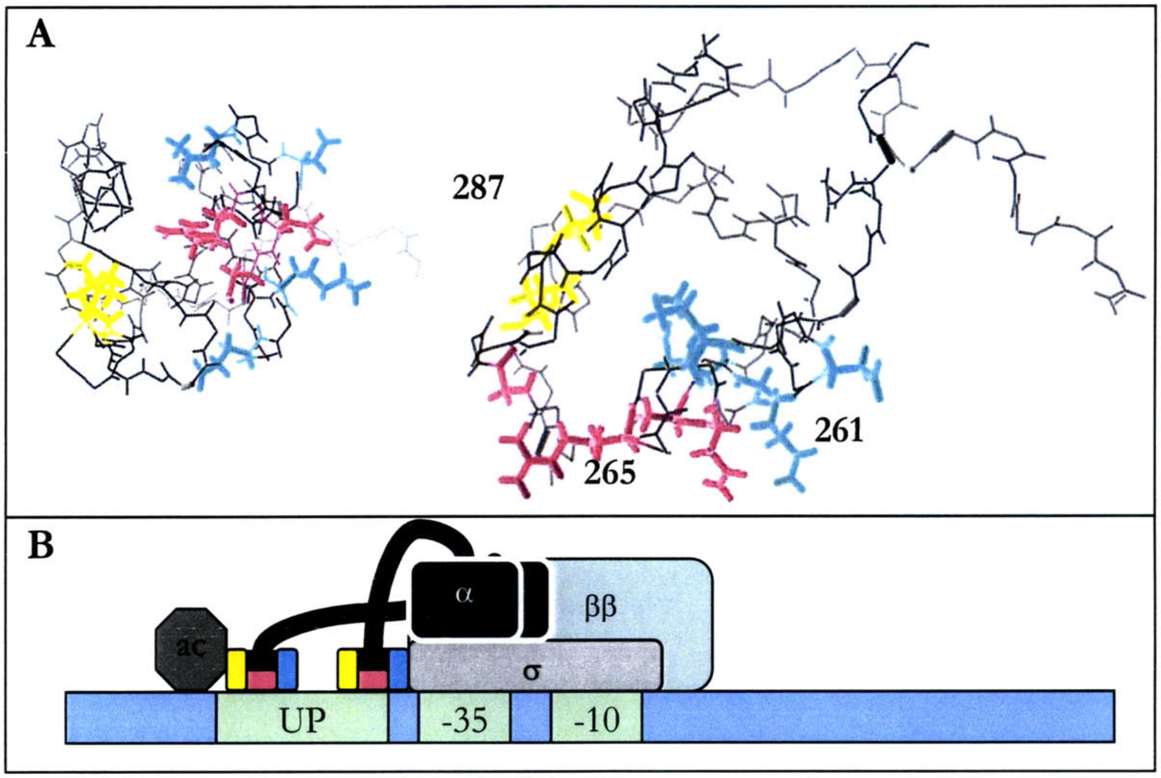

Figure 7.3-3. Location of the determinants mutated in αCTD

(A) The amino acids belonging to the corresponding determinants are shown in the structure of the CTD solved by Jeon *et al.* (1995). (B) Schematic representation of the determinants interacting with different promoter elements. The 261 determinant is shown in contact with the sigma factor, the 265 determinant is shown interacting with DNA at the UP region, and the 287 determinant is shown in contact with an activator ("ac").

7.4 Key Interactions between the RNA Polymerase and the Promoter for Phenotypic Alteration

After gathering the information from all diversity quantification experiments presented in this thesis, the results from the different library designs can be put in a more general context. This information comes from two fronts: the alpha subunit and sigma D subunit libraries. The high diversity of the αCTD*L, and eventually of the αCTD*t libraries, pointed to the surface of the αCTD as a useful determinant for diversity. Previously in this Chapter, we pinpointed further the region with highest regulatory flexibility and ascribed it to the face of the αCTD that interacts with sigma. On the other hand, the optimization experiments with sigma D point to the region 4 of this protein as the most promising one for altering phenotype. In this last Section, we recapitulate the findings and suggest possible implications of the observed trends.

7.4.1 *Residue D259 and the 261 Determinant of the αCTD*

The trends given by considering individual amino acids in the surface of the αCTD, and in particular the high divergence obtained in the case of residue D259, suggest that the diversity introduced by the 261 determinant may be directly or indirectly caused through interactions with sigma. As pointed out earlier, this amino acid in alpha has been implicated in a direct intersubunit interaction with amino acid R603 located in region 4.2 of sigma D, in the vicinity of the contact of the sigma factor with the -35 promoter box (Ross *et al.*, 2003). As noted in Section 4.4, other mutations in this face of the αCTD have also been reported to impact the phenotype globally. In that Section we discussed a paper that noted the pleiotropic effects of a E261K mutation, ranging from inability to grow in minimal media to distinct colony morphology (Jafri *et al.*, 1996).

Considering that *rpoD* is known to control expression of more genes than the number acted upon by activators of αCTD (Martinez-Antonio *et al.*, 2008), an indirect effect through the action of sigma should not be surprising. The result is still significant, however, as the αCTD is known to contact the UP promoter region at most promoters (Ross & Gourse, 2005). The construction and evaluation of targeted libraries of the sigma D region that recognizes and binds the -35 promoter box provides some further proof that the effect of mutating the 261 determinant may be indirect (see Section 7.4.2 below).

Although the hypothesis that the diversity introduced by mutagenesis of the 261 diversity is an indirect consequence of intersubunit interactions at the RNAP-promoter contact seems probable, additional effects may occur concurrently. In fact, the data gathered so far merely indicates that the interaction between the αCTD and sigma is important, but many mechanisms may explain this phenomenon; for example, the αCTD-sigma interaction may alter contacts between regulators and their amino acid targets in region 4 of sigma. Another possibility is that the 261 determinant could contact regulators by itself or simultaneously with sigma. The combination of these mechanisms may result in the high regulatory flexibility we have measured with the use of the divergence metric.

The results offered by the diversity introduced by mutagenesis of the 261 determinant suggest that this face of the protein could be targeted for constructing libraries in the future. The lower diversity of the αCTD*t library compared to that of the -35* libraries (see **Figure 6.6-4**) may be taken to imply that this effort would be in vain. However, this observation is flawed for two reasons. First, the αCTD*t library design allowed mutations in locations other than the 261 determinant, which may act antagonistically to those shown to impart greatest diversity. Second, the αCTD*t library has many amino acids targeted for mutagenesis, so that its search space could not be covered experimentally; the coverage may be significantly increased with a different library design. In fact, a library with mutations restricted to the 261 determinant would explore the significance of both these factors simultaneously.

7.4.2 *The -35-binding region of sigma D*

Optimizing the libraries of sigma D by reducing the search space to regions in charge of transcriptional regulation indicated that region 4 of the protein has most regulatory flexibility. This region is in charge mainly of promoter recognition at the -35 hexamer, but it harbors other regulatory functions as well. For example, it binds activators such as λcI, CRP, RhaS (which controls L-rhamnose metabolism), and others (Dove *et al.*, 2003). It has also been shown to be the target for transcriptional inhibitors, such as the anti-sigma factor Rsd (Jishage & Ishihama, 1998). In addition, region 4 of sigma has been suggested to be regulated through its interaction with the RNAP β-flap, a contact that ensures proper spacing between regions 2 and 4 at the promoter -10 and -35 hexamers (Colland *et al.*, 2001; Dove *et al.*, 2003).

The fact that the principal function of the region that produced the highest phenotypic diversity is contacting the -35 hexamer led us to suggest in the previous Section that the effect of the 261 determinant was caused by an indirect effect through this promoter element. However, the variety of ways in which region 4 functions reiterates the possibility that several effects take place simultaneously in the same cell, though probably the mechanisms that alter the phenotype are distinct at different promoters. If that is the case, then highly targeted mutagenesis of DNA-contacting amino acids, such as that initially proposed during our optimization of the sigma D libraries, would not exploit all possible routes for producing new traits. This is in spite of the improvement in ethanol and hydrolysate tolerance exhibited by mutant E7, which had two substitutions very close to DNA-contacting residues (**Figure 6.6-8**, Section 6.6.3). Also, recall that the S6 mutant isolated in *L. plantarum* libraries has a mutation in a residue located in a DNA-contacting patch of the protein, suggesting that the DNA-binding functions of sigma D in that species are important for phenotypic alteration (**Figure 4.2-4**). Two observations are evidently not sufficient to eliminate the multiple-mechanism hypothesis.

Chapter 8

8. Recommendations and Conclusions

8.1 Summary

We have demonstrated a method for guiding and optimizing the construction of libraries for random search-based strain improvement, and have applied it for the case of transcriptional engineering. The overarching goal was initially to develop tools for advancing this evolutionary approach, but resulted in a method that can be generally applied to any library design.

Firstly, we explored the most practical and simple way of improving transcriptional engineering: we aimed at finding new targets to expand the options of this framework. In addition to using previously-constructed *E. coli* sigma D libraries for improving the production of hyaluronic acid, we investigated other sigma factors, such as sigmas S, E, and H for enhancing traits such as tolerance to carbon dioxide, ethanol, and heat. We were also able to find several new phenotypes with sigma D libraries in a new bacterial species, *L. plantarum*, some with important industrial applications. One new target for transcriptional engineering in *E. coli* proved to be the most promising: the alpha subunit of the RNA polymerase. We successfully used libraries of this protein to improve unrelated phenotypes, such as tolerance to butanol and other solvents, and production of hyaluronic acid and L-tyrosine.

Secondly, we developed a metric to quantitatively compare different targets for transcriptional engineering. The method, based on the premise that highly phenotypically diverse populations are likely to harbor new traits, was first applied to libraries of sigma D and chemical whole-cell mutagenesis in *L. plantarum.* The so-called divergence metric statistically compares the diversity in a measurable phenotype between a library population and a wild-type, clonal population. We showed that this metric correlates qualitatively with the probability that an improved mutant will be found in a library.

Thirdly, we used the divergence metric to optimize transcriptional engineering libraries by successively evaluating and modifying different library designs. We proved that this effort could be used to reduce the search space of a library design so that the overall quality of a population to be screened, as judged by its phenotypic diversity, progressively increases. We applied this idea to the problem of finding a butyrate tolerant mutant in libraries of the alpha subunit of the polymerase, since our initial selection experiments were unsuccessful even though they had delivered other improved phenotypes. This optimization exercise not only resulted in butyrate-tolerant mutants, but it also provided information about which determinants of the alpha subunit had most potential for phenotypic alteration. We also applied our optimization algorithm to libraries of sigma D, and used the resulting populations to improve the survivability of an ethanologenic strain of *E. coli* to biomass hydrolysate and high ethanol conditions.

Lastly, we explored the use of the divergence metric for studying key residues, regions, structures, or functionalities that can confer greatest phenotypic diversity when mutated. We demonstrated that single amino acids can be experimentally considered as determinants of diversity, and explored which locations in the αCTD lead to greatest diversity. We put forth the concept of regulatory flexibility, which refers to the potential of a determinant to create phenotypic diversity by allowing genotypic diversity.

8.2 Conclusions

The results of this thesis permitted us to arrive at the following conclusions:

1. Transcriptional engineering can be performed for strain improvement by mutating various protein targets, in addition to the previously identified sigma factors (mainly sigma D, coded by the gene *rpoD*). In particular, we found great potential for phenotypic alteration in the alpha subunit of the RNA polymerase, coded by the gene *rpoA*.
2. Both *rpoD* and the newly found *rpoA* are good targets for transcriptional engineering, as indicated by (i) the phenotypic diversity of their libraries, and (ii) our ability to obtain several new phenotypes in a few species and strains during the course of this thesis. These include: lactic acid tolerance, increased lactic acid production, low pH tolerance, malic acid tolerance (in *L. plantarum*), butanol tolerance (and general solvent tolerance in the same mutant), enhanced titers and productivity of L-tyrosine, increased hyaluronic acid accumulation, butyrate tolerance (in *E. coli*), ethanol tolerance, increased ethanol productivity, and

tolerance to overlimed bagasse hydrolysate (in an ethanologenic KO11-derivative strain of *E. coli*).

3. Diversity in complex phenotypes can be quantified experimentally and utilized to evaluate the evolutionary potential of strain improvement libraries.
4. The divergence metric derived from such quantification can be used to:
    a. Find and compare targets for mutagenesis.
    b. Restrict mutagenesis to specific regions to probe their potential for affecting phenotype and to study their function.
    c. Optimize the construction of libraries by changing parameters such as size, promoter strength, mutagenesis rate, etc.
    d. Identify key residues, regions, structures, or functionalities that can be targeted for mutagenesis. A comprehensive evaluation at the single-amino acid level opens up the possibility to construct libraries that contain all possible combinations of desired changes.
5. In *L. plantarum*, transcriptional engineering using *rpoD* was more successful at introducing phenotypic diversity, as judged by the divergence metric, than chemical mutagenesis of the entire genome (using NTG at 40-50% killing).
6. Successive evaluation and delimitation of the search space can be achieved by ignoring genetic determinants (or regions) that when altered result in phenotypically redundant variants, but keeping those that result in new phenotypes. This approach led to the construction of optimized *rpoA* and *rpoD* libraries.

7. Thorough study of *rpoA* using the divergence metric led to the following conclusions:

    a. That this target is a useful one for strain improvement in different *E. coli* strains (e.g. K12, DH5α, KO11 derivative, tyrosine-producing parental strain, etc.).

    b. That mutagenesis of the αCTD, with low mutation rate in particular, holds great promise for altering complex phenotypes.

    c. That reducing the search space to surface amino acids in this domain results in a marked increase in divergence.

    d. That expression from the native *spc* promoter leads to improved divergence compared to that from the weaker *lac* promoter.

    e. That the residues in the "261 determinant", responsible for contacting region 4.2 of sigma D, are especially useful for impacting phenotype and thus have highest regulatory flexibility. This fact suggests that an indirect change in specificity, probably through modified interactions with any of the sigmas, results in the observed trends. Other effects related to sigma or αCTD interactions (e.g. with activators, inhibitors, etc.) may play a concurrent role in the high regulatory flexibility.

8. We saw a good correlation between the phenotypic diversity of *rpoA* libraries in complex and simple phenotypes. This indicates that the transcriptional diversity of these libraries is likely due to the alteration of the transcription process itself, and not due to an indirect effect (e.g. nonspecific protein-protein interactions, secondary responses, etc.)

9. A similar approach with *rpoD* led to the following conclusions:
    a. That this target is a useful one for strain improvement in different bacterial species and different strains (e.g. *L. plantarum* and *E. coli* K12, DH5α, KO11 derivative).
    b. That increasing the mutation rate when targeting the entire coding region increases the level of divergence.
    c. That targeting the -10 and -35-binding regions (corresponding to parts of regions 2 and 4 of sigma D, respectively), which are known to be responsible for protein-DNA interactions and thus for much transcriptional regulation, produces the libraries with highest diversity. Furthermore, we learned that increasing the mutagenesis rate to an average of ~4-5 base pair changes per sequence also has a positive effect on diversity.
    d. That targeting the -35-binding region holds greatest potential for altering phenotype, even when comparing with previously isolated and optimized *rpoA* libraries. This fact is also supported by conclusion 7e, above.
10. That amino acids impacting transcriptional regulation through direct contacts are not indisputably desired as targets for mutagenesis, as they may be under strong selection pressures. The result is low genotypic diversity, which in turn leads to low phenotypic diversity.
11. Simultaneous consideration of the divergence data for all libraries constructed so far indicates that the regions of the RNA polymerase in charge of its interaction with the promoter at or close to the -35 element have a larger regulatory flexibility than either of the regions that contact the -10 or UP elements. This may suggest

that the -35 promoter region (with all its DNA and protein elements) has a more important role in differential expression across different operons than the other promoter regions.

8.3 Recommendations and Future Work

The work of this thesis could branch into many directions, depending on the goal or application one is interested in pursuing. In order to keep the recommendations rooted in the results presented here, the list of possible directions given by this Section is restricted to short-term prospects. These are in addition to the many recommendations made throughout this document.

