A MOLECULAR BIOLOGY APPROACH TO THE STUDY OF SINGLE CELL PROTEIN PRODUCED FROM BREWERY WASTEWATER

by

Jackson Zan Lee
A thesis submitted to the Faculty and Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Environmental Science and Engineering).

Golden, Colorado

Date \textit{Mar. 7, 2012}

Signed: \textit{Jackson Z. Lee}

Signed: \textit{Dr. John R. Spear}
Advisor

Golden, Colorado

Date \textit{Mar. 7, 2012}

Signed: \textit{John M.}

Dr. John E. McCray
Professor and Department Head
Department of Civil and Environmental Engineering
ABSTRACT

Advances in DNA and protein sequencing techniques have enabled biotechnologists to collect information on a scale never before seen. These new molecular biology tools can be combined with applications of existing technologies in the wastewater field. In this project we use these bioinformatics, genomics, and proteomics tools to evaluate the manufacture and use of a Single Cell Protein (SCP) product. This SCP is intended as a sustainable aquaculture fish meal replacement product and was made in a brewery wastewater treatment pilot plant facility in conjunction with several industry partners.

The use of DNA sequencing tools to study engineering problems is not well established. Therefore, a survey, identify, examine, and propose approach is taken to study this application. As a first step, an entire wastewater treatment plant facility was sampled over time, as well as the SCP production pilot bioreactor. Results from this study indicate that reactor type, size, and time influence a natural turnover of microbial consortia, and that disruptions can greatly influence logarithmic distributions of microbial communities. In this analysis, sequences related to several species of diazotrophs were identified as being significantly enriched by the pilot bioreactor. However, as diazotrophs possess a variety of metabolic lifestyles, this indicated several possibilities relating to nitrogen metabolism, depending on environmental conditions.

In the next stage of the project, the metabolic lifestyle of these diazotrophic organisms was studied with protein and proteomic techniques to elaborate what relationships exist between environment, consortia, and SCP protein content. This study identified the role of organic acids as a major substrate for diazotrophs in the bioreactor. SCP protein quantity and quality were mostly steady over time, but proteomes had substantial variation of proteins. Furthermore, proteomic results indicate that growth on organic acids was responsible for the major enrichment of proteins in SCP, primarily as non-essential amino acids, but overall, amino acid composition is tightly regulated in this system. This suggests that acidogenic pretreatment can be a useful addition to produce consistent SCP, and that improvement efforts should target increased production over amino acid variation.
TABLE OF CONTENTS

ABSTRACT ............................................................................................................................. iii
LIST OF FIGURES ................................................................................................................ viii
LIST OF TABLES .................................................................................................................... x
ACKNOWLEDGEMENTS ........................................................................................................ xi
DEDICATION ........................................................................................................................... xii

CHAPTER 1  INTRODUCTION ................................................................................................. 1
  1.1  Research Motivation ........................................................................................................ 1
  1.2  Project Overview and Approach ...................................................................................... 3
  1.3  Research Questions .......................................................................................................... 7
  1.4  Research Tasks ................................................................................................................ 8
    1.4.1  Task 1: Microbial Ecology Survey and Identification of Significant Organisms in Wastewater Treatment ......................................................... 8
    1.4.2  Task 2: Protein in Single Cell Protein ...................................................................... 8
  1.5  Organization of Dissertation .......................................................................................... 8

CHAPTER 2  SINGLE CELL PROTEIN FROM BREWERY WASTEWATER FOR AQUACULTURE FEEDS: CURRENT CHALLENGES, FUTURE PROSPECTS .... 11
  2.1  Abstract ........................................................................................................................ 11
  2.2  Introduction .................................................................................................................... 11
  2.3  Activated Sludge as Single Cell Protein ........................................................................ 13
  2.4  Engineering SCP from Brewery Wastewater ............................................................... 15
  2.5  Process Design and Optimization ................................................................................ 18
  2.6  Current Application of BSCP to Farmed Fish ............................................................... 20
  2.7  SCP-based Feeds .......................................................................................................... 21
  2.8  BSCP in the Global Perspective .................................................................................... 25
  2.9  Future Prospects and Research .................................................................................... 27
  2.10 Conclusion ................................................................................................................... 29

CHAPTER 3  METHODS FOR HIGH-THROUGHPUT SEQUENCING AND BIOINFORMATICS ................................................................................................. 31
3.1 Summary ........................................................................................................... 31
3.2 Background of High-Throughput Sequencing .................................................. 31
3.3 Differences between Sanger Sequencing and Pyrosequencing .......................... 35
3.4 Preparation of Samples for PyroPCR ................................................................ 37
3.5 Bioinformatics .................................................................................................. 39
  3.5.1 Flowgram Clustering and Denoising .......................................................... 41
  3.5.2 Chimera Checking ..................................................................................... 41
  3.5.3 Sequence Alignment, Clustering, Classifications, and Tree-building ....... 46
3.6 Microbial Ecology ............................................................................................. 50
  3.6.1 Alpha Diversity and Evenness Metrics .................................................... 51
  3.6.2 Beta Diversity .......................................................................................... 52
3.7 Multivariate Statistics ....................................................................................... 56
  3.7.1 Principle Component Analysis / Principle Coordinate Analysis (PCA/PCoA) .................................................................................. 57
  3.7.2 Co-abundance ......................................................................................... 59

CHAPTER 4 MICROBIAL TURNOVER, PERTURBATION, AND SELECTION ACROSS A BREWERY WASTEWATER TREATMENT WORKS AND PILOT BIOREACTOR ................................................................. 63
  4.1 Abstract ........................................................................................................ 63
  4.2 Introduction .................................................................................................... 64
  4.3 Materials and Methods .................................................................................. 65
    4.3.1 Sampling and Sanger Sequencing .......................................................... 65
    4.3.2 Pyrosequencing & Bioinformatics Pipeline .......................................... 66
    4.3.3 Alpha and Beta Diversity Analysis ....................................................... 67
    4.3.4 Co-abundance Analysis ........................................................................ 68
    4.3.5 Nucleotide Accession Numbers ........................................................... 69
  4.4 Results ........................................................................................................... 70
    4.4.1 Microbial Diversity Across a Treatment Plant ...................................... 70
    4.4.2 Temporal Dynamics of a Pilot Reactor ................................................. 75
    4.4.3 Co-abundance of OTUs Across Time .................................................... 77
  4.5 Discussion ..................................................................................................... 83
4.5.1 Insights into Microbial Diversity in Wastewater Treatment Plants........83
4.5.2 Alpha Diversity and Coverage..............................................88
4.5.3 Normal Variation vs. Disruption..........................................89
4.5.4 Common, Cosmopolitan, and Endemic Microorganisms in Wastewater Treatment.................................................................90
4.5.5 Sources of Immigration in a Pilot Reactor...............................90
4.5.6 The Core Microbiome and the “Rarer” Biosphere of Wastewater ......91
4.6 Conclusion.................................................................................92

CHAPTER 5 PROTEINS AND PROTEOMICS OF A SUSTAINABLE FEED PRODUCED FROM BREWERY WASTEWATER TREATMENT..................93
5.1 Abstract ..................................................................................93
5.2 Introduction .............................................................................94
5.3 Proteins in the Wastewater, Feed, and Biotechnology Sectors..........97
5.3.1 Feed Industry Practice..............................................................97
5.3.2 Laboratory Detection of Protein...............................................99
5.3.3 Fractionation of Proteins..........................................................100
5.3.4 Identification of Proteins..........................................................101
5.4 Materials and Methods ................................................................101
5.4.1 Overview of Production and Study Site......................................101
5.4.2 Crude Protein and Amino Acid Profiles.....................................103
5.4.3 Excitation-Emission Spectra.....................................................104
5.4.4 Fractionation of Proteins for 2D-PAGE.....................................104
5.4.5 LC-MS/MS Peptide Digest Preparation......................................106
5.4.6 Peptide Identification...............................................................106
5.4.7 Organic Acid Analysis.............................................................107
5.5 Results .....................................................................................108
5.5.1 Results from Holistic Analysis..................................................108
5.5.2 Amino Acid Profiles...............................................................111
5.5.3 2D-PAGE..............................................................................111
5.5.4 Peptide Identification...............................................................113
5.6 Discussion ................................................................................120
5.6.1 Amino Acid Variation ................................................................. 120
5.6.2 Metabolic Activities Related to SCP Production ....................... 120
5.6.3 Proteome Variation ................................................................. 121
5.6.4 Three Strategies for Protein Enhancement ............................... 121
5.7 Conclusion .................................................................................. 122

CHAPTER 6 CONCLUSIONS AND FUTURE WORK ............................... 123
6.1 Summary of Conclusions for Research Tasks ............................... 123
6.1.1 Task 1: Microbial Ecology Survey and Identification of Significant Organisms in Wastewater Treatment ....................................... 123
6.1.2 Task 2: Protein in Single Cell Protein ...................................... 124
6.2 Conclusions and Recommendations ........................................... 125
6.3 Future Work ............................................................................... 125
6.3.1 Making a Better SCP Product ................................................ 126
6.3.2 Remaining Questions on the Structure of Microbial Communities .... 126
6.3.3 Intentional Microbial Perturbation ......................................... 127

REFERENCES CITED ........................................................................ 131

APPENDIX ....................................................................................... 153
Personal Motivation ........................................................................ 153
Supplemental Figures and Tables ..................................................... 157
## LIST OF FIGURES

Figure 1.1. Collapse of Atlantic cod stocks off the East Coast of Newfoundland in 1992 [4] ........................................................................................................ 1

Figure 1.2. Project site overview schematic with treatment plant, pilot plant, and final product images .............................................................................. 4

Figure 1.3. Microscopy of pilot bioreactor mixed liquor ............................................................................................................................... 5

Figure 1.4. The increase of known divisions of organisms with no cultured representatives, (Courtesy Kirk Harris, CU-Boulder) ........................................ 6

Figure 2.1. Schematic of the SCP formulated feed production process as applied in a brewing process. Adapted from [23], [29], [47]: ........................................ 17

Figure 2.2. Global nitrogen cycling with a focus on aquaculture (global flows and magnitudes derived from [72], [75]) ........................................................................ 26

Figure 3.1 The functioning of 454 pyrotag sequencing, reproduced from [96] .......... 36

Figure 3.2. Primer composition for 454 GS FLX Platform .......................................................... 38

Figure 3.3. Results of sequence distribution across samples for two runs ................. 40

Figure 3.4 Chimera breakeven points along V1 and V2 regions of the 16S gene using moth implementation of ChimeraSlayer ........................................ 44

Figure 3.5. Taxonomic recovery of the customized RDP Bayesian classifier trained on a customized Silva 104 NR dataset .................................................. 49

Figure 3.6. Rarefaction curves of sampling effort in the study of the next chapter with environment types clustered based on the sampling legend in Figure 4.1 ......................................................................................................... 53

Figure 3.7. Effect of OTU sharing and sample similarity on OTU distances .............. 61

Figure 4.1. Overview of the study site showing the brewery treatment works (top) and the pilot scale reactor (bottom) ................................................. 65

Figure 4.2. High-rise figure of treatment works microbial diversity ........................ 72

Figure 4.3. Phylum level distribution clustered by pairwise UniFrac unweighted distance ....................................................................................... 74

Figure 4.4. A novel genus related to genus Arcicella ........................................ 76

Figure 4.5. Alpha diversity dynamics of reactor startup and disruption ................. 78

Figure 4.6. Microbial turnover across a reactor during startup and disruption ...... 79

Figure 4.7. OTUs in a pilot reactor and SCP product clustered over time............... 84

Figure 4.8. Relationship between abundance and temporal dispersion ................. 85

Figure 4.9 Phylogenetic tree and sequence abundance of depleted and enriched OTUs ......................................................................................... 86
Figure 5.1. Distribution of diazotrophic sequences as a fraction of all sequences over the study period in the pilot reactor. ................................................................. 95

Figure 5.2. Overview of the study site showing the brewery treatment works (top) and the pilot scale reactor (bottom) ................................................................. 102

Figure 5.3. Organic acid composition of the Oberon pilot plant (solid lines) and acidogenic basin (dashed lines) by mass basis. ......................................................... 109

Figure 5.4. Crude protein and total amino acid over the study period (top) and production (bottom), with 3-day moving averages ........................................... 110

Figure 5.5. Excitation-emission matrix of pilot bioreactor mixed liquor .................. 110

Figure 5.6. Amino acid (AA) composition over time, total amino acid basis ........... 111

Figure 5.7. Amino acid variation analysis .......................................................... 112

Figure 5.8. 2D-PAGE gels of fractionated proteins from pilot bioreactor Day 175 (A), Day 335 (B), Day 365 (C), SCP product Day 365 (D), plant influent Day 365 (E). Numbers mark spots selected for sequencing... 114

Figure 5.9. Peptide identification table for 2D-PAGE gel A .................................. 115

Figure 5.10. Peptide identification table for 2D-PAGE gel B ............................... 116

Figure 5.11. Peptide identification table for 2D-PAGE gel C ............................... 117

Figure 5.12. Peptide identification table for 2D-PAGE gel E ............................... 118

Figure A1. The Nile Tilapia (photo: wiki commons) ............................................. 155

Figure A2. Tilapia cultivation, in thousands of tonnes per year harvested (FAO) ...... 156
LIST OF TABLES

Table 2.1. COD loadings of the alcohol fermentation and food processing industries .......................................................... 14
Table 2.2. Amino acid profile and dietary information of several fish meal replacements ..................................................... 22
Table 3.1. Current High-Throughput Sequencing Technologies and their per run and per base pair yields and costs. Reproduced from [101] ........................................ 34
Table 3.2. Chimeras detected using three different detection algorithms ................................................................. 45
Table 4.1. Alpha diversity estimates, sampling schedule, and operating parameters of each unit process. Note: Operating parameters are approximate. ..... 71
Table 5.1. Common characteristics of organic acid utilizing diazotrophs and acidogenic bacteria detected in the pilot bioreactor .................................................. 96
Table 5.2. Different types of fish meal replacement alternatives and their limiting amino acids (LAA, right column). Reproduced from [28] ...................... 98
Table 5.3. Standard industry protein analysis techniques ......................................................... 102
Table A1. Chemical Protein Methods Summary ................................................................. 157
Table A2. Full Peptide Identification Tables ................................................................. 158
ACKNOWLEDGEMENTS

They say that we scientists stand on the shoulders of giants, but it is better said that we stand on the shoulders of friends, without whose support and goodwill we would have faltered, fallen, and given up long ago. Though I suppose having giant friends isn’t bad either.
DEDICATION

For my parents and my grandparents.
CHAPTER 1
INTRODUCTION

1.1 Research Motivation

Already half of all fish stocks have been deemed fully exploited [1], meaning “nearly half of the main marine stocks offer no reasonable expectation for further expansion” [2]. This has led to the collapse of several fisheries including the Atlantic Cod fishery (Figure 1.1), and the expected collapse of many others over the next several decades [3].

![Graph showing fish landings in tons from 1850 to 2000, with a significant drop in 1992.](Source: Millennium Ecosystem Assessment)

Figure 1.1. Collapse of Atlantic cod stocks off the East Coast of Newfoundland in 1992 [4]

Concomitantly, the aquaculture industry (the industrial farming of fish) has grown at an annual rate of 14% since 1970 [5-6]. Because aqua feed production relies on significant amounts of fishmeal protein unsustainably harvested from ocean fisheries, further aquaculture growth will result in fish meal shortages and environmental degradation. Therefore, it will become increasingly necessary and economic to begin the reuse and reprocessing of waste materials from multiple industries into the aquaculture
feed sector, both as a consequence of increasing prices for decreasing fish stocks and to safeguard the environment through recycling of potential wastes [7]. For example, solid food byproducts of various forms from the fermentation industry (beer, cheese, wine, etc.) have been changed from a costly landfill waste to a cheap commercial feedstock.