Earlier in our discussion, we framed transcriptional engineering in the context of a variety of random search-based approaches for strain improvement. When the trait of interest is thought to be complex, then the randomization strategy should have the capability of altering the cellular physiology globally. In this way, many nodes of the network can be manipulated at once, giving rise to otherwise unreachable phenotypes. In fact, the divergence metric relies both in the global character of the perturbation and in the interconnectivity of the physiological network. This suggests that other global regulators could be targeted for randomization, and their regulatory flexibility could be measured with the divergence metric. For example, determinants in charge of regulating transcript termination, mRNA degradation or protection (mRNA chaperones), translation initiation, translation elongation, protein degradation, etc. could be considered.

The effort of considering and deciding between targets could be aided by judicious use of the divergence metric. For example, one could evaluate and compare libraries of the several subunits of the transcription machinery of eukaryotes, which is significantly more complex than that of bacteria. One could also compare mechanisms of physiological control by testing targets responsible for different regulatory routes (e.g. compare transcription initiation vs. transcript degradation).

Promising targets should have a characteristic in addition to their ability to perturb the network globally: they should have a discriminatory mechanism to distinguish between the cellular components on which they act. Many of the mutant sigmas that were isolated previous to this work are truncated versions of the protein or have many mutations. The sigma factor has many functions located throughout the protein, and it is unlikely that the previously-isolated variants retain most of these. In fact, expert opinions gathered from scientific meetings during the term of this research seem to suggest that these mutants would not act as transcription factors, but impart the observed phenotype through an indirect effect. Alper and Stephanopoulos (2007) recognize this in their paper, and we do the same for a *L. plantarum rpoD* mutant described in Section 4.2.2.3. These observations, summed to the conclusion that determinants that function in transcription have high regulatory flexibility, lead to the importance of targeting the regions in charge of discriminating between loci. The success of zinc-finger protein-based artificial transcription factors for strain improvement enforces this hypothesis. Therefore, a key recommendation put forth by this Section is to compare the potential of artificial transcription factors to native ones for phenotypic alteration, and decide which is most effective.

Our work with the alpha and sigma subunits already opens several possibilities for future strain improvement approaches through the use of transcriptional engineering. First, because each RNAP complex contains two alpha subunits, cooperation between two mutant alphas could take place within the same RNAP or, alternatively, could result in several versions of RNAP within the same cell. Thus, the introduction of two (or more) mutant alphas could potentially allow exploration of a larger phenotypic space. Second, since the alpha and sigma units regulate promoter preferences by different mechanisms, combining these mutants within the same strain could result in synergistic transcriptional responses unachievable by either subunit tested separately. Although such combinations significantly increase the number of possible libraries, the use of the phenotypic diversity metric could aid in designing better ones, e.g., by increasing or decreasing the mutagenesis rate or reducing the search space (as explained in Chapter 6) in these expanded libraries. Yet another possibility, suggested by the experiments describing promoter replacement of the butyrate-tolerant mutants (Section 6.2.4.2), is to tune the levels of mutant and wild-type subunits in the cell to find an optimal balance between them.

One more recommendation is to use the methods developed by the present thesis to reduce the search space of targets that have been already proven worthwhile. This could be done either in successive steps (like the effort described for the isolation of butyrate tolerant clones), or by considering individual amino acids. For example, this could be applied to improve the quality of libraries of the sigma D, sigma S, and the TATA-binding protein of yeast (the latter described in Alper *et al.*, (2006)). It could also be applied to the 261 determinant of the αCTD. In general, it would be possible, and sometimes desirable, to reduce the search space to the point it can be comprehensively covered experimentally (based on the transformation efficiency of the species of interest).

Finally, the divergence metric could be applied to random search-based libraries to be used for purposes other than strain improvement. For example, it could be used in protein engineering efforts to distinguish determinants that have “enzymatic flexibility” (to borrow the term from that introduced by this thesis). The key remaining challenge would be to develop high throughput assays for the phenotypes that will vary across the members of the protein population to use in the diversity calculation.

Chapter 9

9. Materials and Methods

This Chapter contains the experimental protocols used throughout this thesis. In order to render the previous Chapters comprehensible, they were purposely written without making much reference to the Materials and Methods Section. Therefore, in this Chapter we merely complement the details for those experiments that were not fully described by the main text or the figure legends.

9.1 Reagents and Enzymes

All DNA manipulations, such as genomic DNA isolation, restriction enzyme digestion and ligation, were performed by standard procedures (Sambrook & Russell, 2001a; Sambrook *et al.*, 2006). Some protocols were adapted following the product-specific instructions provided by the manufacturer.

Restriction enzymes, Antarctic phosphatase, and Phusion DNA polymerase were generally obtained from New England Biolabs (Ipswich, MA). Antibiotics and biochemicals such as lysozyme, mutanolysin, and penicillin G, and other organic reagents such as lactic acid were from Sigma-Aldrich (St. Louis, MO). Media was usually from Difco (Sparks, MD). Primers were designed with Vector NTI (version 10.1.1) and ordered either from Invitrogen (Carlsbad, CA) or from Integrated DNA Technologies (Coralville, IA). Fastlink ligase was from Epicentre Biotechnologies (Madison, WI). For error-prone PCR (epPCR), the GeneMorph II Kit from Stratagene (La Jolla, CA) was used according to manufacturer's instructions. For inserting or deleting restriction sites, directed base pair changes were introduced with the QuikChange Multi Site-Directed Mutagenesis Kit (Stratagene). PCR purification of DNA was done with QIAprep kits (Qiagen, Valencia, CA). Gel purification was done using a GeneClean kit (Qbiogene, Morgan Irvine, CA), unless otherwise noted.

9.2 Methods for *L. plantarum*

9.2.1 *Bacterial Strains, Plasmids and Growth Conditions*

L. plantarum was obtained from ATCC (BAA-793) and *E. coli* DH5α for subcloning of some of the plasmids (see below) was obtained from Invitrogen. *Lactobacillus* was routinely grown in MRS (bioMerieux, France) medium and *E. coli* in LB. Media was supplemented with chloramphenicol to 8 μg/mL for *Lactobacillus* and 5 μg/mL for *E. coli* as needed.

Plasmid pGK12 (Kok *et al.*, 1984), obtained from Todd R. Klaenhammer, confers erythromycin and chloramphenicol resistance and was propagated unmethylated in *E. coli* GM1829. Plasmid pDK12 was constructed by inserting the multiple cloning site

(MCS) of plasmid pUC18 into the NsiI and ClaI sites of pGK12. Primers MCSs and MCSa (primer sequences for *L. plantarum* are given in Section 9.2.7 and 9.2.8) were used to amplify the MCS, the PCR product was cut along with pGK12 and the two fragments were ligated.

The new plasmid (pDK12) is capable of alpha-complementation in DH5α. The control plasmid, PDK12D, has the unmutated rpoD gene amplified from *L. plantarum* genomic DNA ((Kleerebezem *et al.*, 2003); NCBI Accession No. AL935257, region 219202-220308) with primers Xma-rpoprom and Xba-rpoterm. The reverse primer includes the transcriptional terminator of the *pln* operon (NCBI Accession No. X94434). The insert and pDK12 were cut with XmaI and XbaI and ligated. The correct structure of pDK12D was confirmed by sequencing (**Figure 9.2-1**).

To co-express the rpoD mutants, we fused them into the same plasmid. They were amplified with either Asc-H13s and H13a or with Asc-S6a and S6s primers so that each was expressed from its own promoter. The first insert was cut with XmaI and AscI and the second with AscI and XbaI. Cut inserts were simultaneously ligated to cut pDK12 and then electroporated into *Lactobacillus*. The correct structure was confirmed by PCR and sequencing with primers pGK12s and pGK12a, located in pDK12 external to the insertion site, and others (see below).

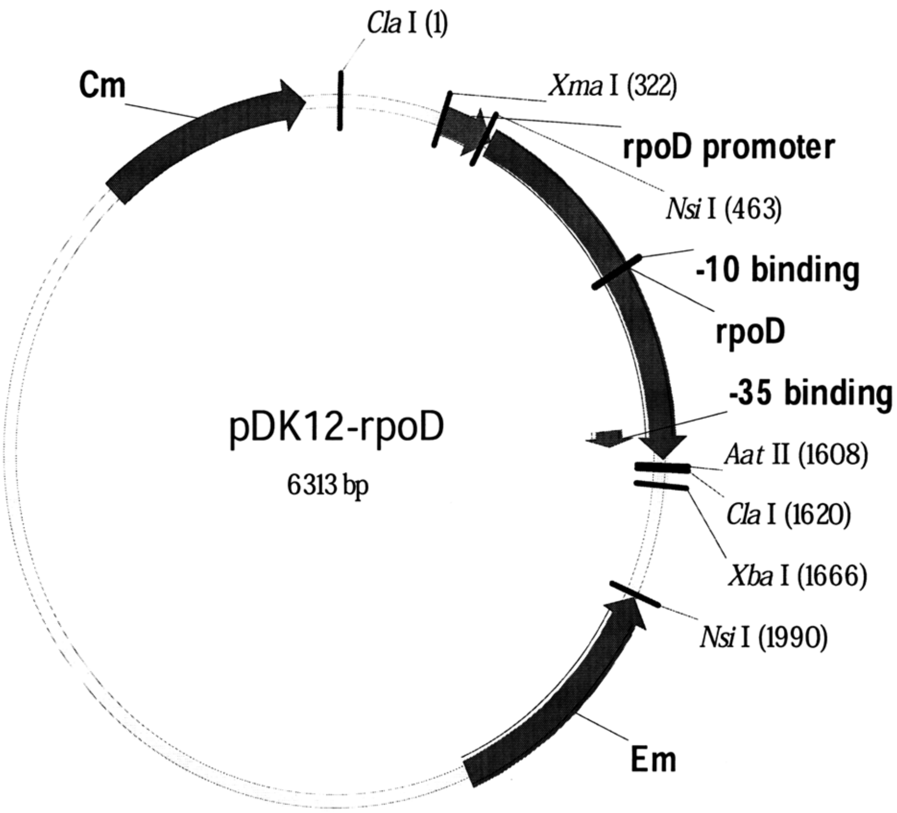

Figure 9.2-1. Schematic illustration of plasmid pDK12D

The restriction sites relevant for the cloning of the MCS and the *rpoD* gene are shown, as well as the relative location of the regions of sigma that bind the -10 and -35 promoter hexamers. The markers for chloramphenicol and erythromycin resistance are also indicated (Cm and Em, respectively).

9.2.2 *DNA Extraction and Purification*

For plasmid extraction, the QIAprep kit was used for both *Lactobacillus* and *E. coli*, except that for *Lactobacillus* the overnight culture (5 mL) was first washed with EDTA buffer (50mM pH 8.0), resuspended in the same (2.4 mL) and lysozyme and mutanolysin were added to a final concentration of 2 mg/mL and 42 U/mL, respectively. The mixture was incubated for at least 1 hr at 37 °C with shaking, and then the plasmid prep protocol was followed using this mixture. Genomic DNA from *L. plantarum* was obtained using an UltraClean microbial DNA isolation kit (Mo Bio Laboratories, Carlsbad, CA) with no pretreatment of the culture. PCR products were purified using the QIAquik kit (Qiagen) prior to restriction and ligation reactions. Gel purification of the products of epPCR was done using a GeneClean kit.

9.2.3 *Transformation by Electroporation*

Transformation efficiency is a key determinant of library size. Therefore, an electroporation protocol previously described (Aukrust *et al.*, 1995; Posno *et al.*, 1991) was optimized prior to library construction. An overnight culture was diluted (1:50) in fresh MRS, incubated with shaking at 37 °C, and penicillin was added to a final concentration of 10 μg/mL after 1 hr of inoculation. The OD_{600} was monitored until it reached ~0.5 (usually 2.5 hr after penicillin addition), and the culture was immediately placed on ice. All subsequent steps were done at 4 °C. The chilled cells were spinned once for 5 min at 1500 × g, washed twice with 3.5X EB (Sucrose 1M, $MgCl_2$ 3.5mM) and then resuspended in 1/100 of the original culture volume. Electroporation was done in a Gene Pulser (Bio-Rad Laboratories, Hercules, CA) at 2.5 kV and 100 Ω, using a 0.2 cm cuvette. Immediately after the pulse, cells were resuspended in 1 mL MRSSM (MRS media supplemented with 1 M sucrose and 100mM $MgCl_2$), grown for 2 hr at 37 °C with shaking, and plated in MRS agar with 8 μg/mL chloramphenicol.

9.2.4 *Library Construction and Phenotype Selection*

Plasmid pDK12D was used as the template for the epPCR reaction, using primers Xma-rpoprom and Xba-rpoterm. Mutation frequency was varied by using different amounts of target; 560 ng for low, 280 ng for medium, and 28 ng for high, as suggested by the manufacturer. The inserts were cut with XmaI and XbaI, gel-purified, and inserted into linearized and dephosphorilated pDK12. The ligation reaction was electroporated into freshly prepared electrocompetent cells as described above. After overnight incubation, the colonies were scraped off from the plates and the liquid libraries were stored at -80 °C until phenotype selection. The total library size was $>10^5$. The NTG library was prepared from an unmutated strain as previously described (Miller, 1972).

Each library was challenged either in 5.5 g/L of L-lactate at an initial pH of 4.60 ± 0.05 (LA condition) or at an initial pH of 3.85 ± 0.05 ($\sim pK_a$) adjusted with HCl without added lactate (HCl condition). The pH was measured using a Symphony pH meter (VWR, West Chester, PA). Libraries were subcultured twice 20-30 hr after inoculation, and then plated to isolate individual clones. The plasmids carrying the mutant sigma factors were extracted, retransformed into fresh cells by electroporation, and the phenotypes were confirmed using the same conditions used for challenging.

9.2.5 *Measurement of Colony Size of* **L. plantarum** *for Diversity Quantification*

Colony area was measured by plating cells in one of four conditions (all in MRS agar with chloramphenicol): 900 mM NaCl (high osmotic pressure), 60 mM HCl, 4 g/L L-lactate, or no stress. Cells were diluted and plated in low enough concentration to be able to distinguish individual clones. Plates were put at 4 °C overnight to stop growth before photographing using an AlphaImager 3400 system (Alpha Innotech, San Leandro, CA). Images were processed using MetaMorph version 6.2 (Molecular Devices, Sunnyvale, CA). All data analysis was done with MATLAB (MathWorks, Natick, MA).

9.2.6 *Fermentations*

Overnight cultures of each clone were diluted in shake-flasks to an OD_{600}=0.02 in either MRS supplemented with glucose to 100 g/L (pH not adjusted) or MRS with no added glucose and initial pH adjusted to 3.85 ± 0.05 with HCl (same as HCl condition described previously). Glucose supplementation was added to ensure that this nutrient was not limiting, following previously established practices (Giraud *et al.*, 1991; Patnaik *et al.*, 2002). L-lactate in the supernatant was measured with a YSI 2700 Select Biochemistry Analyzer (YSI, Yellow Springs, OH) as in (Patnaik *et al.*, 2002).

9.2.7 *Primers for Amplification and Cloning*

All primers are named as follows:

Name: Sequence (5' → 3')

MCSs: GCGCGCATCGATTGAGTGAGCTGATACCGCTCGCC

MCSa: GCGCATGCATCGTCAGCGGGTGTTGGCG

Xma-rpoprom:GCGCCCCGGGTTTGGTTCAGCAGTTAACGTTGGC

Xba-rpoterm: GCGCTCTAGAAAAATAGCCCAAAACCTCGTTAGGA

GATTTTGGGCTATTTTATCGATGGTTAGTCAGACGTCATCATCTGGTGATTAT

Asc-H13s: GGCGCGCCTTTGGTTCAGCAGTTAACGTTGGC

H13a: TAAAACGACGGCCAGTGCCAAG

Asc-S6a: GGCGCGCCAAAATAGCCCAAAACCTCGTTAGGAGATT

S6s: AGGAAACAGCTATGACATGATTACGAATTC

9.2.8 *Primers for Sequencing*

pGK12s: TACTTTTTACAGTCGGTTTTCTAATGTCACTAACCT

pGK12a: AATTGACGATTTAAACAATATTAGCTTTGAACAATT

seq1a: TTTCATCAACAACACTAATCCCAGCA

seq2s: GAAATTGGTCGTGTCGACTTGTTAACG

seq3a: GGCGAATCCACCAAGTCGCG

seq4s: GGTTCGTGAAATCTTGAAGATCGCAC

seq5a: CAATTGGCGTTTCCAATGAAACGG

seq6s: GCAAAGGCAAAAGCAACGACGGAATA

9.3 Methods for *E. coli*

9.3.1 *Construction of sigma D and sigma S Libraries*

9.3.1.1 Error-prone PCR Libraries

A low copy host plasmid (pHACM) was constructed as previously described (Alper and Stephanopoulos, 2007). The genes encoding the sigma D subunit and the sigma S subunit of RNA polymerase, *rpoD* and *rpoS*, respectively, were amplified from *E. coli* genomic DNA, using the following primers: rpoD-F-SacI and rpoD-R-HindIII for sigma D, and rpoS-F-SacI and rpoS-R-HindIII for sigma S (primers for *E. coli* can be found in Sections 9.3.10 and 9.3.11).

Fragment mutagenesis was performed using the Genemorph II Random Mutagenesis kit (Stratagene) with various concentrations of initial template to obtain low, medium, and high mutation rates as described in the product protocol as well as previously described (Alper and Stephanopoulos, 2007). Following the error-prone PCR, the mutated fragments of *rpoD* and *rpoS* were purified using a Qiagen PCR cleanup kit, digested by the respective restriction enzymes overnight (HindIII/SacI), ligated overnight into a digested pHACM backbone, and finally transformed into *E. coli* DH5α competent cells. Cells were plated on LB-agar plates and scraped off to create a liquid library. The total library size was approximately 10^6. The plasmid library was extracted using the Qiagen Miniprep kit (Qiagen) and stored at -80 °C.

For diversity quantification experiments, new *rpoD* libraries were constructed with a similar design, except the epPCR products were gel-purified. This was done to eliminate the possibility of truncated factors that were previously reported.

9.3.1.2 Targeted and Position-Specific Libraries of sigma D

For the position-specific libraries of *rpoD*, primers rpoD-10A and B or rpoD-35A and B were used to amplify the pHACM harboring the gene and cut with either BsaI or ClaI, respectively (in addition to DpnI to digest unamplified vector). The chosen bases (those shown in **Figure 6.6-1**) were spiked so that the total average mutation rate of each library corresponds to that of the αCTD*t library with 6% spiked bases (see Section 9.3.3.3, below). The plasmid was re-circularized overnight by ligation and transformed into DH10B.

For libraries with targeted mutagenesis to the -10 and -35 binding regions, primers rpoD-BsaI-10a and rpoD-BstBI-10s or rpoD-Dra-35s and rpoD-Hind-35a were used for error-prone amplification, respectively. The protocol was adapted so that the two mutation frequencies were similar to the previously-constructed αCTD*L and αCTD*H libraries (see Section 9.3.3.3, below). Then, they were cut with BsaI/BstBI or DraIII/HindIII and cloned into similarly cloned into *rpoD*-bearing pHACM vectors. The resulting libraries were transformed into DH10B, miniprepped, and re-transformed into K12 *recA*$^-$ for diversity quantification.

9.3.2 *Construction of sigma E-sigma H Libraries*

The sigma E and H fused libraries were constructed first by constructing an artificial operon containing the *rpoH* and the *rpoE* genes cloned between the KpnI and MluI sites of the pZE-Q plasmid. The *rpoH* gene was amplified with Phusion polymerase using primers rpoH-KpnI-s and rpoH-ClaI-a and was cut with NEB's ClaI and KpnI, while the *rpoE* gene was amplified with primers rpoE-ClaI-s and rpoE-AscI-a and cut with ClaI and AscI (the latter with complementary overhangs to the MluI site of the plasmid). The resulting fragments were fused and cloned in a triple ligation behind the Q-promoter of plasmid pZE (from Alper *et al.*, 2005).

Error-prone PCR was performed as before on the entire operon with primers rpoH-Kpn1-s and rpoE-AscI-a, but cut with restriction sites KpnI and MfeI. The library of fragments was cloned to the plasmid cut with the same restriction sites, and transformed into *E. coli* DH10B (Invitrogen).

9.3.3 *Construction of the alpha Subunit Libraries*

9.3.3.1 Error-prone PCR Libraries

The native *rpoA* gene was amplified from genomic DNA using Phusion DNA polymerase (NEB, Ipswich, USA) and cloned into the ApaLI and XmaI sites of the multi-cloning site of pHACM (Alper & Stephanopoulos, 2007), using NEB restriction enzymes. Primers rpoA-XmaI-s and rpoA-ApaLI-a were used for cloning. The correct insert was verified by sequencing and strains transformed with this plasmid are denoted 'wild-type' throughout our work.

Error-prone PCR was carried out with the same primers using the GeneMorph II kit (Stratagene) as before, and the mutation frequency was varied by changing the initial amount of target DNA from 700 ng, 250 ng, and 25 ng for rpoA*L (low), rpoA*M (medium), and rpoA*H (high), respectively. After ligation with Fast-link ligase (Epicentre, Madison, USA), the libraries were transformed into DH10B cells (Invitrogen), plated in LB agar, and pooled together after overnight growth. The plasmids were recovered by miniprep (Qiagen) and used to re-transform the three host strains. The original size of the library was approximately 10^5.

9.3.3.2 Saturation Mutagenesis Libraries of the L-tyrosine Producing Mutant (rpoA14)

The saturation mutagenesis library for rpoA14 was constructed with the QuickChange Multi Site-directed mutagenesis kit (Stratagene, La Jolla, USA) by designing primers according to the manufacturer's instructions with degenerate bases to substitute for the codons corresponding to V257 and L281.

9.3.3.3 Targeted and Position-Specific Libraries of the αCTD

For αCTD*H and αCTD*L, a BsiWI restriction site was introduced by a point mutation T707C (slightly upstream of the CTD) using a QuikChange Multi Site-Directed Mutagenesis Kit (Stratagene). The CTD sequence was amplified by error-prone PCR with primers rpoA-B and rpoA-C (resulting in ~5-6 and ~1-2 mutations per sequence, for αCTD*H and αCTD*L respectively) and cloned between the newly-introduced BsiWI and the ApaLI present at the 3'-end.

For the αCTD*t libraries, two oligonucleotides (rpoA-D and rpoA-E) either containing degenerate codons or spiked at the target positions with 6% (or 3%) of non-wild-type bases were constructed, and an artificial BglII site was introduced at the 5'-end of each primer to allow for re-circularization of the plasmid (the BglII site was introduced by a T835A mutation between amino acids E273 and E286). The residues targeted for mutagenesis in αCTD*t were: D259, L262, R265, N268, C269, K271, E273, E286, L290, G296, K298, and S299. The entire plasmid was amplified with Phusion DNA polymerase using the spiked oligonucleotides rpoA-D and rpoA-E and cut with BglII and DpnI to rid the mix of the unmutated plasmid. Neither of the newly-introduced BsiWI nor BglII sites changed the amino acid sequence of *rpoA*.

The exact same protocol was used for the αCTD*t library expressed from the P_{spc} promoter (Post *et al.*, 1978), except a pCL1920 vector was used (Lerner & Inouye, 1990); the *rpoA* gene was first cloned using primers rpoA-F and rpoA-G, which include the P_{spc} promoter and T1 terminator, respectively, for efficient use in the pCL1920 vector.

9.3.3.4 Saturation Mutagenesis of Individual Amino Acids in αCTD

Degenerate codons were substituted at the specified locations using the QuikChange Multi Site-Directed Mutagenesis Kit from Stratagene, with primers rpoA-R265, rpoA-V242, rpoA-D259, etc. The libraries were isolated by miniprep from XL10-Gold (from the kit), and retransformed into K12 *recA*$^{-}$ for diversity quantification.

9.3.4 *Hyaluronic Acid Methods*

9.3.4.1 Host Strain

Hyaluronic acid production experiments were completed with recombinant *E. coli* Top10 /pMBAD-*sseABC* in LB liquid medium supplemented with Mg^{2+}, using *L*-arabinose as inducer. Plasmid pMBAD (Yu & Stephanopoulos, 2007) was constructed by the introduction of a 62 bp multi-cloning site (MCS) sequence containing XbaI-BamHI-StuI-KpnI-SacI-EcoRI-HindIII restriction sites into the plasmid of pBAD (Invitrogen) with an ampicillin resistance marker. *E. coli* Top10 (Invitrogen) was used as the expression host of the plasmid pMBAD-*sseABC*, which was constructed by the insertion of the fragment *sseABC* into the backbone of pMBAD (Yu & Stephanopoulos, 2007). The *sseABC* operon consists of the genes *sehasA*, *hasB* and *hasC*. The *sehasA* was synthesized by assembly PCR (Hoover & Lubkowski, 2002) according to the protein sequence of the HA synthase from *Steptococcus equisimilis* (NCBI-AAB87874.1, GI:2655100). The functionality of *hasB* and *hasC* were provided by the genes *ugd* and *galF* of *E. coli* K12 MG1655, coding for the UDP-glucose 6-dehygrogenase and the glucose-1-P uridyltransferase, respectively. *E. coli* Top10 /pMBAD-*sseABC* is an *L*-arabinose inducible recombinant strain for HA production (Yu & Stephanopoulos, 2007), while *E. coli* Top10 /pMBAD was used as the null control. *E. coli* DH5α (Invitrogen) was used for routine transformations as described in the protocol. An approximately equal concentration of the plasmid libraries was transformed into *E. coli* Top10 /pMBAD-*sseABC* by electroporation and plated on selective plates after dilution.

9.3.4.2 High-Throughput Screen and Quantification with Alcian Blue

LBSMA (Bellemann *et al.*, 1994) medium (LB Medium supplemented with sorbitol, $MgCl_2$, ampicillin and *L*-arabinose) was used for the translucent colony identification step. This constituted the first stage of the screen, and was followed by quantification with the alcian blue dye.

The alcian blue solution was prepared by the following procedure: 1.0 g alcian blue 8GX (Sigma Aldrich) was dissolved in 100 ml 3% glacial acetic acid and the pH was adjusted to 2.5 using acetic acid (WebPath: Internet Pathology Laboratory). The solution was filtered through a 0.45 μm syringe filter (VWR, USA), and a crystal of thymol was added. It was stored at room temperature and found to be stable for 6 months. The optimized procedure for high throughput HA quantification is as follows: 400 μL of fermentation broth containing HA was aliquoted into a 1.5 mL centrifuge tube pre-filled with 550 μL 3% acetic acid. Then, 50 μL Alcian blue solution was added followed by vortexing, and the mixture was microwaved for 30 seconds; after centrifugation, the tube was cooled at room temperature for 2.5 h. The solution was centrifuged at 10,000 rpm for 1 min, and 200 μL of supernatant were loaded into a 96-well plate, and the OD_{540} was measured using the plate reader. A standard curve was generated using 400 μL of 50, 100, 200, 300 and 500 mg/L commercial HA standards (VWR). All experiments were repeated 3 times except where specifically noted.

9.3.4.3 HA Quantification by HPLC

HA titers were measured by a modified HPLC method (Kakizaki *et al.*, 2002). Fermentation broth samples were incubated first with an equal volume of 0.1% w/v sodium-dodecyl-sulfate (SDS) at room temperature for 10 min to free the capsular HA

(Chong & Nielsen, 2003). Subsequently, the HA product was precipitated out from the medium samples with 1.5 volumes of ethanol (Ogrodowski *et al.*, 2005) incubating at 4 °C for 1 h. The precipitate was collected by centrifugation (2,000 *g* for 20 min at room temperature) and resuspended in 1 volume of 0.2 M NaCl for 10 min. Then the re-dissolved samples were centrifuged for 8 min at 3000 *g*, filtered through a 0.45 μm syringe filter (VWR), and applied to the modified HPLC assay. Gel Filtration Chromatography (GFC) in combination with a UV photodiode array detector (Waters 2695-996) was used to determine the concentration of the HA products in the broth. The column was a model Shodex SB-806M OHpak (8×300mm, Thompson, USA) supporting molecular weight (MW) analyses from 10^3 to 2×10^7 Da. HA products with MW of 6.8×10^5 Dalton, purchased from Lifecore Biomedical Inc., were prepared into around 300 mg/L aqueous standards in 0.2 M NaCl. The detection was carried out at wavelength of 206 nm and room temperature, with 0.2 M NaCl as the effluent buffer at flow rate of 0.5 mL/min.

9.3.5 *L-Tyrosine Methods*

9.3.5.1 Host Strain and Screening

L-tyrosine production experiments used a parental strain of *E. coli* K12 Δ*pheA* *tyrR*::$\mathrm{P_{LTET\text{-}O1}}$ *tyrA*$^{\mathrm{fbr}}$*aroG*$^{\mathrm{fbr}}$ *lacZ*:: $\mathrm{P_{LTET\text{-}O1}}$ *tyrA*$^{\mathrm{fbr}}$*aroG*$^{\mathrm{fbr}}$ /pTrcmelA$^{\mathrm{mut1}}$ ((Lutke-Eversloh & Stephanopoulos, 2005; Lutke-Eversloh & Stephanopoulos, 2007) and C.N.S. Santos unpublished data). These were performed at 37°C with 225 rpm orbital shaking in 50 mL

MOPS minimal medium (Teknova, Hollister, USA) cultures supplemented with 5 g/L glucose and an additional 4 g/L NH_4Cl.

Libraries of L-tyrosine production mutants were constructed by transforming the parental strain *E. coli* strain with the *rpoA* libraries. Approximately 7.5x10^5 viable colonies were obtained and subsequently screened for L-tyrosine production as described previously (Santos & Stephanopoulos, 2008b).

9.3.5.2 Quantification of L-Tyrosine

Cell-free culture supernatants were filtered through 0.2 μm PTFE membrane syringe filters (VWR International) and used for HPLC analysis with a Waters 2690 Separations module connected with a Waters 996 Photodiode Array detector (Waters) set to a wavelength of 278 nm. Samples were separated on a Waters Resolve C18 column with 0.1 % (vol/vol) trifluoroacetic acid (TFA) in water (solvent A) and 0.1 % (vol/vol) TFA in acetonitrile (solvent B) as the mobile phase. The following gradient was used at a flow rate of 1 ml/min: 0 min, 95 % solvent A + 5 % solvent B; 8 min, 20 % solvent A + 80 % solvent B; 10 min, 80 % solvent A + 20 % solvent B; 11 min, 95 % solvent A + 5 % solvent B.

9.3.6 *Selection Experiments*

9.3.6.1 Carbon Dioxide Tolerance

Selection of CO_2-tolerant mutants in sigma S libraries consisted of serial subculturing of the libraries in MOPS media prepared with carbonated water. Capped 32-mL borosilicate glass tubes were purchased from VWR to hold the pressure and CO_2 inside the tube. This water was purchased pressurized to ~250 kPa (according to manufacturer) from Polar Beverages, Inc. (MA, USA). The composition of the media was as follows:

Media component	mL added
MOPS 10X	3.2
Glucose 200 g/L	0.32
Yeast extract 50 g/L	0.32
Thiamine, 1M	0.032
Dipotassium phosphate (4M)	0.01056
Adjust volume to:	4
Carbonated water added	28

The libraries were subcultured approximately every 12 hrs for 6-8 rounds, at which there was appreciable and perdurable enrichment (>10-20%) of a library compared to the control. At this point, cells were plated in rich media supplemented with antibiotics to isolate individual clones. These were tested prior and subsequent to re-transformation in order to confirm the phenotype of interest.

9.3.6.2 Heat and Ethanol Tolerance

We screened the epPCR rpoH-rpoE libraries in 50 g/L of ethanol at 42 °C. The challenge involved resuspending the cells in LB media subject to these conditions and plating them after 24 hr. Single clones were chosen from the plates with highest survivability, and they were tested individually in the same conditions used for selection. Two controls were used for comparison, one bearing an empty plasmid, and one bearing plasmid with a non-mutated operon.

9.3.6.3 Butanol and Solvent Tolerance

DH5α cells transformed with the alpha subunit libraries (5x10^6 colonies were obtained) were pooled together, and cultured on LB liquid medium at 37 °C with shaking for 2 hr before placing them under selection conditions. Selection for butanol tolerance was carried out in screw-cap shake flasks in 0.9% butanol (v/v). The culture was grown overnight, and used to re-inoculate fresh selection medium. This process was repeated twice before plating cells to test individual colonies. For verification of the phenotype, plasmids were isolated and reintroduced into a clean background, grown overnight in LB and inoculated to OD=0.1 in 5 mL of medium in the presence of different concentrations of 1-butanol (1C4), 2-butanol (2C4), 3-pentanol (3C5) or 1-pentanol (1C5). Tubes were sealed with parafilm to avoid evaporation of alcohols.