However, as food manufacturers learn to separate useful organic solid fractions from their food waste streams, the usefulness of the waste streams decrease until only residual wastewater remains to be sent to wastewater treatment works. One emerging frontier is the reuse of residual industrial wastewater itself to produce useful byproducts. Industrial wastewater typically contains large untapped amounts of useful dissolved carbon measured as Biological Oxygen Demand (BOD) that can be utilized for protein production for fish feed in the form of Single Cell Protein (SCP). Protein is produced from the growth of bacteria as it consumes the BOD and is harvested and processed into an SCP product. Industrial wastewater is advantageous for its low cost and also for its lack of toxic contaminants such as heavy metals and human wastes typically found in domestic wastewater.

Researchers have recognized SCP’s potential as an animal feed for at least two decades, though its successful use on a large scale has not been developed [8]. Changing factors today combine to render a favorable environment for renewed efforts to exploit this feed ingredient. First, stricter environmental regulations have led to higher operating costs for treatment facilities. Second, rising fishmeal costs have made alternative protein meal sources competitive. Finally, the aquaculture industry has expanded over the last 30 years and currently includes a variety of species that were not of significant commercial interest previously [7]. The potential market for such fish meal replacement is vast. Approximately 200 million tons of protein meal is sold in the world per year for various human and animal feed purposes, of which fishmeal accounted for roughly 6-7 million tons [7]. By comparison a large brewery can potentially produce 2,000-30,000 tons of SCP a year. Although it is difficult to extrapolate from brewery and other food processing sources (molasses, yeast, juicers, vegetable processors, breweries, distillers) the possible worldwide production capacity of SCP could reach hundreds of thousands of
tons a year in recaptured protein from wastewater. Therefore, SCP can help retain biomass within the world’s fisheries and ecosystems thereby preserving diversity and providing an additional value-added market for industrial food byproducts.

1.2 Project Overview and Approach

The research of this dissertation was completed at a single field site, a pilot bioreactor established at New Belgium Brewery from 2008-2009 for the purpose of producing an SCP product from brewery waste. An overview of the project site Wastewater Treatment Plant (WWTP) is shown in Figure 1.2. The brewery waste was received by holding basins for flow normalization, followed by methanogenesis Upflow Anaerobic Sludge Blanket (UASB) treatment. Finally, remaining waste was sent to a conventional anaerobic treatment and clarification process prior to discharge to the municipal Publicly Owned Treatment Works (POTW). Preliminary microscopy of the mixed liquor and final product showed no floc formation (typical for young sludge), though it did show a cohesive film, possibly extracellular polymeric substances (EPS) attaching cells, and the presence of plant debris in the mixed liquor and a variety of cell morphotypes (Figure 1.3). A main objective of this study was to promote higher rates of protein production within the reactor. The key to this protein quality is the addition of micronutrients to the wastewater during aerobic growth. It has been known for some time that industrial wastewater lacks key micronutrients and that addition of these micronutrient cocktails results in increased carbon metabolism, increased microbial diversity, and an increase in the microbial biomass that serves as the basis for SCP [9]. By supplementing with these micronutrients, a pilot plant at New Belgium Brewing showed an increase in protein content up to and exceeding the economic target of 55-60% (w/w) crude protein. However, even though this target has been achieved in a pilot plant study, questions related to how biochemistry and environment influence microbial consortia are unresolved. This means that the ability to understand changes in protein content due to upsets and the ability to scale this technique to other sites is largely undeveloped particularly given the “black box” nature of conventional wastewater treatment. The ultimate goal of this research effort is to develop an understanding of how to create a better SCP product with technology that scales to multiple sites.
Figure 1.3. Microscopy of pilot bioreactor mixed liquor. (A) Mixed liquor plant matter (bright field), (B) EPS attaching cells (100x, phase contrast (PC) 3), (C) Eukaryote, spirilla, and cocci in mixed liquor (1400x, PC 3), (D) Dried SCP final product (Courtesy, Xiaoyu Fang).
Rapid technological leaps in high throughput sequencing have markedly increased our ability to detect environmental microorganisms from only a few hundred sequences per run to hundreds of thousands of sequences per run with the potential for millions of sequences per run in the immediate future. This development is critical since the vast majority of microorganisms cannot be effectively cultured and are only studied from a molecular biology standpoint (Figure 1.4). This increase in sequencing throughput with decreased cost has greatly expanded the detection coverage of microbial diversity in environmental samples through the use of pyrotag technology [10]. This technology has already been able to make significant discoveries of the structure and role of microbial consortia in the human microbiome [11], deep-sea ecosystems [12], and in raw sewage [13] to name a few examples. Coupled with this increased sequencing capability has

Figure 1.4. The increase of known divisions of organisms with no cultured representatives, (Courtesy Kirk Harris, CU-Boulder)
been the crossover application of tools related to the purification and characterization of proteins from the clinical biotechnology field [14]. The improved understanding of the characteristics of protein behavior has led to advances in wastewater constituent separation and fluorescence imaging techniques [15–17]. Additionally, traditional biochemistry tools for protein separation such as 2D-PAGE are now being applied to wastewater [18] and are being coupled with new protein sequencing technologies [19] to directly obtain the protein composition found in environmental samples [20]. In this dissertation, genomic and proteomic tools were used to study the environmental microbial processes utilized for SCP production to determine and develop targets for product enhancement.

Due to the fundamentally complex nature of microbial ecology problems, significant qualitative surveys are first needed to characterize novel systems. From there, an analysis of the organization of these communities must be made before important microbial subpopulations of these systems can be identified. Next, the significance and relevance of these identified populations to the problem of study need to be examined. We used this survey, identify, and examine approach to study environmental microbiological datasets derived from the wastewater treatment plant described above. From this we then seek to propose testable hypothesis, despite the limited ability to model such systems.

1.3 Research Questions

The research questions in this thesis are primarily descriptive and try to ask what is the holistic nature of wastewater systems, what parts of these systems are important, and what kind of information can be collected.

1. What kinds of microbes are associated with wastewater and SCP in particular?
2. What is the normal variation of microorganisms in a wastewater treatment plant?
3. What kinds of microbes are consistently enriched in SCP production?
4. Do proteins isolated from SCP production match the expected metabolism of these organisms?
5. What process or mechanism is responsible for SCP protein formation?
1.4 Research Tasks

The dissertation research tasks are listed below. These tasks are then re-addressed in the conclusion chapter.

1.4.1 Task 1: Microbial Ecology Survey and Identification of Significant Organisms in Wastewater Treatment
1. Define patterns of alpha diversity measurements that indicate changes in the functioning of the plant
2. Determine the pattern of beta diversity and co-abundance that indicate succession and disruption of bioreactors.
3. Determine the most important factors of variation of microbial ecology across a treatment plant.
4. Identify the most commonly enriched organisms in SCP production and their distribution.
5. Examine common metabolism of enriched organisms.

1.4.2 Task 2: Protein in Single Cell Protein
1. Examine SCP by traditional colorimetric, fluorometric, HPLC, and feed industry techniques
3. Identify proteins related to SCP production and organisms from which SCP proteins derive.
4. Examine common metabolisms from Task 1 to see if there is agreement.

1.5 Organization of Dissertation
Chapter 1: Introduction
Chapter 1 is an introduction to the research motivations of this dissertation and the objectives and organization of the dissertation.
Chapter 2: Single Cell Protein From Brewery Wastewater for Aquaculture Feeds: Current Challenges, Future Prospects
Chapter 2 is a literature review of the current state of affairs with SCP production and testing. This chapter is a review paper accepted to Reviews in Aquaculture.

Chapter 3: Methods for High-Throughput Sequencing and Bioinformatics
Chapter 3 describes methods of pyrotag laboratory preparation and bioinformatics analysis used in this dissertation.

Chapter 4: Microbial Turnover, Perturbation, and Selection Across a Brewery Wastewater Treatment Works and Pilot Bioreactor
Chapter 4 addresses the first task, the complex and rapidly evolving process of how we actually use DNA sequencing technology to survey an increasing amount of microbial diversity. This chapter is an original research article submitted to The International Society of Microbial Ecology (ISME) Journal.

Chapter 5: Proteins and Proteomics of a Sustainable Feed Produced from Brewery Wastewater Treatment
Chapter 5 addresses the second task, and how the infant field of environmental proteomics can be used to directly measure ecosystem activities at the protein level. This chapter is an original research article in submission to Biotechnology and Bioengineering.

Chapter 6: Conclusions and Future Work
Conclusions about the dissertation and potential areas of future research and concluding remarks are presented here.
CHAPTER 2
SINGLE CELL PROTEIN FROM BREWERY WASTEWATER FOR AQUACULTURE FEEDS: CURRENT CHALLENGES, FUTURE PROSPECTS

Author’s Note: This chapter presents the work of a paper in review at *Reviews in Aquaculture*. Authors: Jackson Lee, Andrew Logan, Seth Terry, and John Spear. Co-authors provided expert resources on SCP production and wastewater treatment.

2.1 Abstract

The production and use of single cell protein, an alternative protein source derived from the growth of microbes, has been a developing field for several decades. Recent interest in fisheries sustainability and increased prices of fish meal have turned new attention to the concept of alternative protein replacement, particularly from single cell protein grown from brewery or other food processing wastes. With global fish stocks in decline, SCP used in aquaculture can ultimately become an integral part of sustainable fisheries management. However, a number of challenges exist with the production and use of SCP such as site selection, performance optimization, and digestibility. In this review, we outline the overall production process of single cell protein from brewery wastewater and discuss some of the current engineering design considerations, economics, global cycling of nitrogen due to aquaculture, and nutritional effectiveness in growth trials, as well as the potential pitfalls and future research possibilities of this kind of fish meal protein replacement.

2.2 Introduction

Single Cell Protein (SCP) refers to microbial biomass produced for use as a protein supplement for animal consumption. In the 1960s, researchers invested considerable effort to find ways to grow protein from petroleum sources [21] as a solution to the predicted protein shortfall expected to impact world hunger. However, petroleum-based SCP fell out of favor as high oil prices of the 1970’s, increased capital costs, and increased food production made the economics unfavorable [22]. Meanwhile, research on other less expensive carbon substrates for SCP production continued. By the 1980s,
SCP research centered on processing activated sludge microbial biomass from various domestic and industrial wastewater treatment works as a potential method to increase animal feed protein supply economically, particularly for aquaculture [23].

Today, two very powerful drivers have spurred renewed development of SCP: fisheries depletion and aquaculture growth. Advancements in harvesting technology have led to losses of major fisheries due to overfishing of the world's oceans. Today, 63% of 166 stocks examined are in need of rebuilding, and there is evidence that by 2048, overfishing will lead to the collapse of the majority of fish stocks (though the magnitude is still debated) [3], [24]. According to the United Nations Food and Agriculture Organization (FAO), approximately 90 million tonnes of fish were taken from the world's oceans in 2010 and many species are now considered over-exploited or depleted [25].

At the same time, global aquaculture has grown rapidly to satisfy the basic protein needs of global population growth. In the period of 1970-2008 aquaculture cultivation increased at an average rate of 8.3% per year, with 45.7% of all seafood consumed in 2008 coming from aquaculture [25]. However, aquaculture production consumed 3 million tonnes (56%, in 2006 landings) of global fish meal production [26]. This is a major 'sticking point' for the rationale of using aquaculture to sustainably address global protein needs [1], as increasing amounts of fish harvested from oceans are needed to sustain aquaculture growth. The conclusions from Worm et al. (2006) coupled with increased coverage in news media on the overexploitation of the world's fisheries, has cast public attention on the issue of fisheries management and has led to a demand for 'environmentally friendly' seafood. Yet, by currently available practices, aquaculture still requires about 0.63 pounds of wild caught fish to produce one pound of farmed fish on average, with some species requiring much higher ratios of fish inputs [27]. Up to 70% of feeds for carnivorous fish and 50% for shrimp can be comprised of fish meal and fish oil [28]. With fish meal production flat and the price of fish meal peaking at 1800 USD / tonne in 2010 [25], feed costs normally contribute the single largest production cost for commercial aquaculture [27]. While there are a number of cheaper alternatives
to fish meal currently in the market, e.g. slaughterhouse wastes and agricultural residues [28], many have problems with their use, e.g. essential amino acid deficiency [29]. This has led researchers to explore numerous other potential sources, with SCP-based alternatives as one of the more economical, faster-growing, and easy to produce alternatives [8], [28], [30].

2.3 Activated Sludge as Single Cell Protein

By far the largest amount of SCP research (aside from petroleum substrates) has centered upon using wastewater to produce SCP. In these systems, microbes feed on nutrients in the waste streams to form high-protein aggregates that can be harvested as a protein source for consumption. Proposed schemes for SCP production span different microbe varieties from all three domains of life (Bacteria, Archaea and Eucarya) grown on different varieties of substrates derived from industrial wastes. Much of this research was performed in the Western Hemisphere before 1990 and has since shifted to emerging economies such as India and China. A number of excellent review articles and book chapters exist that summarize this body of work spanning several decades [22], [28], [31–36].

To date, however, few of these SCP-from-wastewater concepts have been commercially implemented and none have achieved large commercial success. This has been attributed to a combination of lower protein prices, poor wastewater composition, high variability of wastewater flows, health risks, or heavy metal toxicity [30]. Despite these attributes, several researchers have suggested that food-processing wastes in particular possess distinct characteristics that might make the technology profitable [23], [37]. Domestic wastewater biomass sources are often plagued by heavy metal contamination and pathogens as part of their process stream [29], [38]. Food-processing wastewaters do not have these issues [23]. Unlike the growth of pure culture organisms on wastewater (which require costly processing steps like pre-sterilization [23] or pH adjustment to prevent contamination [30]), mixed-culture systems based on conventional aerobic wastewater treatment are more stable, easily maintained, and less susceptible to contamination [33]. Additionally, brewery and food processing wastewater chemical
oxygen demand (COD) levels tend to be significantly higher than domestic wastewater (Table 2.1), which should translate into higher biomass production.

Further advantages for brewery wastewaters over other types of food-processing wastewaters are that they typically produce year-round (though it should be noted that variable day-to-day flow rates can impact the process wastewater microbial biomass [39]), and brewery-derived SCP has been reported to have a lysine to methionine ratio (two essential amino acids absent from many plant sources) that is comparable to fish meal [23]. This last point is critical for the production, implementation and use of SCP as a viable alternative for a more sustainable fish meal replacement.

Table 2.1. COD loadings of the alcohol fermentation and food processing industries

<table>
<thead>
<tr>
<th>Process</th>
<th>COD</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple Processing Waste</td>
<td>18,700</td>
<td>[40]</td>
</tr>
<tr>
<td>Tomato Processing Waste</td>
<td>1,500</td>
<td>[40]</td>
</tr>
<tr>
<td>Sugar Beet Processing Waste</td>
<td>600-3,000</td>
<td>[33]</td>
</tr>
<tr>
<td>Brewery Wastewater</td>
<td>1,800-17,000</td>
<td>[33], [41]</td>
</tr>
<tr>
<td>Distillery Wastewater</td>
<td>10,000-190,000</td>
<td>[33], [42]</td>
</tr>
<tr>
<td>Winery Wastewater</td>
<td>13,230-23,170</td>
<td>[43]</td>
</tr>
<tr>
<td>Domestic Wastewater</td>
<td>259-800</td>
<td>[44]</td>
</tr>
</tbody>
</table>

Because of these advantages, this review will focus mainly on the challenges of the production, economics, pollution, and nutritional quality of mixed culture SCP from food-processing wastes in general, and brewery wastes specifically. This kind of product has been called activated sludge in feed [23], or Activated Sludge Single Cell Protein (AS-SCP) [45], or even a Biofloc protein source. All of these terms are fairly broad and can include multiple technologies in different fields of study. De Muyl and colleagues referred to brewery-derived dried activated sludge as Brewery SCP (BSCP) [46], therefore, in this paper we will use the term BSCP to distinguish the optimized growth of Brewery wastewater activated sludge to produce Single Cell Protein fish meal.
alternatives specifically. All other microbial protein production from previous research, some times called floc, sludge, or other terms will be referenced as SCP in general.