9.3.6.4 Butyrate Tolerance

MOPS medium with 15 g/L butyrate was used for both selection and growth assays (initial pH adjusted to 7.0 with 6N HCl), except when trying the conditions described in **Figure 6.2-3**. For selection, 30 mL of media were inoculated and cells were grown for about 20-24hr, then a sample was transferred to a fresh batch of media. This procedure was repeated thrice, after which cells were spread in solid media overnight and individual colonies were picked for further study. Clones #1 and #16 in αCTD*L were chosen for their faster growth in butyrate, and their plasmids were purified and re-transformed into a clean K12 *recA*$^-$ background to confirm the phenotype (**Figure 6.2-5**).

For growth assays, cells were cultured overnight in 15g/L butyrate to avoid adaptation-related distortion of the first few measurements and then diluted in the same media to obtain their growth curves. The mutant genes from clones #1 and #16 and the wild-type *rpoA* were transferred to a pCL1920 plasmid (which has the same origin of replication than pHACM, but confers streptomycin resistance, (Lerner & Inouye, 1990)) and expressed from the P_{spc} promoter (Post *et al.*, 1978). Primers rpoA-F and rpoA-G were used as explained in Section 9.3.3.3 above.

9.3.6.5 Ethanol and Hydrolysate Tolerance of KO11 Derivative

Mutants in the αCTD*t libraries were selected by subjecting the cells to 40 g/L for 48 hr in media containing 5% corn-steep liquor (CSL) and 25 g/L glucose at 37 °C. After stressing them, cells were plated in LB medium supplemented with antibiotics. Individual retransformed clones were tested for their ability to produce ethanol in overlimed bagasse hydrolysate (provided by Verenium, San Diego , CA) with 5% CSL, and pH-adjusted to 6.5 with HCl (6N). Xylose was added initially to obtain 100 g/L of total sugars, and was also added in the fed-batch phase using a 50 % sterile solution at an equivalent rate of about 2 g/L hr.

Mutants in the *rpoD* -35*H and -10H libraries were directly selected for survivors in overlimed bagasse hydrolysate with 5% CSL and 50 g/L ethanol, adjusted to pH of 6.5. Dilutions were plated in LB medium initially and at several time points to quantify survivability of libraries. The same conditions and procedures were used to test individual clones before and after re-transformation.

9.3.7 *Intracellular pH Determination for Measuring Ion Leakage of Butanol-Tolerant Mutant*

The intracellular pH was monitored by expressing a pH-responsive GFP (a present from G. Miesenbock and J. Rothman, (Miesenbock *et al.*, 1998)). The response of the GFP variant was monitored as the excitation ratio at 395 nm over 475 nm, and the emission was measured at 530 nm. A standard curve was constructed by resuspending DH5α cells in various buffers as described previously (Karagiannis & Young, 2001), without a carbon source for 45 min at 37 °C. For pH_i difference quantification, cells were grown in LB and resuspended in potassium phosphate buffer (50 mM, $pH_e = 4.7$) with 0.5% glucose as an energy source in the presence and absence of butanol.

9.3.8 *Intracellular pH Determination for Phenotypic Diversity Quantification*

All libraries were quantified for diversity in an *E. coli* K12 *recA$^-$* background. We used the intracellular pH in growing and non-growing cells as phenotypes contained in the divergence metric (see Chapter 5 for equations). For determination of pH_i during growth, cells were stained with CFSE (Invitrogen) as suggested by the product manual and grown in MOPS media with 250 mg/L of each D-xylose, D-galactose, L-arabinose, and glycine. Several carbon sources were used to prevent favoring the growth of a subset of mutants, while at the same time allowing for full induction of the plasmid-borne *rpoA*. We kept the same conditions when expressing this gene from the constitutive promoter or for the case of *rpoD,* in order to allow for a fair comparison. Variability introduced by the choice of carbon sources or other details in the protocol was accounted for by normalization with the control.

Media was withdrawn at different time points from each library and control cultures, put on ice and measured by flow cytometry (using a BD FACScan). The pH_i was calculated as the ratio of 585 to 530 nm emission when excited at 488 nm (Spilimbergo *et al.*, 2005). Each time point was considered an entry in the distance vector for quantification of divergence.

Two more entries of the distance vector were composed of pH_i values in non-growing cells, except when indicated in the main text. For the case of non-growing pH_i determination, cells were stained with BCECF-AM (Invitrogen), and resuspended in 10mM phosphate buffer at either pH 5.0 or 7.0 immediately before FACS analysis (pH_i with this probe was calculated as the ratio of 650 to 530 nm emission when excited at 488 nm, as per manual recommendations). A sub-sample of 1500 data points was taken at random from each library and control data sets, and this sub-set was used to calculate the divergence; the algorithm was run 50 times and the divergence was averaged to smooth out the effects of sub-sampling.

For saturation mutagenesis libraries of individual residues in the αCTD, the quantification protocol was modified to deal with the fact that different amino acid positions are likely to play a more important role in some growth conditions than in others (Benoff *et al.*, 2002; Fritsch *et al.*, 2000; Lochowska *et al.*, 2004). Therefore, we measured the pH_i in different media that varied on the choice of carbon source, on the absence or presence of amino acids, and availability of complex nutrients. Four conditions were used: LB, MOPS supplemented with L-arabinose instead of glucose, MOPS-L-arabinose supplemented with 0.5% casamino acids, and MOPS with mixed carbon sources as used for the other libraries.

9.3.9 *Phenotypic Diversity Quantification at a Single Locus using GFP*

The plasmids bearing the libraries were transformed into a K12 *recA*$^-$ strain that harbored an pKLJ03 plasmid (Jones & Keasling, 1998), modified to express an unstable variant of GFP (Andersen *et al.*, 1998). The fluorescent protein was cloned downstream of the P_{trc} promoter of pKLJ03. After transformation, the cells were plated in M9 minimal media with galactose (instead of glucose) and 1mM IPTG to fully induce both the plasmid-borne *rpoA* and the GFP genes.

Fluorescence of the library populations was determined by flow cytometry (using a BD FACScan) by exciting at 488 nm. The emission from the FL1 channel (at 530 nm) was taken as the phenotype for diversity quantification.

9.3.10 *Primers for Amplification and Cloning*

All primers are named as follows:

Name: Sequence (5' → 3')

9.3.10.1 Sigma D and Sigma S

rpoD-F-SacI: AACCTAGGAGCTCTGATTTAACGGCTTAAGTGCCGAAGAGC

rpoD-R-HindIII: TGGAAGCTTTAACGCCTGATCCGGCCTACCGATTAAT

rpoS-F-SacI: AACCTAGGAGCTCAGACTGGCCTTTCTGACAGATGCTTACT

rpoS-R-HindIII: AACCTAGGAGCTCAGACTGGCCTTTCTGACAGATGCTTACT

(A star implies the preceding base is either spiked or fully randomized)

rpoD-10A: GCATATGATTGAGACCATCAACAAGCTCAACCGTATTTCTCG

rpoD-10B: TTGTTGATGGTCTCAATCATATGCACCGGAATACGGATGGTGCGC
GCCTGATCCGCGATAGAG*C*G*G*T*GATCGCC*T*G*ACGGATCCAC*C*
A*GGTTGCG*T*A*GGTGGAGAACTTGTAACCACGGCGGTATTC

rpoD-35A: TTTCGGTATCGATATGAACACCGACTACACGC*T*G*G*A*A*GAAG
TGGGTAAACAGTTCGACGTTA*C*C*C*G*C*G*A*A*C*G*T*ATCC*G*T*C*A*
G*ATCGAAGCGA*A*G*GCGCTGCGCAAACTGCGTCACCCG

rpoD-35B: TGTTCATATCGATACCGAAA*C*G*CATACGCAGAACTTTTGCTTC
A*C*G*CGCGGTCAGGCCAGCCAGCACG

rpoD-BsaI-10a: CGGTTGAGCTTGTTGATGGTCTCAATC

rpoD-BstBI-10s: GATGAAAGCGGTTGATAAATTCGAATACC

rpoD-Dra-35s: GGCAACGCACGACGTGCTGG

rpoD-Hind-35a: CTATGACCATGATTACGCCAAGCTTTAACG

9.3.10.2Sigma E and Sigma H

rpoE_AscI-a: GCGCCCGGCGCGCCCCCAATTGCCACGCCTGATAAGCG
GTTGAACTTTGTT

rpoE_ClaI-s: CGCGCTAAATCGATATGAGCGAGCAGTTAACGGAC

rpoH_ClaI-a: CACACCTATCGATTTACGCTTCAATGGCAGCA

rpoH_KpnI-s: GAGAAAGGTACCATGACTGACAAAATG

9.3.10.3Alpha Subunit

rpoA_ApaLI-a: GCGCGGTGCACTGGCGCATGACCTTATCCTTCTCAGTA rpoA_XmaI-s: GCGCGCCCGGGACGTTGTAAGCATTCGTGAGAAAGCG

rpoA-B: GCGCGGTGCACTGGCGCATGACCTTATCCTTCTCAGTA

rpoA-C: ACGTGACGTACGTCAGCCTGAAGTGAAAGAAGAGAAACC

(A star implies the preceding base is either spiked or fully randomized)

rpoA-D: TATCGGAGATCTGGTACAGCGTACCG*A*G*GTT
GAGCTCC*T*T*AAAACGCCTAACCTTG*G*T*AAAA*A*A*T*C*T*CTTACTGA
GATTAAAGACGTGCTGGCTTCCCGT

rpoA-E: TGTACCAGATCTCCGATATAGTGGATACGT*T*C*TGCT*T*T*AA
GG*C*A*G*T*T*AGCAGAG*C*G*GACAGTC*A*A*TTCCAGA*T*C*GTCAACA
GGGCGCAGCAGGATCGGAT

rpoA-F: GCGAGCGATCTAGACTCAGAAATGAGCCGTTTATTTTTCTACCC
ATATCCTTGAAGCGGTGTTATAATGCCGCGCCCTCGATATGGGGATTTTTGTG
T ATGCTGGCAAGATGGAAGGTACGTTTA AG

rpoA-G: CGGCGCGCCCGGGTTTATAAAACGAAAGGCCCAGTCTTTCGACTG
AGCCTTTCGTTTTATGTGCACTGGCGCATGACCTTATCCTTCTC

rpoA-R265: CGATCTGGAATTGACTGTCNNSTCTGCTAACTGCCTTAAAGCAG

rpoA-E286: CTGGTACAGCGTACCNNSGTTGAGCTCCTTAAAACGCC

rpoA-L290: CGTACCGAGGTTGAGCTCNNSAAAACGCCTAACCTTGG

rpoA-D259: CCTGCTGCGCCCTGTTGACNNSCTGGAATTGACTGTCC

rpoA-L262: CCTGTTGACGATCTGGAANNSACTGTCCGCTCTGC

rpoA-N268: GGAATTGACTGTCCGCTCTGCTNNSTGCCTTAAAGCAGAAGC

rpoA-C269: GACTGTCCGCTCTGCTAACNNSCTTAAAGCAGAAGCTATCC

rpoA-K271: GTCCGCTCTGCTAACTGCCTTNNSGCAGAAGCTATCCAC
TATATCG

rpoA-E273: CTGCTAACTGCCTTAAAGCANNSGCTATCCACTATATCGG

rpoA-V242: CTTACGTGATGTACGTCAGCCTGAANNSAAAGAAGAGAAACCA
GAGTTC

9.3.11 *Sequencing Primers*

9.3.11.1 Sigma D

seq1a: ACCCGATGTCCTTCAGCCGTTAA

seq2s: CAGATCAGATCGAAGACATCATCCAAATG

seq3a: CAATTTCGCCTTCGCGGGTCA

seq4s: GCCACTCACGTCGGTTCTGAGCTT

seq5a: GACAGTTTCAGGATCTCTTCCTGAGCG

seq6s: CCAGCGATACCTGGTTCAACGC

seq7a: AACCAGACGTAAGTTCGCTTCAACCATC

seq8s: GATCAGGCGCGCACCATCC

seq9a: AGATGCGAATCTTCATCATCACCGA

seq10s: AAACAGTTCGACGTTACCCGCGAA

9.3.11.2 Sigma S

seq1s: CGAAGAGGAACTGTTATCGCAGGG

seq2a: TTACTCTCGATCATCCGGCGGC

seq3s: ATACGCAACCTGGTGGATTCGCC

seq4a: TCCAGTTGCTCTGCGATCTCTTCC

seq5s: ACCACGCAAGATGACGATATGAAGCAGAG

seq6a: CTCGCGGAACAGCGCTTCG

9.3.11.3 Sigma E

seq1s: ATGAGCGAGCAGTTAACGGACC

seq2a: GCACTACCAGTAAGTTAAAGGCTTTCTGA

seq3s: GCTATGTGCCGTCGGGTGATG

seq4a: CGCCCCTGAGCAACCAGGTAAT

seq5s: AGTCCCTCCCGGAAGATTTACGC

seq6a: ACGCCTGATAAGCGGTTGAACTTTG

9.3.11.4 Sigma H

seq1s: ATGACTGACAAAATGCAAAGTTTAGCTTTA – 3’

seq2a: GCCATGGTAATGCAGCTTTTCAGC – 3’

seq3s: CTATGGCCTGCCACAGGCG – 3’

seq4a: GCTTTGATCCAGTGAACGGCGA – 3’

seq5s: GTAACCAGCAAAGACGTACGTGAGATGG – 3’

seq6a: CATCTTCAATGCCGTCGGCAA – 3’

seq7s: GCGCGCTGGCTGGACG – 3’

seq8a: CGCTTCAATGGCAGCACGCAAT – 3’

9.3.11.5 Alpha Subunit

seq1s: GCTTTACTCCAAGTAAAGCTTAGTACCAAAGAGAG – 3’

seq2a: ATTTCCAGGATATCTTCCTGAACGCCT – 3’

seq3s: CGGTGATGTCGAAATCGTCAAGC – 3’

seq4a: CTCTTCAGGATCGATTGTGCCGTT – 3’

seq5s: GTTGACGATCTGGAATTGACTGTCCG – 3’

seq6a: GCGCGGTTTTAGAAACTCTGTCACA – 3’

Chapter 10

10. References

Abdel-Fattah, W. R., Fadil, M., Nigam, P. & Banat, I. M. (2000). Isolation of thermotolerant ethanologenic yeasts and use of selected strains in industrial scale fermentation in an Egyptian distillery. *Biotechnol Bioeng* **68**, 531-535.

Acedo-Felix, E. & Perez-Martinez, G. (2003). Significant differences between Lactobacillus casei subsp. casei ATCC 393T and a commonly used plasmid-cured derivative revealed by a polyphasic study. *Int J Syst Evol Microbiol* **53**, 67-75.

Ades, S. E. (2004). Control of the alternative sigma factor sigmaE in Escherichia coli. *Curr Opin Microbiol* **7**, 157-162.

Alegre, M. T., Rodriguez, M. C. & Mesas, J. M. (2004). Transformation of Lactobacillus plantarum by electroporation with in vitro modified plasmid DNA. *FEMS Microbiol Lett* **241**, 73-77.

Alper, H. & Stephanopoulos, G. (2007). Global transcription machinery engineering: a new approach for improving cellular phenotype. *Metab Eng* **9**, 258-267.

Alper, H. & Stephanopoulos, G. (2008). Uncovering the gene knockout landscape for improved lycopene production in E. coli. *Appl Microbiol Biotechnol* **78**, 801-810.

Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. (2005). Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A* **102**, 12678-12683.

Alper, H., Moxley, J., Nevoigt, E., Fink, G. R. & Stephanopoulos, G. (2006). Engineering yeast transcription machinery for improved ethanol tolerance and production. *Science* **314**, 1565-1568.

Altaras, N. E. & Cameron, D. C. (1999). Metabolic engineering of a 1,2-propanediol pathway in Escherichia coli. *Appl Environ Microbiol* **65**, 1180-1185.

Altaras, N. E. & Cameron, D. C. (2000). Enhanced production of (R)-1,2-propanediol by metabolically engineered Escherichia coli. *Biotechnol Prog* **16**, 940-946.

Andersen, J. B., Sternberg, C., Poulsen, L. K., Bjorn, S. P., Givskov, M. & Molin, S. (1998). New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Appl Environ Microbiol* **64**, 2240-2246.