2.4 Engineering SCP from Brewery Wastewater

Due to the need to prevent wastewater discharges from harming the environment, large breweries and other food-processing facilities often have existing wastewater treatment facilities to comply with local water regulations. The fundamentals of converting this treatment process to the production of BSCP from wastewater have not changed particularly in the several decades since the paper described by Vriens and colleagues [23]. In the case of breweries (Figure 2.1A), a number of solid and liquid organic wastes can be produced as part of the malting stage or brewing stage of the brewing process. Malting is the preparation of barley by steeping, germination, and kilning (heated drying) to produce the enzymes and sugars needed for the brewing process. Each stage of this process produces a number of draining and wash wastes as well as solid screenings such as malt sprouts. The brewing process involves the crushing and mashing of the malted barley followed by boiling of the extract (wort), then fermentation with yeast, and finally filtration of the resulting mature beer. Solid process wastes typically consist of spent hops and barley, yeast solids, and diatomaceous earth (DE) used in filtration. Liquid wastes that remain consist of waste beer or first and last runnings, tank bottom liquids, and tank wash or drainings. Much of the solid waste can be dried and pelleted or used directly for animal feed with some of the wet wastes mixed in with the feeds or recycled in the brewing process [39], [41], [47]. It is the remaining liquid wastes, often with dissolved and suspended organic solids, which must be treated at an on-site process wastewater treatment facility.

Wastewater treatment facilities are complex, but generally follow a basic layout (Figure 2.1B) [23], [44]. After being received at the wastewater treatment plant, brewery wastewater is typically screened and sent to a holding basin to buffer the flow rates (Figure 2.1B: (1)). A constant stream is then sent to the aeration basins (2) for conventional activated sludge treatment; a process where a community of microorganisms metabolizes influent wastewater carbon and nutrients, aided by mechanically supplied air, with the overall aim of encouraging the formation of flocs
(large aggregates of microbes and inert or dead particulate matter) that settle out of the mixed liquor in a secondary clarification stage (3). This settled material is sludge, a thickened mixture of suspended flocs consisting mostly of water. A fraction of this sludge (Returned Activated Sludge (RAS)) is returned to the aeration basin (2) as a way to increase the biological solids concentration in the aeration basins (measured as Mean Cell Residence Time (MCRT), in days, representing the average age of the sludge). This reuse of older sludge has at least two effects: to inoculate incoming wastewater with mature active microorganisms and to reduce the size of wastewater treatment reactors and their associated capital costs. The fraction of the sludge not reused (Waste Activated Sludge (WAS)) is then processed for either disposal or recovery. The fundamental equations for MCRT and F/M governing the design of such processes are as follows for a settling clarifier (3) (adapted from [23], [44]):

\[
MCRT = \frac{X}{\dot{X}} = \frac{vX}{(Q-Q_w)x_e+Q_wx_r}
\]  

(2.1)

\[
F/M = \frac{Q_{S_e}}{vX}
\]  

(2.2)

F/M is the Food (i.e., the carbon source) to Microorganism ratio and can be thought of as the ratio between these primary components in the system [47] and importantly, the F/M ratio influences the rate of treatment. MCRT is the average cell age and is defined as the total amount of biomass in a system (\(X\)) divided by the rate of removal of biomass (\(\dot{X}\)) and affects system performance parameters such as growth rates, treatment type, and cell yields. MCRT fixes the ultimate capital costs of constructing treatment basins for production of BSCP. From the above equations, several characteristics of the waste and several design parameters constrain the overall BSCP production design. The substrate concentration (\(S_e\)), and the influent flow rate (Q) are specific to each individual site. The design of the clarifier Figure 2.1B: (3) determines the operational range of the biomass concentration in the effluent (\(X_e\)), the re-circulated waste (\(X_r\)), and the fraction of flow wasted (\(Q_w\)). MCRT and F/M are typically selected to optimize treatment performance based on best practice (anywhere from 1-50 days MCRT, 0.04-1.0 g substrate/g biomass-
Figure 2.1. Schematic of the SCP formulated feed production process as applied in a brewing process. Adapted from [23], [29], [47]:

1. Holding Basin  
2. Activated Sludge Tank  
3. Clarifier or Membrane  
4. Wet Well  
5. Dewatering Stage(s)  
6. Drying / Baking  
7. Grinding and Mixing  
8. Steam Pelleting  
9. Air Drying  
10. Extrusion Forming  
11. Drying & Cooling  
12. Fat Coating & Packaging
day for F/M depending on the type of treatment desired [44]). These factors ultimately determine the volume (V) of the aeration basin (2). The treated effluent is sent to a wet well (Figure 2.1B: (4)) where it can be reused for non-potable brewery process operations (i.e. cooling water) or is discharged from the treatment facility. In particular, the high amount of COD in brewery wastes may require the treatment of wastes over two or more stages due to the large amount of COD elimination required prior to discharge [44]. In the BSCP production process, the conversion of WAS to BSCP requires that the majority of the water in the sludge be removed. Dewatering typically involves both mechanical dewatering and / or a drying process (such as by centrifugation, pressing, open-air / vacuum drying or a combination of these devices [44]) (5), followed by thermal drying to remove the final moisture (6) and can contribute considerably to the overall processing cost of the final product. Presuming the prior existence of basic activated sludge infrastructure, dewatering unit operation expenses combined with the costs related to aeration make up the bulk of capital and processing costs for wastewater SCP production.

2.5 Process Design and Optimization

Increasing disposal costs for sludge from wastewater treatment plants means a number of value-added sludge recovery technologies are now being investigated, from more mature practices such as land amendment and methanogenesis, to developing concepts such as bioproducts (bioplastics, bioflocculents, etc.), biosorbents, construction aggregates, fuels, and direct electricity generation [48]. These competing technologies and economic constraints then make optimal selection and design of the appropriate wastewater treatment facility a prime concern. Not all breweries are required by water regulations to have treatment facilities, and many facilities only partially treat their wastes due to local water regulations and the high capital cost of treatment works [41]. Existing facilities with steady flow rate, high organic loading, and strict separation of the domestic and process waste streams are the best candidates to consider for BSCP production. Unit operations that limit energy or other operational costs are also a benefit. At the same time, certain off-products can potentially ‘sour’ the value of the waste stream. Non-segregated brewery wastes can contain indigestible diatomaceous earth (depending on the filtration process) that requires a separator to be installed. Certain
toxins such as wash water containing caustics or surfactants, mixed wastes from domestic sewage, and opportunistic bacterial pathogens also pose potential hazards. Each of these factors can potentially reduce the functionality and value of the final product stream.

The BSCP production process itself must also be optimized to produce as much protein as possible (measured in the feed industry in terms of crude protein (Kjeldahl N x 6.25) [49]), in as short amount of time as possible (known as the rate of biomass production) and as efficiently as possible (termed the cell yield or substrate conversion efficiency) to maximize economic competitiveness. Yield is measured experimentally under specific systems and is a function of the species present, the type of substrate, and the environmental conditions [44]. Two strategies are taken to maximize these factors. First, biomass production can be increased with the addition of nutrients, particularly nitrogen and phosphorus [23], [50]. Second, the careful selection of a lowered MCRT for protein production performance has been shown to be critical for raising the yield and biomass production as well as the protein fraction [23], [51]. However, this has implications for the maturity of flocs in the reactor and can influence settleability in conventional clarifiers, which ultimately affects wastewater treatment plant effluent water quality and may violate local water quality standards. Some have attempted to avoid this problem by using Membrane Biological Reactors (MBRs) [52], [53].