Anderson, A. J. & Dawes, E. A. (1990). Occurrence, metabolism, metabolic role, and industrial uses of bacterial polyhydroxyalkanoates. *Microbiol Rev* **54**, 450-472.

Annous, B. A. & Blaschek, H. P. (1991). Isolation and characterization of Clostridium acetobutylicum mutants with enhanced amylolytic activity. *Appl Environ Microbiol* **57**, 2544-2548.

Antoniewicz, M. R., Kelleher, J. K. & Stephanopoulos, G. (2007). Elementary metabolite units (EMU): a novel framework for modeling isotopic distributions. *Metab Eng* **9**, 68-86.

Applebee, M. K., Herrgard, M. J. & Palsson, B. O. (2008). Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. *J Bacteriol* **190**, 5087-5094.

Atsumi, S., Hanai, T. & Liao, J. C. (2008a). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, 86-89.

Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., Smith, K. M., Brynildsen, M. P., Chou, K. J., Hanai, T. & Liao, J. C. (2008b). Metabolic engineering of Escherichia coli for 1-butanol production. *Metab Eng* **10**, 305-311.

Aukrust, T. W., Brurberg, M. B. & Nes, I. F. (1995). Transformation of Lactobacillus by electroporation. *Methods Mol Biol* **47**, 201-208.

Azcarate-Peril, M. A., Altermann, E., Hoover-Fitzula, R. L., Cano, R. J. & Klaenhammer, T. R. (2004). Identification and inactivation of genetic loci involved with Lactobacillus acidophilus acid tolerance. *Appl Environ Microbiol* **70**, 5315-5322.

Baltz, R. H. (2001). Genetic methods and strategies for secondary metabolite yield improvement in actinomycetes. *Antonie Van Leeuwenhoek* **79**, 251-259.

Baltz, R. H. & Seno, E. T. (1981). Properties of Streptomyces fradiae mutants blocked in biosynthesis of the macrolide antibiotic tylosin. *Antimicrob Agents Chemother* **20**, 214-225.

Barbosa, M. F., Yomano, L. P. & Ingram, L. O. (1994). Cloning, sequencing and expression of stress genes from the ethanol-producing bacterium Zymomonas mobilis: the groESL operon. *Gene* **148**, 51-57.

Bartsevich, V. V. & Juliano, R. L. (2000). Regulation of the MDR1 gene by transcriptional repressors selected using peptide combinatorial libraries. *Mol Pharmacol* **58**, 1-10.

Becker, G. & Hengge-Aronis, R. (2001). What makes an Escherichia coli promoter sigma(S) dependent? Role of the -13/-14 nucleotide promoter positions and region 2.5 of sigma(S). *Mol Microbiol* **39**, 1153-1165.

Beerli, R. R. & Barbas, C. F., 3rd (2002). Engineering polydactyl zinc-finger transcription factors. *Nat Biotechnol* **20**, 135-141.

Bellemann, P., Bereswill, S., Berger, S. & Geider, K. (1994). Visualization of capsule formation by Erwinia amylovora and assays to determine amylovoran synthesis. *Int J Biol Macromol* **16**, 290-296.

Benner, S. A. & Sismour, M. (2005). Synthetic biology. *Nature Reviews Genetics*, 533-543.

Benoff, B., Yang, H., Lawson, C. L., Parkinson, G., Liu, J., Blatter, E., Ebright, Y. W., Berman, H. M. & Ebright, R. H. (2002). Structural basis of transcription activation: the CAP-alpha CTD-DNA complex. *Science* **297**, 1562-1566.

Bernal, P., Munoz-Rojas, J., Hurtado, A., Ramos, J. L. & Segura, A. (2007). A Pseudomonas putida cardiolipin synthesis mutant exhibits increased sensitivity to drugs related to transport functionality. *Environ Microbiol* **9**, 1135-1145.

Betengaugh, M. & Bentley, W. (2008). Metabolic engineering in the 21st century: Meeting global challenges of sustainability and health. *Current Opinion in Biotechnology*, 411-413.

Biles, B. D. & Connolly, B. A. (2004). Low-fidelity Pyrococcus furiosus DNA polymerase mutants useful in error-prone PCR. *Nucleic Acids Res* **32**, e176.

Bock, G., Huber, L. A., Wick, G. & Traill, K. N. (1989). Use of a FACS III for fluorescence depolarization with DPH. *J Histochem Cytochem* **37**, 1653-1658.

Bonomo, J., Lynch, M. D., Warnecke, T., Price, J. V. & Gill, R. T. (2008). Genome-scale analysis of anti-metabolite directed strain engineering. *Metab Eng* **10**, 109-120.

Booth, I. R. (1985). Regulation of cytoplasmic pH in bacteria. *Microbiol Rev* **49**, 359-378.

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M. & Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815-819.

Botsford, J. L. & Harman, J. G. (1992). Cyclic AMP in prokaryotes. *Microbiol Rev* **56**, 100-122.

Braun, V., Mahren, S. & Sauter, A. (2006). Gene regulation by transmembrane signaling. *Biometals* **19**, 103-113.

Brink, M. F., Verbeet, M. P. & de Boer, H. A. (1995). Specialized ribosomes: highly specific translation in vivo of a single targetted mRNA species. *Gene* **156**, 215-222.

Browning, D. F. & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**, 57-65.

Brunner, M. & Bujard, H. (1987). Promoter recognition and promoter strength in the Escherichia coli system. *Embo J* **6**, 3139-3144.

Burgard, A. P., Pharkya, P. & Maranas, C. D. (2003). OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering* **84**, 647-657.

Burgess, R. R. & Anthony, L. (2001). How sigma docks to RNA polymerase and what sigma does. *Curr Opin Microbiol* **4**, 126-131.

Busby, S. & Ebright, R. H. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* **79**, 743-746.

Busby, S. & Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *J Mol Biol* **293**, 199-213.

Cakar, Z. P., Seker, U. O., Tamerler, C., Sonderegger, M. & Sauer, U. (2005). Evolutionary engineering of multiple-stress resistant Saccharomyces cerevisiae. *FEMS Yeast Res* **5**, 569-578.

Cameron, D. C., Altaras, N. E., Hoffman, M. L. & Shaw, A. J. (1998). Metabolic engineering of propanediol pathways. *Biotechnol Prog* **14**, 116-125.

Campbell, E. A., Muzzin, O., Chlenov, M., Sun, J. L., Olson, C. A., Weinman, O., Trester-Zedlitz, M. L. & Darst, S. A. (2002). Structure of the bacterial RNA polymerase promoter specificity sigma subunit. *Mol Cell* **9**, 527-539.

Cerretti, D. P., Dean, D., Davis, G. R., Bedwell, D. M. & Nomura, M. (1983). The spc ribosomal protein operon of Escherichia coli: sequence and cotranscription of the ribosomal protein genes and a protein export gene. *Nucleic Acids Res* **11**, 2599-2616.

Chand, P., Aruna, A., Maqsood, A. M. & Rao, L. V. (2005). Novel mutation method for increased cellulase production. *J Appl Microbiol* **98**, 318-323.

Choi, J. I., Lee, S. Y. & Han, K. (1998). Cloning of the Alcaligenes latus polyhydroxyalkanoate biosynthesis genes and use of these genes for enhanced production of Poly(3-hydroxybutyrate) in Escherichia coli. *Appl Environ Microbiol* **64**, 4897-4903.

Chong, B. F. & Nielsen, L. K. (2003). Amplifying the cellular reduction potential of Streptococcus zooepidemicus. *J Biotechnol* **100**, 33-41.

Choy, H. E., Park, S. W., Aki, T., Parrack, P., Fujita, N., Ishihama, A. & Adhya, S. (1995). Repression and activation of transcription by Gal and Lac repressors: involvement of alpha subunit of RNA polymerase. *Embo J* **14**, 4523-4529.

Choy, H. E., Hanger, R. R., Aki, T., Mahoney, M., Murakami, K., Ishihama, A. & Adhya, S. (1997). Repression and activation of promoter-bound RNA polymerase activity by Gal repressor. *J Mol Biol* **272**, 293-300.

Cirz, R. T., Jones, M. B., Gingles, N. A., Minogue, T. D., Jarrahi, B., Peterson, S. N. & Romesberg, F. E. (2007). Complete and SOS-mediated response of Staphylococcus aureus to the antibiotic ciprofloxacin. *J Bacteriol* **189**, 531-539.

Colland, F., Rain, J. C., Gounon, P., Labigne, A., Legrain, P. & De Reuse, H. (2001). Identification of the Helicobacter pylori anti-sigma28 factor. *Mol Microbiol* **41**, 477-487.

Cummings, M. P., Neel, M. C. & Shaw, K. L. (2008). A genealogical approach to quantifying lineage divergence. *Evolution* **62**, 2411-2422.

da Silva, E. A., de Melo, W. F., Antunes, D. F., dos Santos, S. K. B., Resende, A. D., Simoes, D. A. & de Morais, M. A. (2005). Isolation by genetic and physiological characteristics of a fuel-ethanol fermentative Saccharomyces cerevisiae strain with potential for genetic manipulation. *Journal of Industrial Microbiology & Biotechnology* **32**, 481-486.

Dalbow, D. G. & Bremer, H. (1975). Metabolic regulation of beta-galactosidase synthesis in Escherichia coli. A test for constitutive ribosome synthesis. *Biochem J* **150**, 1-8.

Dangi, B., Gronenborn, A. M., Rosner, J. L. & Martin, R. G. (2004). Versatility of the carboxy-terminal domain of the alpha subunit of RNA polymerase in transcriptional activation: use of the DNA contact site as a protein contact site for MarA. *Mol Microbiol* **54**, 45-59.

Demain, A. L. & Solomon, N. A. (1986). *Manual of industrial microbiology and biotechnology*. Washington, D.C.: American Society for Microbiology.

Demain, A. L. & Fang, A. (2000). The natural functions of secondary metabolites. *Adv Biochem Eng Biotechnol* **69**, 1-39.

Demain, A. L., Davies, J. E. & Atlas, R. M. (1999). *Manual of industrial microbiology and biotechnology*, 2nd edn. Washington, D.C.: ASM Press.

Demain, A. L., Newcomb, M. & Wu, J. H. (2005). Cellulase, clostridia, and ethanol. *Microbiol Mol Biol Rev* **69**, 124-154.

Dombroski, A. J., Walter, W. A., Record, M. T., Jr., Siegele, D. A. & Gross, C. A. (1992). Polypeptides containing highly conserved regions of transcription initiation factor sigma 70 exhibit specificity of binding to promoter DNA. *Cell* **70**, 501-512.

Dove, S. L., Darst, S. A. & Hochschild, A. (2003). Region 4 of sigma as a target for transcription regulation. *Mol Microbiol* **48**, 863-874.

Drummond, D. A., Iverson, B. L., Georgiou, G. & Arnold, F. H. (2005). Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J Mol Biol* **350**, 806-816.

Durot, M., Bourguignon, P. Y. & Schachter, V. (2009). Genome-scale models of bacterial metabolism: reconstruction and applications. *Fems Microbiology Reviews* **33**, 164-190.

Durre, P. (2007). Biobutanol: an attractive biofuel. *Biotechnol J* **2**, 1525-1534.

Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science*, 1967–1971.

Ebright, R. H. (2000). RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol* **304**, 687-698.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Egler, M., Grosse, C., Grass, G. & Nies, D. H. (2005). Role of the extracytoplasmic function protein family sigma factor RpoE in metal resistance of Escherichia coli. *J Bacteriol* **187**, 2297-2307.

el Hawrani, A. S., Sessions, R. B., Moreton, K. M. & Holbrook, J. J. (1996). Guided evolution of enzymes with new substrate specificities. *J Mol Biol* **264**, 97-110.

Ellinger, T., Behnke, D., Knaus, R., Bujard, H. & Gralla, J. D. (1994). Context-dependent effects of upstream A-tracts. Stimulation or inhibition of Escherichia coli promoter function. *J Mol Biol* **239**, 466-475.

Ellington, A. D. & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822.

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science* **297**, 1183-1186.

Errington, J. (1991). Possible intermediate steps in the evolution of a prokaryotic developmental system. *Proc Biol Sci* **244**, 117-121.

Estrem, S. T., Ross, W., Gaal, T., Chen, Z. W., Niu, W., Ebright, R. H. & Gourse, R. L. (1999). Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev* **13**, 2134-2147.

Ezeji, T., Qureshi, N. & Blaschek, H. P. (2007). Butanol production from agricultural residues: Impact of degradation products on Clostridium beijerinckii growth and butanol fermentation. *Biotechnol Bioeng* **97**, 1460-1469.

Falke, D., Fisher, M., Ye, D. & Juliano, R. L. (2003). Design of artificial transcription factors to selectively regulate the pro-apoptotic bax gene. *Nucleic Acids Res* **31**, e10.

Fang, F. C. (2005). Sigma cascades in prokaryotic regulatory networks. *Proc Natl Acad Sci U S A* **102**, 4933-4934.

Featherstone, M. (2002). Coactivators in transcription initiation: here are your orders. *Curr Opin Genet Dev* **12**, 149-155.

Ferenci, T. (2003). What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli? *Trends Microbiol* **11**, 457-461.

Fiocco, D., Capozzi, V., Goffin, P., Hols, P. & Spano, G. (2007). Improved adaptation to heat, cold, and solvent tolerance in Lactobacillus plantarum. *Appl Microbiol Biotechnol* **77**, 909-915.

Firth, A. E. & Patrick, W. M. (2005). Statistics of protein library construction. *Bioinformatics* **21**, 3314-3315.

Fischer, C. R. & Peterson, A. (2008). Conversion of natural products including cellulose to hydrocarbons, hydrogen and/or related compounds. US.

Fischer, C. R., Klein-Marcuschamer, D. & Stephanopoulos, G. (2008). Selection and optimization of microbial hosts for biofuels production. *Metab Eng* **10**, 295-304.

Follstad, B. D., Balcarcel, R. R., Stephanopoulos, G. & Wang, D. I. (1999). Metabolic flux analysis of hybridoma continuous culture steady state multiplicity. *Biotechnol Bioeng* **63**, 675-683.

Forster, A., Aurich, A., Mauersberger, S. & Barth, G. (2007a). Citric acid production from sucrose using a recombinant strain of the yeast Yarrowia lipolytica. *Appl Microbiol Biotechnol* **75**, 1409-1417.

Forster, A., Jacobs, K., Juretzek, T., Mauersberger, S. & Barth, G. (2007b). Overexpression of the ICL1 gene changes the product ratio of citric acid production by Yarrowia lipolytica. *Appl Microbiol Biotechnol* **77**, 861-869.

Franck, P., Petitipain, N., Cherlet, M., Dardennes, M., Maachi, F., Schutz, B., Poisson, L. & Nabet, P. (1996). Measurement of intracellular pH in cultured cells by flow cytometry with BCECF-AM. *J Biotechnol* **46**, 187-195.

Fritsch, P. S., Urbanowski, M. L. & Stauffer, G. V. (2000). Role of the RNA polymerase alpha subunits in MetR-dependent activation of metE and metH: important residues in the C-terminal domain and orientation requirements within RNA polymerase. *J Bacteriol* **182**, 5539-5550.

Gaal, T., Ross, W., Blatter, E. E., Tang, H., Jia, X., Krishnan, V. V., Assa-Munt, N., Ebright, R. H. & Gourse, R. L. (1996). DNA-binding determinants of the alpha subunit of RNA polymerase: novel DNA-binding domain architecture. *Genes Dev* **10**, 16-26.

Gall, S., Lynch, M. D., Sandoval, N. R. & Gill, R. T. (2008). Parallel mapping of genotypes to phenotypes contributing to overall biological fitness. *Metab Eng* **10**, 382-393.

Gantet, P., Hubac, C. & Brown, S. C. (1990). Flow Cytometric Fluorescence Anisotropy of Lipophilic Probes in Epidermal and Mesophyll Protoplasts from Water-Stressed Lupinus albus L. *Plant Physiol* **94**, 729-737.

Gardella, T., Moyle, H. & Susskind, M. M. (1989). A mutant Escherichia coli sigma 70 subunit of RNA polymerase with altered promoter specificity. *J Mol Biol* **206**, 579-590.

Gerngross, T. U. (2004). Advances in the production of human therapeutic proteins in yeast and filamentous fungi. *Nature Biotechnology*, 1409 - 1414

Giraud, E., Lelong, B. & Raimbault, M. (1991). Influence of Ph and Initial Lactate Concentration on the Growth of Lactobacillus-Plantarum. *Applied Microbiology and Biotechnology* **36**, 96-99.

Goa, K. L. & Benfield, P. (1994). Hyaluronic acid. A review of its pharmacology and use as a surgical aid in ophthalmology, and its therapeutic potential in joint disease and wound healing. *Drugs* **47**, 536-566.

Gorsich, S. W., Dien, B. S., Nichols, N. N., Slininger, P. J., Liu, Z. L. & Skory, C. D. (2006). Tolerance to furfural-induced stress is associated with pentose phosphate pathway genes ZWF1, GND1, RPE1, and TKL1 in Saccharomyces cerevisiae. *Appl Microbiol Biotechnol* **71**, 339-349.

Gourse, R. L., Ross, W. & Gaal, T. (2000). UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *Mol Microbiol* **37**, 687-695.

Graca da Silveira, M., Vitoria San Romao, M., Loureiro-Dias, M. C., Rombouts, F. M. & Abee, T. (2002). Flow cytometric assessment of membrane integrity of ethanol-stressed Oenococcus oeni cells. *Appl Environ Microbiol* **68**, 6087-6093.

Green, E. M., Boynton, Z. L., Harris, L. M., Rudolph, F. B., Papoutsakis, E. T. & Bennett, G. N. (1996). Genetic manipulation of acid formation pathways by gene inactivation in Clostridium acetobutylicum ATCC 824. *Microbiology* **142 (Pt 8)**, 2079-2086.

Gruber, T. M. & Bryant, D. A. (1997). Molecular systematic studies of eubacteria, using sigma70-type sigma factors of group 1 and group 2. *J Bacteriol* **179**, 1734-1747.

Gruber, T. M. & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* **57**, 441-466.

Gustafson, B. L., Wehner, P. S., Nelson, G. O. & Mercer, P. N. (1989). Low pressure catalytic hydrogenation of carbonyl-containing compounds. USA: Eastman Kodak Company.

Hamilton, S. R., Bobrowicz, P., Bobrowicz, B., Davidson, R. C., Li, H., Mitchell, T., Nett, J. H., Rausch, S., Stadheim, T. A., Wischnewski, H., Wildt, S. & Gerngross, T. U. (2003). Production of complex human glycoproteins in yeast. *Science* **301**, 1244-1246.

Hammer, K., Mijakovic, I. & Jensen, P. R. (2006). Synthetic promoter libraries: Tuning of gene expression. *Trends in Biotechnology* **24**, 53-55.

Hanai, T., Atsumi, S. & Liao, J. C. (2007). Engineered synthetic pathway for isopropanol production in Escherichia coli. *Appl Environ Microbiol* **73**, 7814-7818.

Hansen, M. E., Lund, F. & Carstensen, J. M. (2003). Visual clone identification of Penicillium commune isolates. *J Microbiol Methods* **52**, 221-229.

Harding, K. G., Dennis, J. S., von Blottnitz, H. & Harrison, S. T. (2007). Environmental analysis of plastic production processes: comparing petroleum-based polypropylene and polyethylene with biologically-based poly-beta-hydroxybutyric acid using life cycle analysis. *J Biotechnol* **130**, 57-66.

Hengge-Aronis, R. (2002). Signal transduction and regulatory mechanisms involved in control of the sigma(S) (RpoS) subunit of RNA polymerase. *Microbiol Mol Biol Rev* **66**, 373-395, table of contents.

Hofvendahl, K. & Hahn-Hagerdal, B. (2000). Factors affecting the fermentative lactic acid production from renewable resources. *Enzyme and Microbial Technology* **26**, 87-107.

Hoover, D. M. & Lubkowski, J. (2002). DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res* **30**, e43.

Hou, L. (2009). Improved Production of Ethanol by Novel Genome Shuffling in Saccharomyces cerevisiae. *Appl Biochem Biotechnol.*

Hui, A. & de Boer, H. A. (1987). Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc Natl Acad Sci U S A* **84**, 4762-4766.

Igarashi, K. & Ishihama, A. (1991). Bipartite functional map of the E. coli RNA polymerase alpha subunit: involvement of the C-terminal region in transcription activation by cAMP-CRP. *Cell* **65**, 1015-1022.

Igarashi, K., Fujita, N. & Ishihama, A. (1991). Identification of a subunit assembly domain in the alpha subunit of Escherichia coli RNA polymerase. *J Mol Biol* **218**, 1-6.

Imashimizu, M., Hanaoka, M., Seki, A., Murakami, K. S. & Tanaka, K. (2006). The cyanobacterial principal sigma factor region 1.1 is involved in DNA-binding in the free form and in transcription activity as holoenzyme. *FEBS Lett* **580**, 3439-3444.

Ingram, L. O. & Conway, T. (1988). Expression of Different Levels of Ethanologenic Enzymes from Zymomonas mobilis in Recombinant Strains of Escherichia coli. *Appl Environ Microbiol* **54**, 397-404.

Ingram, L. O., Conway, T., Clark, D. P., Sewell, G. W. & Preston, J. F. (1987). Genetic engineering of ethanol production in Escherichia coli. *Appl Environ Microbiol* **53**, 2420-2425.

Ingram, L. O., Gomez, P. F., Lai, X., Moniruzzaman, M., Wood, B. E., Yomano, L. P. & York, S. W. (1998). Metabolic engineering of bacteria for ethanol production. *Biotechnol Bioeng* **58**, 204-214.

Ingram, L. O., Aldrich, H. C., Borges, A. C., Causey, T. B., Martinez, A., Morales, F., Saleh, A., Underwood, S. A., Yomano, L. P., York, S. W., Zaldivar, J. & Zhou, S. (1999). Enteric bacterial catalysts for fuel ethanol production. *Biotechnol Prog* **15**, 855-866.

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M. & Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840-845.

Isett, K., George, H., Herber, W. & Amanullah, A. (2007). Twenty-four-well plate miniature bioreactor high-throughput system: assessment for microbial cultivations. *Biotechnol Bioeng* **98**, 1017-1028.

Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol* **54**, 499-518.

Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G. & Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* **21**, 635-637.

Jafri, S., Urbanowski, M. L. & Stauffer, G. V. (1996). The glutamic acid residue at amino acid 261 of the alpha subunit is a determinant of the intrinsic efficiency of RNA polymerase at the metE core promoter in Escherichia coli. *J Bacteriol* **178**, 6810-6816.

Jamieson, A. C., Miller, J. C. & Pabo, C. O. (2003). Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov* **2**, 361-368.

Jensen, P. R. & Hammer, K. (1998a). Artificial promoters for metabolic optimization. *Biotechnol Bioeng* **58**, 191-195.

Jensen, P. R. & Hammer, K. (1998b). The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Appl Environ Microbiol* **64**, 82-87.

Jeon, Y. H., Negishi, T., Shirakawa, M., Yamazaki, T., Fujita, N., Ishihama, A. & Kyogoku, Y. (1995). Solution structure of the activator contact domain of the RNA polymerase alpha subunit. *Science* **270**, 1495-1497.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651-654.

Jin, Y. S. & Stephanopoulos, G. (2007). Multi-dimensional gene target search for improving lycopene biosynthesis in Escherichia coli. *Metab Eng* **9**, 337-347.

Jishage, M. & Ishihama, A. (1998). A stationary phase protein in Escherichia coli with binding activity to the major sigma subunit of RNA polymerase. *Proc Natl Acad Sci U S A* **95**, 4953-4958.

Jishage, M., Kvint, K., Shingler, V. & Nystrom, T. (2002). Regulation of sigma factor competition by the alarmone ppGpp. *Genes Dev* **16**, 1260-1270.

Jones, K. L. & Keasling, J. D. (1998). Construction and characterization of F plasmid-based expression vectors. *Biotechnol Bioeng* **59**, 659-665.

Jones, K. L., Kim, S. W. & Keasling, J. D. (2000). Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. *Metab Eng* **2**, 328-338.

Jurgen, B., Hanschke, R., Sarvas, M., Hecker, M. & Schweder, T. (2001). Proteome and transcriptome based analysis of Bacillus subtilis cells overproducing an insoluble heterologous protein. *Appl Microbiol Biotechnol* **55**, 326-332.

Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**, 451-464.

Kakizaki, I., Takagaki, K., Endo, Y., Kudo, D., Ikeya, H., Miyoshi, T., Baggenstoss, B. A., Tlapak-Simmons, V. L., Kumari, K., Nakane, A., Weigel, P. H. & Endo, M. (2002). Inhibition of hyaluronan synthesis in Streptococcus equi FM100 by 4-methylumbelliferone. *Eur J Biochem* **269**, 5066-5075.

Kalia, V. C. & Purohit, H. J. (2008). Microbial diversity and genomics in aid of bioenergy. *J Ind Microbiol Biotechnol* **35**, 403-419.

Karagiannis, J. & Young, P. G. (2001). Intracellular pH homeostasis during cell-cycle progression and growth state transition in Schizosaccharomyces pombe. *J Cell Sci* **114**, 2929-2941.

Kayser, O. & Quax, W. J. (2007). *Medicinal plant biotechnology : from basic research to industrial applications.* Weinheim: Wiley-VCH.

Kedzierska, B., Szambowska, A., Herman-Antosiewicz, A., Lee, D. J., Busby, S. J., Wegrzyn, G. & Thomas, M. S. (2007). The C-terminal domain of the Escherichia coli RNA polymerase alpha subunit plays a role in the CI-dependent activation of the bacteriophage lambda pM promoter. *Nucleic Acids Res* **35**, 2311-2320.

Keenan, T. M., Nakas, J. P. & Tanenbaum, S. W. (2006). Polyhydroxyalkanoate copolymers from forest biomass. *J Ind Microbiol Biotechnol* **33**, 616-626.

Kim do, Y., Rha, E., Choi, S. L., Song, J. J., Hong, S. P., Sung, M. H. & Lee, S. G. (2007). Development of bioreactor system for L-tyrosine synthesis using thermostable tyrosine phenol-lyase. *J Microbiol Biotechnol* **17**, 116-122.

Kim, J. H., Yoo, S. J., Oh, D. K., Kweon, Y. G., Park, D. W., Lee, C. H. & Gil, G. H. (1996). Selection of a Streptococcus equi mutant and optimization of culture conditions for the production of high molecular weight hyaluronic acid. *Enzyme and Microbial Technology* **19**, 440-445.

Kimura, M. & Ishihama, A. (1995). Functional map of the alpha subunit of Escherichia coli RNA polymerase: insertion analysis of the amino-terminal assembly domain. *J Mol Biol* **248**, 756-767.

Kiss, R. D. & Stephanopoulos, G. (1991). Metabolic-Activity Control of the L-Lysine Fermentation by Restrained Growth Fed-Batch Strategies. *Biotechnology Progress* **7**, 501-509.

Kitano, H. (2004). Biological robustness. *Nat Rev Genet* **5**, 826-837.

Kittell, J., Borup, B., Voladari, R. & Zahn, K. (2005). Parallel capillary electrophoresis for the quantitative screening of fermentation broths containing natural products. *Metab Eng* **7**, 53-58.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O. P., Leer, R., Tarchini, R., Peters, S. A., Sandbrink, H. M., Fiers, M. W., Stiekema, W., Lankhorst, R. M., Bron, P. A., Hoffer, S. M., Groot, M. N., Kerkhoven, R., de Vries, M., Ursing, B., de Vos, W. M. & Siezen, R. J. (2003). Complete genome sequence of Lactobacillus plantarum WCFS1. *Proc Natl Acad Sci U S A* **100**, 1990-1995.

Klein-Marcuschamer, D., Ajikumar, P. K. & Stephanopoulos, G. (2007). Engineering microbial cell factories for biosynthesis of isoprenoid molecules: beyond lycopene. *Trends Biotechnol* **25**, 417-424.

Klinke, H. B., Thomsen, A. B. & Ahring, B. K. (2004). Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Appl Microbiol Biotechnol* **66**, 10-26.

Kok, J., van der Vossen, J. M. & Venema, G. (1984). Construction of plasmid cloning vectors for lactic streptococci which also replicate in Bacillus subtilis and Escherichia coli. *Appl Environ Microbiol* **48**, 726-731.

Kresnowati, M. T., Suarez-Mendez, C., Groothuizen, M. K., van Winden, W. A. & Heijnen, J. J. (2007). Measurement of fast dynamic intracellular pH in Saccharomyces cerevisiae using benzoic acid pulse. *Biotechnol Bioeng* **97**, 86-98.

Kumazawa, J. & Yagisawa, M. (2002). The history of antibiotics: the Japanese story. *J Infect Chemother* **8**, 125-133.

Lacoursiere, A., Thompson, B. G., Kole, M. M., Ward, D. & Gerson, D. F. (1986). Effects of Carbon-Dioxide Concentration on Anaerobic Fermentations of Escherichia-Coli. *Applied Microbiology and Biotechnology* **23**, 404-406.

Larsson, S., Nilvebrant, N. O. & Jonsson, L. J. (2001). Effect of overexpression of Saccharomyces cerevisiae Pad1p on the resistance to phenylacrylic acids and lignocellulose hydrolysates under aerobic and oxygen-limited conditions. *Appl Microbiol Biotechnol* **57**, 167-174.

Lauren, T. (1998). *The chemistry, biology and medical applications of hyaluronan and its derivatives.* . London: Portland Press Ltd.

Lawford, H. G. & Rousseau, J. D. (1998). Improving fermentation performance of recombinant Zymomonas in acetic acid-containing media. *Appl Biochem Biotechnol* **70-72**, 161-172.

Lee, D. K., Kim, Y. H., Kim, J. S. & Seol, W. (2004). Induction and characterization of taxol-resistance phenotypes with a transiently expressed artificial transcriptional activator library. *Nucleic Acids Res* **32**, e116.

Lee, J. Y., Sung, B. H., Yu, B. J., Lee, J. H., Lee, S. H., Kim, M. S., Koob, M. D. & Kim, S. C. (2008). Phenotypic engineering by reprogramming gene transcription using novel artificial transcription factors in Escherichia coli. *Nucleic Acids Res* **36**, e102.

Lee, S. J. & Gralla, J. D. (2004). Osmo-regulation of bacterial transcription via poised RNA polymerase. *Mol Cell* **14**, 153-162.

Lee, S. Y. (1998). Poly(3-hydroxybutyrate) production from xylose by recombinant Escherichia coli. *Bioprocess Engineering* **18**, 397-399.

Lerner, C. G. & Inouye, M. (1990). Low copy number plasmids for regulated low-level expression of cloned genes in Escherichia coli with blue/white insert screening capability. *Nucleic Acids Res* **18**, 4631.

Li, M., Moyle, H. & Susskind, M. M. (1994). Target of the transcriptional activation function of phage lambda cI protein. *Science* **263**, 75-77.

Li, S. C., Goto, N. K., Williams, K. A. & Deber, C. M. (1996). Alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment. *Proc Natl Acad Sci U S A* **93**, 6676-6681.

Lin, C. F. & Chung, T. C. (1999). Cloning of erythromycin-resistance determinants and replication origins from indigenous plasmids of Lactobacillus reuteri for potential use in construction of cloning vectors. *Plasmid* **42**, 31-41.

Lin, Y. & Tanaka, S. (2006). Ethanol fermentation from biomass resources: current state and prospects. *Appl Microbiol Biotechnol* **69**, 627-642.

Liu, P. Q., Rebar, E. J., Zhang, L., Liu, Q., Jamieson, A. C., Liang, Y., Qi, H., Li, P. X., Chen, B., Mendel, M. C., Zhong, X., Lee, Y. L., Eisenberg, S. P., Spratt, S. K., Case, C. C. & Wolffe, A. P. (2001). Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J Biol Chem* **276**, 11323-11334.

Liu, Z. H., Yang, Q. & Ma, J. (2007). A heat shock protein gene (hsp22.4) from Chaetomium globosum confers heat and Na(2)CO (3) tolerance to yeast. *Appl Microbiol Biotechnol* **77**, 901-908.

Liu, Z. L., Slininger, P. J. & Gorsich, S. W. (2005). Enhanced biotransformation of furfural and hydroxymethylfurfural by newly developed ethanologenic yeast strains. *Appl Biochem Biotechnol* **121-124**, 451-460.