Unreasonable production costs at any point in BSCP production can easily drive potential manufacturers from the market and careful assessments are needed before capital funds are invested. A major hurdle for existing wastewater operators is the capital cost of retrofitting plants as opposed to constructing a new system. Existing treatment facilities are often part of an aging infrastructure that was not designed with resource recovery and environmental sustainability in mind. Considering competition from a host of other wastewater utilization technologies, technologies that can be more easily retrofitted to existing plants will have a competitive advantage.
2.6 Current Application of BSCP to Farmed Fish

Once BSCP is produced it must be incorporated into a feed product at a feed mill before it can be used. Feed manufacturers and large scale farming operations receive feed ingredients and evaluate each ingredient to determine a blending ratio (Figure 2.1.C). These feed ingredients are ground and blended together (7) and either go through a steam pelleting process, by combination with a binding agent (such as high-gluten wheat flour) followed by cooling (8-9), or a cold extrusion process followed by a drying and fat-coating process (10-11). Before shipping, pellets may be coated in a thin layer of lipids to enhance palatability and provide fat prior to packaging (12). The ultimate feed can contain protein, fat, and nutrients from a variety of locations and sources tailored for the type and age class of the specific animal being grown [29].

Novel feeds often require testing of nutritional quality before commercial adoption to ensure that they meet standards expected for measurements of digestibility, water stability, and growth [54]. This is typically done in large replicate trials in environmentally isolated fish pens where feed composition, species, and water quality conditions are carefully controlled [29]. Table 2.2 shows the amino acid profile and crude protein, crude fat, and crude ash of some commonly regarded commercial replacements. No general trend exists for how much fish meal can be replaced but a number of health concerns are noted in literature. For bone meal and blood meal alternatives in shrimp cultivation, as more material replaces fish meal in feed formulations growth and survival rate are lowered (tested up to 75% fish meal replacement) [55]. No significant differences were found for fish meal replacement by blood and bone meal, up to 80% replacement for grouper (tested up to 100% replacement) [56].

Some suggested reasons as to why certain alternative feed materials do not work as well as others include problems with essential amino acid (EAA) ratios, digestibility, and/or high ash content [56-57]. Most fish require the same 10 essential amino acids: arginine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine matched to an established feeding target [29]. Lysine, cysteine, and methionine are typically the limiting amino acids seen in plant-based diets [28].
Additionally, the inclusion of bone products has been shown to increase the ash content in alternative protein sources and has been cited as a detrimental factor in proper feed digestion. In Table 2.2 the meat and bone meal shows high protein, but also very high ash values indicating high amounts of mineral material. High calcium and phosphorus and other bone mineral ashes have been found to reduce mineral absorption [58], particularly zinc, which can have developmental consequences on rainbow trout and other fishes [59]. Finally, animal digestive systems are composed of diverse physical and chemical mechanisms that are involved in the breakdown and incorporation of nutrients into tissue. The species, age, habitat, life-style, physiology, meal size and schedule, feed formulation [54] and the gut microbiome may all change the apparent digestibility of experimental feeds.

2.7 SCP-based Feeds

These same issues of digestibility, essential amino acids, and ash content have also been mentioned for SCP-based feeds, where levels of cysteine and methionine have been reported as the limiting amino acids commonly found in SCP products [28]. Despite this, SCP products tend to have methionine or cysteine quantities more similar to fish meal than their conventional feed counterparts (Table 2.2), and amino acid profiles are comparable to fish meal [60] and is an excellent source of vitamin B12 [45]. High ash content (containing Calcium, Phosphorus, and Iron) has also been found in domestic waste activated sludge SCP [45]. In feeding trials using process-wastewater produced SCP, the replacement of up to 10% fish meal with SCP (from brewery, dairy, citrus, cannery, and abattoir sludge) did not have a negative impact for a variety of animals and conditions tested [23], though in terms of the total amino acid profile of protein, waste-derived SCP is comparable with other industry alternatives such as soybean meal and meat-and-bone meal [60]. Nevertheless, digestibility has been cited as a possible limiting concern due to the resistance of bacterial cell walls to enzymatic breakdown [23]. This has also been cited as a concern for the poor digestibility of algal SCP in particular [61]. The trend seen from anaerobic poultry waste SCP has been that a modest amount of replacement (20% for channel catfish) results in increased growth and survival, but beyond that, performance was severely degraded [62]. Inclusion of bacterial SCP in
Table 2.2. Amino acid profile and dietary information of several fish meal replacements

<table>
<thead>
<tr>
<th>Product Type</th>
<th>Fish Meal&lt;sup&gt;a&lt;/sup&gt;</th>
<th>BioProtein&lt;sup&gt;b&lt;/sup&gt;</th>
<th>ProFloc&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Spirulina maxima&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Chlorella pyrenoidosa&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Soymeal&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Meat and Bone Meal&lt;sup&gt;e&lt;/sup&gt;</th>
<th>DDGS&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Norse LT-94 SCP</td>
<td>BSCP</td>
<td>Cyanobacteria</td>
<td>Algae</td>
<td>Hi-Pro</td>
<td>Beef and Bone</td>
<td></td>
<td>1.78&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>Ala</td>
<td>4.18</td>
<td>4.43</td>
<td>3.82</td>
<td>4.45</td>
<td>3.36</td>
<td>2.01</td>
<td>3.90</td>
<td>1.09</td>
</tr>
<tr>
<td>Arg</td>
<td>4.04</td>
<td>3.82</td>
<td>3.60</td>
<td>4.26</td>
<td>3.19</td>
<td>3.43</td>
<td>3.90</td>
<td>1.75</td>
</tr>
<tr>
<td>Asp</td>
<td>6.10</td>
<td>5.52</td>
<td>6.36</td>
<td>5.63</td>
<td>3.36</td>
<td>5.23</td>
<td>3.30</td>
<td>0.69</td>
</tr>
<tr>
<td>Cys</td>
<td>0.69</td>
<td>0.55</td>
<td>0.55</td>
<td>0.26</td>
<td>n.r.</td>
<td>0.72</td>
<td>0.30</td>
<td>0.56</td>
</tr>
<tr>
<td>Glu</td>
<td>9.25</td>
<td>6.54</td>
<td>8.04</td>
<td>8.25</td>
<td>5.30</td>
<td>n.r.</td>
<td>5.70</td>
<td>3.49</td>
</tr>
<tr>
<td>Gly</td>
<td>4.04</td>
<td>3.13</td>
<td>2.81</td>
<td>3.14</td>
<td>2.74</td>
<td>1.93</td>
<td>7.70</td>
<td>n.r.</td>
</tr>
<tr>
<td>His</td>
<td>1.58</td>
<td>1.29</td>
<td>1.46</td>
<td>1.18</td>
<td>0.80</td>
<td>1.25</td>
<td>0.80</td>
<td>0.69</td>
</tr>
<tr>
<td>Ile</td>
<td>3.01</td>
<td>2.73</td>
<td>3.38</td>
<td>3.93</td>
<td>1.94</td>
<td>2.05</td>
<td>1.30</td>
<td>0.97</td>
</tr>
<tr>
<td>Leu</td>
<td>5.14</td>
<td>5.04</td>
<td>5.06</td>
<td>5.24</td>
<td>2.28</td>
<td>3.61</td>
<td>2.70</td>
<td>3.05</td>
</tr>
<tr>
<td>Lys</td>
<td>5.48</td>
<td>3.47</td>
<td>4.34</td>
<td>3.01</td>
<td>4.50</td>
<td>2.90</td>
<td>2.50</td>
<td>0.71</td>
</tr>
<tr>
<td>Met</td>
<td>2.06</td>
<td>1.50</td>
<td>1.41</td>
<td>0.92</td>
<td>1.03</td>
<td>0.65</td>
<td>0.70</td>
<td>0.54</td>
</tr>
<tr>
<td>Phe</td>
<td>2.81</td>
<td>3.00</td>
<td>3.29</td>
<td>3.21</td>
<td>2.57</td>
<td>2.39</td>
<td>1.50</td>
<td>1.31</td>
</tr>
<tr>
<td>Pro</td>
<td>2.74</td>
<td>2.45</td>
<td>2.77</td>
<td>2.55</td>
<td>2.28</td>
<td>2.28</td>
<td>5.00</td>
<td>1.99</td>
</tr>
<tr>
<td>Ser</td>
<td>2.81</td>
<td>2.38</td>
<td>2.82</td>
<td>2.75</td>
<td>1.25</td>
<td>2.15</td>
<td>1.60</td>
<td>1.09</td>
</tr>
<tr>
<td>Thr</td>
<td>2.81</td>
<td>2.79</td>
<td>3.11</td>
<td>3.01</td>
<td>1.82</td>
<td>1.80</td>
<td>1.10</td>
<td>0.96</td>
</tr>
<tr>
<td>Trp</td>
<td>0.75</td>
<td>0.29</td>
<td>0.98</td>
<td>0.92</td>
<td>0.80</td>
<td>0.67</td>
<td>n.r.</td>
<td>0.20</td>
</tr>
<tr>
<td>Tyr</td>
<td>2.19</td>
<td>2.25</td>
<td>2.83</td>
<td>2.35</td>
<td>1.54</td>
<td>1.66</td>
<td>2.10</td>
<td>0.96</td>
</tr>
<tr>
<td>Val</td>
<td>3.36</td>
<td>3.75</td>
<td>3.52</td>
<td>4.26</td>
<td>2.91</td>
<td>2.22</td>
<td>2.00</td>
<td>1.33</td>
</tr>
<tr>
<td>Total AAs</td>
<td>63.02</td>
<td>55.93</td>
<td>60.15</td>
<td>59.54</td>
<td>41.67</td>
<td>36.94</td>
<td>46.10</td>
<td>22.47</td>
</tr>
</tbody>
</table>
Continued Table 2.2 Continued