Lochowska, A., Iwanicka-Nowicka, R., Zaim, J., Witkowska-Zimny, M., Bolewska, K. & Hryniewicz, M. M. (2004). Identification of activating region (AR) of Escherichia coli LysR-type transcription factor CysB and CysB contact site on RNA polymerase alpha subunit at the cysP promoter. *Mol Microbiol* **53**, 791-806.

Lunzer, M., Miller, S. P., Felsheim, R. & Dean, A. M. (2005). The biochemical architecture of an ancient adaptive landscape. *Science* **310**, 499-501.

Luscombe, N. M. & Thornton, J. M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* **320**, 991-1009.

Lutke-Eversloh, T. & Stephanopoulos, G. (2005). Feedback inhibition of chorismate mutase/prephenate dehydrogenase (TyrA) of Escherichia coli: generation and characterization of tyrosine-insensitive mutants. *Appl Environ Microbiol* **71**, 7224-7228.

Lutke-Eversloh, T. & Stephanopoulos, G. (2007). L-tyrosine production by deregulated strains of Escherichia coli. *Appl Microbiol Biotechnol* **75**, 103-110.

Lutke-Eversloh, T., Santos, C. N. & Stephanopoulos, G. (2007). Perspectives of biotechnological production of L-tyrosine and its applications. *Appl Microbiol Biotechnol* **77**, 751-762.

Lutz, R. & Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res* **25**, 1203-1210.

Lynch, M. D., Warnecke, T. & Gill, R. T. (2007). SCALEs: multiscale analysis of library enrichment. *Nat Methods* **4**, 87-93.

Lynd, L. R., Wyman, C. E. & Gerngross, T. U. (1999). Biocommodity Engineering. *Biotechnol Prog* **15**, 777-793.

Lynd, L. R., Weimer, P. J., van Zyl, W. H. & Pretorius, I. S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev* **66**, 506-577, table of contents.

Lynd, L. R., van Zyl, W. H., McBride, J. E. & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: an update. *Curr Opin Biotechnol* **16**, 577-583.

Ma, H. W., Buer, J. & Zeng, A. P. (2004). Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**, 199.

Magnusson, L. U., Farewell, A. & Nystrom, T. (2005). ppGpp: a global regulator in Escherichia coli. *Trends Microbiol* **13**, 236-242.

Mahishi, L. H., Tripathi, G. & Rawal, S. K. (2003). Poly(3-hydroxybutyrate) (PHB) synthesis by recombinant Escherichia coli harbouring Streptomyces aureofaciens PHB biosynthesis genes: effect of various carbon and nitrogen sources. *Microbiol Res* **158**, 19-27.

Malumbres, M. & Martin, J. F. (1996). Molecular control mechanisms of lysine and threonine biosynthesis in amino acid-producing corynebacteria: redirecting carbon flow. *FEMS Microbiol Lett* **143**, 103-114.

Marten, M. R., Park, T. H. & Nagamune, T. (2002). *Biological systems engineering.* Washington, DC: American Chemical Society : Distributed by Oxford University Press.

Martinez-Antonio, A., Janga, S. C. & Thieffry, D. (2008). Functional organisation of Escherichia coli transcriptional regulatory network. *J Mol Biol* **381**, 238-247.

Matsushima, P., McHenney, M. A. & Baltz, R. H. (1989). Transduction and transformation of plasmid DNA in Streptomyces fradiae strains that express different levels of restriction. *J Bacteriol* **171**, 3080-3084.

McAloon, A., Taylor, F., Yee, W., Ibsen, K. & Wooley, R. (2000). Determining the Cost of Producing Ethanol from Corn Starch and Lignocellulosic Feedstocks. Edited by N. A. R. Service. Golden, CO: NREL.

McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annu Rev Biochem* **54**, 171-204.

McDaniel, R., Licari, P. & Khosla, C. (2001). Process development and metabolic engineering for the overproduction of natural and unnatural polyketides. *Adv Biochem Eng Biotechnol* **73**, 31-52.

McDonald, L. C., Fleming, H. P. & Hassan, H. M. (1990). Acid Tolerance of Leuconostoc mesenteroides and Lactobacillus plantarum. *Appl Environ Microbiol* **56**, 2120-2124.

McLeod, S. M., Aiyar, S. E., Gourse, R. L. & Johnson, R. C. (2002). The C-terminal domains of the RNA polymerase alpha subunits: contact site with Fis and localization during co-activation with CRP at the Escherichia coli proP P2 promoter. *J Mol Biol* **316**, 517-529.

Miesenbock, G., De Angelis, D. A. & Rothman, J. E. (1998). Visualizing secretion and synaptic transmission with pH-sensitive green fluorescent proteins. *Nature* **394**, 192-195.

Miller, J. H. (1972). Experiments in molecular genetics. pp. 125-129. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Missiakas, D. & Raina, S. (1997). Protein folding in the bacterial periplasm. *J Bacteriol* **179**, 2465-2471.

Miyagishi, M., Matsumoto, S., Akashi, H., Kawasaki, H., Fukao, T., Fukuda, Y., Sano, M., Kato, Y., Takagi, Y., Tanaka, Y., Warashina, M., Kuwabara, T., Sawata, S. Y., Ikeda, Y., Kawahara, S., Sunil, K. C., Wadhwa, R. & Taira, K. (2005). Chemistry-based RNA technologies: demonstration of usefulness of libraries of ribozymes and short hairpin RNAs (shRNAs). *Nucleic Acids Symp Ser (Oxf)*, 91-92.

Miyazaki, K. & Arnold, F. H. (1999). Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J Mol Evol* **49**, 716-720.

Molina, N. & van Nimwegen, E. (2008). Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res* **18**, 148-160.

Mooney, R. A., Darst, S. A. & Landick, R. (2005). Sigma and RNA polymerase: an on-again, off-again relationship? *Mol Cell* **20**, 335-345.

Morel, F., Delmas, F., Jobin, M. P., Divies, C. & Guzzo, J. (2001). Improved acid tolerance of a recombinant strain of Escherichia coli expressing genes from the acidophilic bacterium Oenococcus oeni. *Letters in Applied Microbiology* **33**, 126-130.

Moreno-Sanchez, R., Bravo, C. & Westerhoff, H. V. (1999). Determining and understanding the control of flux. An illustration in submitochondrial particles of how to validate schemes of metabolic control. *Eur J Biochem* **264**, 427-433.

Muller, S., Ullrich, S., Losche, A., Loffhagen, N. & Babel, W. (2000). Flow cytometric techniques to characterise physiological states of Acinetobacter calcoaceticus. *J Microbiol Methods* **40**, 67-77.

Munir, K. M., French, D. C. & Loeb, L. A. (1993). Thymidine kinase mutants obtained by random sequence selection. *Proc Natl Acad Sci U S A* **90**, 4012-4016.

Murakami, K., Fujita, N. & Ishihama, A. (1996). Transcription factor recognition surface on the RNA polymerase alpha subunit is involved in contact with the DNA enhancer element. *Embo J* **15**, 4358-4367.

Murphy, M. G., O'Connor, L., Walsh, D. & Condon, S. (1985). Oxygen dependent lactate utilization by Lactobacillus plantarum. *Arch Microbiol* **141**, 75-79.

Mytelka, D. S. & Chamberlin, M. J. (1996). Escherichia coli fliAZY operon. *J Bacteriol* **178**, 24-34.

Ng, T. K., Ben-Bassat, A. & Zeikus, J. G. (1981). Ethanol Production by Thermophilic Bacteria: Fermentation of Cellulosic Substrates by Cocultures of Clostridium thermocellum and Clostridium thermohydrosulfuricum. *Appl Environ Microbiol* **41**, 1337-1343.

Nickels, B. E., Dove, S. L., Murakami, K. S., Darst, S. A. & Hochschild, A. (2002). Protein-protein and protein-DNA interactions of sigma70 region 4 involved in transcription activation by lambdacI. *J Mol Biol* **324**, 17-34.

Nigam, J. N. (2001a). Development of xylose-fermenting yeast Pichia stipitis for ethanol production through adaptation on hardwood hemicellulose acid prehydrolysate. *J Appl Microbiol* **90**, 208-215.

Nigam, J. N. (2001b). Ethanol production from hardwood spent sulfite liquor using an adapted strain of Pichia stipitis. *J Ind Microbiol Biotechnol* **26**, 145-150.

Nikel, P. I., de Almeida, A., Melillo, E. C., Galvagno, M. A. & Pettinari, M. J. (2006). New recombinant Escherichia coli strain tailored for the production of poly(3-hydroxybutyrate) from agroindustrial by-products. *Appl Environ Microbiol* **72**, 3949-3954.

Niu, W., Kim, Y., Tau, G., Heyduk, T. & Ebright, R. H. (1996). Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. *Cell* **87**, 1123-1134.

Nystrom, T. (2004). Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition? *Mol Microbiol* **54**, 855-862.

Ogrodowski, C. S., Hokka, C. O. & Santana, M. H. A. (2005). Production of hyaluronic acid by Streptococcus. *Applied Biochemistry and Biotechnology* **121**, 753-761.

Ohnuma, S., Narita, K., Nakazawa, T., Ishida, C., Takeuchi, Y., Ohto, C. & Nishino, T. (1996). A role of the amino acid residue located on the fifth position before the first aspartate-rich motif of farnesyl diphosphate synthase on determination of the final product. *J Biol Chem* **271**, 30748-30754.

Ohta, K., Beall, D. S., Mejia, J. P., Shanmugam, K. T. & Ingram, L. O. (1991). Genetic improvement of Escherichia coli for ethanol production: chromosomal integration of Zymomonas mobilis genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. *Appl Environ Microbiol* **57**, 893-900.

Onken, U. & Liefke, E. (1989). Effect of total and partial pressure (oxygen and carbon dioxide) on aerobic microbial processes. *Adv Biochem Eng Biotechnol* **40**, 137-169.

Paget, M. S. & Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biol* **4**, 203.

Panikov, N. S., Belova, S. E. & Dorofeev, A. G. (2002). [Nonlinearity in the growth of bacterial colonies: conditions and causes]. *Mikrobiologiia* **71**, 59-65.

Parekh, S., Vinci, V. A. & Strobel, R. J. (2000). Improvement of microbial strains and fermentation processes. *Appl Microbiol Biotechnol* **54**, 287-301.

Park, J. H., Lee, S. Y., Kim, T. Y. & Kim, H. U. (2008). Application of systems biology for bioprocess development. *Trends Biotechnol* **26**, 404-412.

Park, K. S., Jang, Y. S., Lee, H. & Kim, J. S. (2005a). Phenotypic alteration and target gene identification using combinatorial libraries of zinc finger proteins in prokaryotic cells. *J Bacteriol* **187**, 5496-5499.

Park, K. S., Lee, D. K., Lee, H., Lee, Y., Jang, Y. S., Kim, Y. H., Yang, H. Y., Lee, S. I., Seol, W. & Kim, J. S. (2003). Phenotypic alteration of eukaryotic cells using randomized libraries of artificial transcription factors. *Nat Biotechnol* **21**, 1208-1214.

Park, K. S., Seol, W., Yang, H. Y., Lee, S. I., Kim, S. K., Kwon, R. J., Kim, E. J., Roh, Y. H., Seong, B. L. & Kim, J. S. (2005b). Identification and use of zinc finger transcription factors that increase production of recombinant proteins in yeast and mammalian cells. *Biotechnol Prog* **21**, 664-670.

Patnaik, R., Louie, S., Gavrilovic, V., Perry, K., Stemmer, W. P., Ryan, C. M. & del Cardayre, S. (2002). Genome shuffling of Lactobacillus for improved acid tolerance. *Nat Biotechnol* **20**, 707-712.

Paul, B. J., Ross, W., Gaal, T. & Gourse, R. L. (2004). rRNA transcription in Escherichia coli. *Annu Rev Genet* **38**, 749-770.

Penney, D. P., Powers, J. M., Frank, M., Willis, C. & Churukian, C. (2002). Analysis and testing of biological stains - The Biological Stain Commission procedures. *Biotechnic & Histochemistry* **77**, 237-275.

Perlach, R. D., Wright, L. L., Turhollow, A. F., Graham, R. L., Stokes, B. J. & Erbach, D. C. (2005). Biomass as a Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply. Edited by Energy & Agriculture.

Petersson, A., Almeida, J. R., Modig, T., Karhumaa, K., Hahn-Hagerdal, B., Gorwa-Grauslund, M. F. & Liden, G. (2006). A 5-hydroxymethyl furfural reducing enzyme encoded by the Saccharomyces cerevisiae ADH6 gene conveys HMF tolerance. *Yeast* **23**, 455-464.

Pfleger, B. F., Pitera, D. J., Newman, J. D., Martin, V. J. & Keasling, J. D. (2007). Microbial sensors for small molecules: development of a mevalonate biosensor. *Metab Eng* **9**, 30-38.

Pieterse, B., Leer, R. J., Schuren, F. H. & van der Werf, M. J. (2005). Unravelling the multiple effects of lactic acid stress on Lactobacillus plantarum by transcription profiling. *Microbiology* **151**, 3881-3894.

Posno, M., Leer, R. J., van Luijk, N., van Giezen, M. J., Heuvelmans, P. T., Lokman, B. C. & Pouwels, P. H. (1991). Incompatibility of Lactobacillus Vectors with

Replicons Derived from Small Cryptic Lactobacillus Plasmids and Segregational Instability of the Introduced Vectors. *Appl Environ Microbiol* **57**, 1822-1828.

Post, L. E., Arfsten, A. E., Reusser, F. & Nomura, M. (1978). DNA sequences of promoter regions for the str and spc ribosomal protein operons in E. coli. *Cell* **15**, 215-229.

Raab, R. M., Tyo, K. & Stephanopoulos, G. (2005). Metabolic engineering. *Advances in Biochemical Engineering/Biotechnology*, 1-17.

Rallu, F., Gruss, A., Ehrlich, S. D. & Maguin, E. (2000). Acid- and multistress-resistant mutants of Lactococcus lactis : identification of intracellular stress signals. *Mol Microbiol* **35**, 517-528.

Reetz, M. T. & Carballeira, J. D. (2007). Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc* **2**, 891-903.

Reitzer, L. (2003). Nitrogen assimilation and global regulation in Escherichia coli. *Annu Rev Microbiol* **57**, 155-176.

Ren, D., Collingwood, T. N., Rebar, E. J., Wolffe, A. P. & Camp, H. S. (2002). PPARgamma knockdown by engineered transcription factors: exogenous PPARgamma2 but not PPARgamma1 reactivates adipogenesis. *Genes Dev* **16**, 27-32.

Roberts, J. W., Shankar, S. & Filter, J. J. (2008). RNA polymerase elongation factors. *Annu Rev Microbiol* **62**, 211-233.

Ross, W. & Gourse, R. L. (2005). Sequence-independent upstream DNA-alphaCTD interactions strongly stimulate Escherichia coli RNA polymerase-lacUV5 promoter association. *Proc Natl Acad Sci U S A* **102**, 291-296.

Ross, W., Aiyar, S. E., Salomon, J. & Gourse, R. L. (1998). Escherichia coli promoters with UP elements of different strengths: modular structure of bacterial promoters. *J Bacteriol* **180**, 5375-5383.

Ross, W., Schneider, D. A., Paul, B. J., Mertens, A. & Gourse, R. L. (2003). An intersubunit contact stimulating transcription initiation by E coli RNA polymerase: interaction of the alpha C-terminal domain and sigma region 4. *Genes Dev* **17**, 1293-1307.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K. & Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262**, 1407-1413.

Ruiz, N. & Silhavy, T. J. (2005). Sensing external stress: watchdogs of the Escherichia coli cell envelope. *Curr Opin Microbiol* **8**, 122-126.

Sahm, H., Eggeling, L., Eikmanns, B. & Kramer, R. (1996). Construction of L-lysine-, L-threonine-, and L-isoleucine-overproducing strains of Corynebacterium glutamicum. *Ann N Y Acad Sci* **782**, 25-39.

Sakai, S., Tsuchida, Y., Okino, S., Ichihashi, O., Kawaguchi, H., Watanabe, T., Inui, M. & Yukawa, H. (2007). Effect of lignocellulose-derived inhibitors on growth of and ethanol production by growth-arrested Corynebacterium glutamicum R. *Appl Environ Microbiol* **73**, 2349-2353.

Sambrook, J. & Russell, D. W. (2001a). *Molecular cloning : a laboratory manual*, 3rd edn. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Sambrook, J. & Russell, D. W. (2001b). *Molecular cloning : a laboratory manual*, 3rd edn. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

Sambrook, J., Russell, D. W. & Sambrook, J. (2006). *The condensed protocols from Molecular cloning : a laboratory manual.* Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

San, K. Y. & Stephanopoulos, G. (1984). Studies on Online Bioreactor Identification .4. Utilization of Ph Measurements for Product Estimation. *Biotechnology and Bioengineering* **26**, 1209-1218.

Santos, C. N. & Stephanopoulos, G. (2008a). Combinatorial engineering of microbes for optimizing cellular phenotype. *Curr Opin Chem Biol* **12**, 168-176.

Santos, C. N. & Stephanopoulos, G. (2008b). Melanin-based high-throughput screen for L-tyrosine production in Escherichia coli. *Appl Environ Microbiol* **74**, 1190-1197.

Sauer, M., Porro, D., Mattanovich, D. & Branduardi, P. (2008). Microbial production of organic acids: expanding the markets. *Trends in Biotechnology* **26**, 100-108.

Sauer, U. (2001). Evolutionary engineering of industrially important microbial phenotypes. *Adv Biochem Eng Biotechnol* **73**, 129-169.

Savery, N. J., Lloyd, G. S., Busby, S. J., Thomas, M. S., Ebright, R. H. & Gourse, R. L. (2002). Determinants of the C-terminal domain of the Escherichia coli RNA polymerase alpha subunit important for transcription at class I cyclic AMP receptor protein-dependent promoters. *J Bacteriol* **184**, 2273-2280.

Schauer, A. T., Cheng, S. W., Zheng, C., St Pierre, L., Alessi, D., Hidayetoglu, D. L., Costantino, N., Court, D. L. & Friedman, D. I. (1996). The alpha subunit of RNA polymerase and transcription antitermination. *Mol Microbiol* **21**, 839-851.

Schlax, P. J. & Worhunsky, D. J. (2003). Translational repression mechanisms in prokaryotes. *Mol Microbiol* **48**, 1157-1169.

Schneider-Berlin, K. R., Bonilla, T. D. & Rowe, T. C. (2005). Induction of petite mutants in yeast Saccharomyces cerevisiae by the anticancer drug dequalinium. *Mutat Res* **572**, 84-97.

Segal, D. J., Beerli, R. R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., Lund, C. V., Magnenat, L., Valente, D. & Barbas, C. F., 3rd (2003). Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* **42**, 2137-2148.

Shams Yazdani, S. & Gonzalez, R. (2008). Engineering Escherichia coli for the efficient conversion of glycerol to ethanol and co-products. *Metab Eng* **10**, 340-351.

Shang, L., Jiang, M., Ryu, C. H., Chang, H. N., Cho, S. H. & Lee, J. W. (2003). Inhibitory effect of carbon dioxide on the fed-batch culture of Ralstonia eutropha: evaluation by CO2 pulse injection and autogenous CO2 methods. *Biotechnol Bioeng* **83**, 312-320.

Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. (2004). Advanced sequencing technologies: Methods and goals. *Nature Reviews Genetics*, 335-344.

Shi, D. J., Wang, C. L. & Wang, K. M. (2009). Genome shuffling to improve thermotolerance, ethanol tolerance and ethanol productivity of Saccharomyces cerevisiae. *J Ind Microbiol Biotechnol* **36**, 139-147.

Shigapova, N., Torok, Z., Balogh, G., Goloubinoff, P., Vigh, L. & Horvath, I. (2005). Membrane fluidization triggers membrane remodeling which affects the thermotolerance in Escherichia coli. *Biochem Biophys Res Commun* **328**, 1216-1223.

Shuler, M. L. & Kargi, F. (2002). *Bioprocess engineering*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.

Siegele, D. A., Hu, J. C. & Gross, C. A. (1988). Mutations in rpoD, the gene encoding the sigma 70 subunit of Escherichia coli RNA polymerase, that increase expression of the lac operon in the absence of CAP-cAMP. *J Mol Biol* **203**, 29-37.

Siegele, D. A., Hu, J. C., Walter, W. A. & Gross, C. A. (1989). Altered promoter recognition by mutant forms of the sigma 70 subunit of Escherichia coli RNA polymerase. *J Mol Biol* **206**, 591-603.

Sikkema, J., de Bont, J. A. & Poolman, B. (1995). Mechanisms of membrane toxicity of hydrocarbons. *Microbiol Rev* **59**, 201-222.

Sillers, R., Chow, A., Tracy, B. & Papoutsakis, E. T. (2008). Metabolic engineering of the non-sporulating, non-solventogenic Clostridium acetobutylicum strain M5 to produce butanol without acetone demonstrate the robustness of the acid-formation pathways and the importance of the electron balance. *Metab Eng* **10**, 321-332.

Singh, O. V. (2006). Mutagenesis and analysis of mold Aspergillus niger for extracellular glucose oxidase production using sugarcane molasses. *Appl Biochem Biotechnol* **135**, 43-57.

Smolke, C. D., Martin, V. J. & Keasling, J. D. (2001). Controlling the metabolic flux through the carotenoid pathway using directed mRNA processing and stabilization. *Metab Eng* **3**, 313-321.

Sonderegger, M., Jeppsson, M., Larsson, C., Gorwa-Grauslund, M. F., Boles, E., Olsson, L., Spencer-Martins, I., Hahn-Hagerdal, B. & Sauer, U. (2004). Fermentation performance of engineered and evolved xylose-fermenting Saccharomyces cerevisiae strains. *Biotechnol Bioeng* **87**, 90-98.

Spilimbergo, S., Bertucco, A., Basso, G. & Bertoloni, G. (2005). Determination of extracellular and intracellular pH of Bacillus subtilis suspension under CO2 treatment. *Biotechnol Bioeng* **92**, 447-451.

Sprinzak, D. & Elowitz, M. B. (2005). Reconstruction of genetic circuits. *Nature*, 443-448.

Srivastava, R., Cha, H. J., Peterson, M. S. & Bentley, W. E. (2000). Antisense downregulation of sigma(32) as a transient metabolic controller in Escherichia coli: effects on yield of active organophosphorus hydrolase. *Appl Environ Microbiol* **66**, 4366-4371.

Stemmer, W. P. (1994). Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370**, 389-391.

Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. *Metabolic Engineering*, 1-11.

Stephanopoulos, G. & Sinskey, A. J. (1993). Metabolic Engineering - Methodologies and Future-Prospects. *Trends in Biotechnology* **11**, 392-396.

Stephanopoulos, G. & Simpson, T. W. (1997). Flux amplification in complex metabolic networks. *Chemical Engineering Science* **52**, 2607-2627.

Stephanopoulos, G. & Kelleher, J. (2001). How to make a superior cell (vol 292, pg 2024, 2001). *Science* **293**, 1766-1766.

Stephanopoulos, G., Aristidou, A. A. & Nielsen, J. (1998). *Metabolic engineering : principles and methodologies.* San Diego: Academic Press.

Swain, P. S., Elowitz, M. B. & Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A* **99**, 12795-12800.

Taherzadeh, M. J., Niklasson, C. & Liden, G. (1997a). Acetic acid - friend or foe in anaerobic batch conversion of glucose to ethanol by Saccharomyces cerevisiae? *Chemical Engineering Science* **52**, 2653-2659.

Taherzadeh, M. J., Eklund, R., Gustafsson, L., Niklasson, C. & Liden, G. (1997b). Characterization and fermentation of dilute-acid hydrolyzates from wood. *Industrial & Engineering Chemistry Research* **36**, 4659-4665.

Tao, L., Jackson, R. E. & Cheng, Q. (2005). Directed evolution of copy number of a broad host range plasmid for metabolic engineering. *Metab Eng* **7**, 10-17.

Tashiro, Y., Takeda, K., Kobayashi, G., Sonomoto, K., Ishizaki, A. & Yoshino, S. (2004). High butanol production by Clostridium saccharoperbutylacetonicum N1-4 in fed-batch culture with pH-Stat continuous butyric acid and glucose feeding method. *J Biosci Bioeng* **98**, 263-268.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X. & Church, G. M. (2004). Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, 1050-1054.

Turner, P. R. & Denny, W. A. (1996). The mutagenic properties of DNA minor-groove binding ligands. *Mutat Res* **355**, 141-169.

Tyo, K. E., Zhou, H. & Stephanopoulos, G. N. (2006). High-throughput screen for poly-3-hydroxybutyrate in Escherichia coli and Synechocystis sp. strain PCC6803. *Appl Environ Microbiol* **72**, 3412-3417.

Typas, A. & Hengge, R. (2006). Role of the spacer between the -35 and -10 regions in sigmas promoter selectivity in Escherichia coli. *Mol Microbiol* **59**, 1037-1051.

Umemoto, A., Morita, M., Nakazono, N. & Sugino, Y. (1996). The effect of the crp genotypes on the transformation efficiency in Escherichia coli. *DNA Res* **3**, 93-94.

Valli, M., Sauer, M., Branduardi, P., Borth, N., Porro, D. & Mattanovich, D. (2006). Improvement of lactic acid production in Saccharomyces cerevisiae by cell sorting for high intracellular pH. *Appl Environ Microbiol* **72**, 5492-5499.

Vallino, J. J. & Stephanopoulos, G. (1994). Carbon Flux Distributions at the Pyruvate Branch Point in Corynebacterium-Glutamicum during Lysine Overproduction. *Biotechnology Progress* **10**, 320-326.

Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D., Auckenthaler, A. & Schwer, H. (2008). A novel solid phase technology for high-throughput gene synthesis. *Biotechniques* **45**, 340-343.

van Maris, A. J. A., Abbott, D. A., Bellissimi, E., van den Brink, J., Kuyper, M., Luttik, M. A. H., Wisselink, H. W., Scheffers, W. A., van Dijken, J. P. & Pronk, J. T. (2006). Alcoholic fermentation of carbon sources in biomass hydrolysates by Saccharomyces cerevisiae: current status. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **90**, 391-418.

Vijayakumar, S. R., Kirchhof, M. G., Patten, C. L. & Schellhorn, H. E. (2004). RpoS-regulated genes of Escherichia coli identified by random lacZ fusion mutagenesis. *J Bacteriol* **186**, 8499-8507.

Villaverde, A. & Carrio, M. M. (2003). Protein aggregation in recombinant bacteria: biological role of inclusion bodies. *Biotechnol Lett* **25**, 1385-1395.

Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001). Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A* **98**, 3778-3783.

Wackett, L. P. (2008). Biomass to fuels via microbial transformations. *Curr Opin Chem Biol* **12**, 187-193.

Waldburger, C., Gardella, T., Wong, R. & Susskind, M. M. (1990). Changes in conserved region 2 of Escherichia coli sigma 70 affecting promoter recognition. *J Mol Biol* **215**, 267-276.

Wang, T. W., Zhu, H., Ma, X. Y., Zhang, T., Ma, Y. S. & Wei, D. Z. (2006). Mutant library construction in directed molecular evolution: casting a wider net. *Mol Biotechnol* **34**, 55-68.

Wang, Y., Wang, Y. E., Cotticelli, M. G. & Wilson, R. B. (2008). A random shRNA-encoding library for phenotypic selection and hit-optimization. *PLoS ONE* **3**, e3171.

Wang, Y., Li, Y., Pei, X., Yu, L. & Feng, Y. (2007). Genome-shuffling improved acid tolerance and L-lactic acid volumetric productivity in Lactobacillus rhamnosus. *J Biotechnol* **129**, 510-515.

Warnecke, T. E., Lynch, M. D., Karimpour-Fard, A., Sandoval, N. & Gill, R. T. (2008). A genomics approach to improve the analysis and design of strain selections. *Metab Eng* **10**, 154-165.

Weber, H., Polen, T., Heuveling, J., Wendisch, V. F. & Hengge, R. (2005). Genome-wide analysis of the general stress response network in Escherichia coli: sigmaS-dependent genes, promoters, and sigma factor selectivity. *J Bacteriol* **187**, 1591-1603.

Wei, P., Li, Z., He, P., Lin, Y. & Jiang, N. (2007). Genome shuffling of ethanologenic yeast Candida krusei for improved acetic acid tolerance. *Biotechnol Appl Biochem.*

Wei, X. X., Shi, Z. Y., Yuan, M. Q. & Chen, G. Q. (2008). Effect of anaerobic promoters on the microaerobic production of polyhydroxybutyrate (PHB) in recombinant Escherichia coli. *Appl Microbiol Biotechnol.*

Widner, B., Behr, R., Von Dollen, S., Tang, M., Heu, T., Sloma, A., Sternberg, D., DeAngelis, P. L., Weigel, P. H. & Brown, S. (2005). Hyaluronic acid production in Bacillus subtilis. *Applied and Environmental Microbiology* **71**, 3747-3752.

Witkin, E. M. (1976). Ultraviolet mutagenesis and inducible DNA repair in Escherichia coli. *Bacteriol Rev* **40**, 869-907.

Woods, D. R. (1995). The genetic engineering of microbial solvent production. *Trends Biotechnol* **13**, 259-264.

Wyatt, M. D. & Pittman, D. L. (2006). Methylating agents and DNA repair responses: Methylated bases and sources of strand breaks. *Chem Res Toxicol* **19**, 1580-1594.

Yomano, L. P., York, S. W. & Ingram, L. O. (1998). Isolation and characterization of ethanol-tolerant mutants of Escherichia coli KO11 for fuel ethanol production. *J Ind Microbiol Biotechnol* **20**, 132-138.

Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. (2006a). Designed divergent evolution of enzyme function. *Nature* **440**, 1078-1082.

Yoshikuni, Y., Martin, V. J., Ferrin, T. E. & Keasling, J. D. (2006b). Engineering cotton (+)-delta-cadinene synthase to an altered function: germacrene D-4-ol synthase. *Chem Biol* **13**, 91-98.

Young, J. D., Walther, J. L., Antoniewicz, M. R., Yoo, H. & Stephanopoulos, G. (2008). An elementary metabolite unit (EMU) based method of isotopically nonstationary flux analysis. *Biotechnol Bioeng* **99**, 686-699.

Yu, H. & Stephanopoulos, G. (2008). Metabolic engineering of Escherichia coli for biosynthesis of hyaluronic acid. *Metab Eng* **10**, 24-32.

Yu, H., Tyo, K., Alper, H., Klein-Marcuschamer, D. & Stephanopoulos, G. (2008a). A high-throughput screen for hyaluronic acid accumulation in recombinant Escherichia coli transformed by libraries of engineered sigma factors. *Biotechnol Bioeng.*

Yu, L., Pei, X., Lei, T., Wang, Y. & Feng, Y. (2008b). Genome shuffling enhanced L-lactic acid production by improving glucose tolerance of Lactobacillus rhamnosus. *J Biotechnol* **134**, 154-159.

Yura, T. & Nakahigashi, K. (1999). Regulation of the heat-shock response. *Curr Opin Microbiol* **2**, 153-158.

Yura, T., Nagai, H. & Mori, H. (1993). Regulation of the heat-shock response in bacteria. *Annu Rev Microbiol* **47**, 321-350.

Zhang, L., Spratt, S. K., Liu, Q., Johnstone, B., Qi, H., Raschke, E. E., Jamieson, A. C., Rebar, E. J., Wolffe, A. P. & Case, C. C. (2000). Synthetic zinc finger transcription factor action at an endogenous chromosomal site. Activation of the human erythropoietin gene. *J Biol Chem* **275**, 33850-33860.

Zhang, Y. X., Perry, K., Vinci, V. A., Powell, K., Stemmer, W. P. & del Cardayre, S. B. (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415**, 644-646.

Zhao, K., Liu, M. & Burgess, R. R. (2005). The global transcriptional response of Escherichia coli to induced sigma 32 protein involves sigma 32 regulon activation followed by inactivation and degradation of sigma 32 in vivo. *J Biol Chem* **280**, 17758-17768.

Zhong, J. J. & Yue, C. J. (2005). Plant cells: secondary metabolite heterogeneity and its manipulation. *Adv Biochem Eng Biotechnol* **100**, 53-88.

Zigova, J. & Sturdik, E. (2000). Advances in biotechnological production of butyric acid. *Journal of Industrial Microbiology & Biotechnology* **24**, 153-160.