<table>
<thead>
<tr>
<th></th>
<th>Fish Meal&lt;sup&gt;a&lt;/sup&gt;</th>
<th>BioProtein&lt;sup&gt;b&lt;/sup&gt;</th>
<th>ProFloc&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Spirulina maxima&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Chlorella pyrenoidosa&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Soymeal&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Meat and Bone Meal&lt;sup&gt;e&lt;/sup&gt;</th>
<th>DDGS&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moisture (%)</td>
<td>7.5</td>
<td>1.4</td>
<td>&lt;5</td>
<td>n.r.</td>
<td>n.r.</td>
<td>11.3</td>
<td>n.r.</td>
<td>14.0&lt;sup&gt;3&lt;/sup&gt;</td>
</tr>
<tr>
<td>Crude Protein (CP) (%)</td>
<td>68.5</td>
<td>68.1</td>
<td>66.0</td>
<td>60-71</td>
<td>57.0</td>
<td>47.0</td>
<td>49.2</td>
<td>27.0</td>
</tr>
<tr>
<td>Crude Fat (%)</td>
<td>10.4</td>
<td>10.4</td>
<td>6.5</td>
<td>n.r.</td>
<td>n.r.</td>
<td>1.6</td>
<td>9.2</td>
<td>8.8</td>
</tr>
<tr>
<td>Crude Fiber (%)</td>
<td>n.r.</td>
<td>n.r.</td>
<td>1-2</td>
<td>n.r.</td>
<td>n.r.</td>
<td>4.8</td>
<td>n.r.</td>
<td>6.6</td>
</tr>
<tr>
<td>Ash (%)</td>
<td>10.8</td>
<td>8.0</td>
<td>12.5&lt;sup&gt;1&lt;/sup&gt;</td>
<td>n.r.</td>
<td>n.r.</td>
<td>n.r.</td>
<td>36.0</td>
<td>4.4</td>
</tr>
</tbody>
</table>

<sup>a</sup>[60]  
<sup>b</sup>Oberon FMR, Inc.  
<sup>c</sup>[61], estimates only, amino acids reported as g AA/ 16 g N, converted by a factor of 6.25 / 16 * CP (Crude Protein), using values of CP from the same source  
<sup>d</sup>www.soymeal.com, all values averaged for 8 types of soymeal  
<sup>e</sup>[62]  
<sup>f</sup>[63]  
<sup>1</sup>Ash reported as mineral material  
<sup>2</sup>AA as averaged over 8 samples, with 14% moisture assumed  
<sup>3</sup>Crude values averaged over 17 samples  
    n.r., not reported
rainbow trout feed showed neutral effects up to 25% replacement [63]. A methanotrophic biomass SCP tested with salmon showed that beneficial growth occurred at up to 25% replacement of fish meal, with lower dietary amino acid levels cited as the main limiting reason for growth [64].

Specifically addressing BSCP, recent experimental trials have shown these protein alternatives can only replace a fraction of the total protein needed for nutritional requirements before health and growth are impacted. Shrimp grown on a BSCP-formulated feed with soybean meal replacement at 15.6% and on a BSCP-formulated feed with fish meal replacement at 7.8% showed improved growth over fish meal-based feed [65]. A survival study was done for a Sequencing Batch Reactor (SBR) BSCP product and a Membrane Bioreactor (MBR) BSCP product that found a somewhat lower (though not significantly lower) survival rate than the control fish meal diet. Interestingly enough, growth rates and harvest rates may actually increase with increasing BSCP inclusion (tested up to 30% replacement) [53]. BSCP fed to catfish had a positive effect on growth performance up to 14% replacement [46]. Preliminary data from our work on *Litopenaeus vannamei* (white shrimp) suggests that complete replacement of fish meal with BSCP (at 8% protein meal in total feed) was actually superior in terms of growth, survivability, and a feed conversion ratio [66]. However, the exact mechanism of toxicity or growth impairment is not well understood.

Aside from protein applications, algal and microalgal SCP in particular have been discussed as a probiotic and as a fish oil replacement. Algal products from organisms such as *Chlorella* are actively being pursued as biofuel alternatives, but extracted protein by-products from that process can contain high amounts of amino acids, as Table 2.2 suggests. Also, certain species of thraustochytrid algae are known to produce beneficial poly-unsaturated fatty acids and have been tested as fish oil replacements in salmon feeds with positive results [67]. Indeed, the major market for algae (actually *Spirulina*, a cyanobacteria) is currently as a probiotic supplement and an essential food source in aquaculture [68]. A serious need in the aquaculture industry is the ability to counter disastrous infections using such non-antibiotic techniques like probiotics, especially for
eggs and larvae [69]. Like algae, microbial products in general have been shown to contain beneficial bacterial nutrients, though BSCP has not yet been examined in depth as a probiotic.

2.8 BSCP in the Global Perspective

Given the need for sustainability as a driver for fish meal alternatives, we also describe BSCP, vis-à-vis fish meal, from a global nitrogen cycle approach that encompasses production and wastes in the aquaculture industry. In the environment, the major source of natural nitrogen for protein is from the atmosphere via biological nitrogen fixation (BNF) [70] (Figure 2.2). Phytoplankton and rhizospheric organisms have adapted to fix gaseous nitrogen into ammonia and then incorporate ammonia into organic nitrogen by the glutamate synthesis (termed the GS/GOGAT) pathway [71]. These organisms thus serve as primary nitrogen fixers and are then consumed by higher land and aquatic organisms and their fixed nitrogen incorporated into the food chain. Anthropogenic nitrogen used in industry is fixed from gaseous nitrogen by the Haber chemical process into ammonia or urea. The Haber process has fundamentally changed global agriculture as atmospheric nitrogen is directly converted into fertilizer for plants (or nutrient media for BSCP). Fixation of atmospheric nitrogen by the Haber process exceeded 152 million tonnes of ammonia in 2009 [72]. Nitrogen fixation can also be stimulated naturally through the process of intensive agriculture, termed agricultural BNF. Much of this fixed nitrogen ends up as wastes or runoff. In wastewater treatment facilities, nitrogen sources for the aerobic growth of treatment microbiota typically come from the wastewater itself, or from nitrogen originally derived from this agricultural nitrogen via human consumption.

In the marine environment, phytoplankton fix oceanic BNF or microbes scavenge free nitrogen compounds (typically ammonia) from the open ocean. More of this free ammonia now comes from agricultural runoff from rivers and streams and can highly impact the productivity of estuarine systems. Higher organisms consume the microbiota to obtain their own nitrogen and are in turn consumed. The benefits of BSCP will be from reducing the amount of rendered fish meal harvested from the oceans, at a cost of
Figure 2.2. Global nitrogen cycling with a focus on aquaculture (global flows and magnitudes derived from [72], [75]). Note, the magnitudes shown here are subject to high uncertainties. Other sources of nitrogen, such as lightning, energy generation, and burning are not shown.
increased use of artificially fixed nitrogen to grow this protein alternative. Overall, without some sort of nitrogen removal system such as a downstream denitrification treatment plant, the growth of aquaculture operations will additionally increase nitrogen pollution. Ultimately, while the use of alternative protein sources can increase available protein and decrease demand for fish meal, it is only one component to achieving better sustainable aquaculture practices [73], which in itself is but one component of integrated ocean resource management [74].

2.9 Future Prospects and Research

Currently, several companies are attempting to produce BSCP using several different technologies. Historically, the industry has looked quite favorably on the idea of wastewater-derived SCP materials, and as long as aquaculture continues to expand to meet the protein demand gap, there is no reason to believe that fish meal prices will drop [74]. As full-scale reactors for BSCP production come online, a more current estimate of investment and operational costs is needed. Additionally, the digestibility of protein alternatives is poorly understood and requires further research, particularly for factors currently preventing high inclusion in feed.

In the future, the feeding of fish should be considered as part of the global nitrogen cycle solution. Self-enclosed systems for aquaculture, so called Recirculating Aquaculture Systems (RAS) [75] or closed aquaponics systems [79-80] can potentially eliminate the vast majority of nitrogen wastes from waste streams and deliver a constant supply of clean water to aquaculture systems (though at a cost of installing treatment systems) [78]. Additionally, such strategies could allow for the establishment of inland cultivation facilities and decrease the cost and duration of shipping. As a further benefit, the wastes formed in RAS or a Biofloc Technology (BFT) process can be processed into a feed product or even reused directly as a feed source for shrimp [79].

The expansion of MBR technology can also be a potential boon to BSCP production as well. In wastewater treatment, the incorporation of MBRs for water recovery offers
improved separation efficiency, higher loading capacity, more flexible operations regime, and a smaller footprint over conventional sedimentation designs at a tradeoff of increased costs. Despite these costs, more stringent environmental regulations and improvements in power consumption and per unit cost have resulted in a 10-fold increase in the size of new treatment plants over the last 10 years and numerous new treatment works constructed or in development. The main research focus for membrane technology today is in the characterization of bio-fouling constraints of various types of filtration designs, though further improvements in their energy use are needed to offset any expected future energy price increases [83-84].

Another interesting area of research is the application of biomedical technologies and cheap computing analysis power to the environmental and agricultural sectors. The same class of technologies used to complete the human genome and a whole host of new “-omics”-related fields have been developed and targeted on microbial ecology and biotechnology. The synthesis of “-omics” (represented by such technologies as micro-arrays [82], genome [83] and meta-genome [84] sequencing, RNA gene sequencing [85], and protein peptide sequencing [18]) combined with high-power computing bioinformatics and theory-driven analysis, are finally beginning to illuminate the complex interactions found between environmental microbes in wastewater systems [86]. In particular, the field of proteomics involves the separation and quantification of proteins by gel electrophoresis and their identification by the sequencing of their peptides [19]. This can be adapted to study the proteins produced in BSCP to understand the exact blend of proteins being used to formulate aquaculture feeds. Proteomics tools have been used in the medical sciences for quite some time and are now being applied to environmental samples [87]. This kind of knowledge can then be used to identify microbial metabolic pathways that result in higher protein production within cells, essentially a form of “microbial bodybuilding.” The use of high-throughput sequencing techniques to track microbial turnover should be especially noted. Determining probiotic effects has been difficult [69] but the same ability to measure trends in microbial consortia change in human microbiomes [88] can be applied to quantify probiotic effects on microbial consortia in aquaculture.
2.10 Conclusion

At present, the field of SCP fish meal replacement shows great promise to help increase available protein for sustainable aquaculture but remains a nascent industry despite several decades of work. However, the demand for sustainability both by consumers and industry has infused new urgency and energy into this field, and has united experts from disparate fields such as wastewater engineering, microbial ecology, and aquaculture science. The application of new technologies and advanced DNA sequencing with subsequent bioinformatic interpretation will allow for the construction of more complete and high-protein containing fish meal replacements to change fundamentally what is possible for aquaculture. Such an achievement can be accomplished with application to ‘natural’ microbial communities in conventional process wastewater environments with no genetic or recombinant manipulation of genetic material. The future of recoverable bioproducts will be one of stiff competition between a wide array of technologies. But with high demand for alternatives to fish meal, a breakeven point has passed where decades-old ideas are now a feasible part of a sustainable future.
CHAPTER 3
METHODS FOR HIGH-THROUGHPUT SEQUENCING AND BIOINFORMATICS

3.1 Summary

Significant changes in sequencing technology and sequence processing have occurred in only a few short years and have not been fully described in print in many cases. This chapter is a current snapshot of that evolution and describes the development, testing, and implementation of high-throughput sequencing (pyrotag sequencing in particular) to the field of wastewater microbial ecology. First, current methods for the laboratory preparation of pyrotag sequencing samples are presented. Then, bioinformatics techniques of high-throughput sequencing are described in comparison with the sequence processing methods of Sanger sequencing. Next, an examination of ecological methods and multivariate analysis is presented; specifically how can alpha and beta diversity metrics be applied to high-throughput sequencing data. Finally, a method for the co-abundance of species over gradients is examined to understand the problem of identifying significantly co-existing species in samples. Overall, this chapter shows how the complex area of bioinformatics has grown in a short amount of time and highlights the effort taken in characterizing these new techniques for use in this dissertation.

3.2 Background of High-Throughput Sequencing

With the advent of DNA and RNA based molecular techniques, a new view of the world of microbial diversity became accessible. Tools such as clone library sequencing [89], Fluorescence In Situ Hybridization (FISH) [90], and Denaturing Gradient Gel Electrophoresis (DGGE) [91] allowed for more sensitive detection and study of the unculturable microbial consortia critical to the functioning of wastewater treatment. In particular, Polymerase Chain Reaction (PCR)-based techniques began to show the phylogenetic relationships between wastewater indicator or functional organisms that define system performance (such as Nocardia) [92] which before had been identified typically by morphological techniques. PCR-based techniques were also a tool for survey of the broad swath of undetectable microbial consortia of numerous wastewater sources and treatment regimes with 'universal' (i.e., amplify DNA from all three domains.
of life) PCR primers [93]. All of these studies showed that at the time, characterization of microbial ecology was an intractable problem with few tools that could come close to surveying even a fraction of microbes in most ecosystems.

For several decades, a series of technologies based on capillary electrophoresis terminal-dye DNA sequencing technology (termed Sanger sequencing) served as the mainstay of sequencing science. This technology was mainly responsible for the completion of the human genome in 2001 [94] and was used by microbial ecologists primarily to sequence 16S rRNA genes for the purpose of discovering novel diversity and for the identification of certain characteristic functional genes determined from the study of microbial pathways. However, with the advent of microelectronics in the early 2000’s, several companies formed to build a next-generation device. The first team to the finish line, 454 Life Science Corp. (now 454 Roche), sequenced their first proof of concept (a virome) in 2003 [95], and published the first paper in 2005 of the sequencing of a bacterial genome using this technique [96]. The 454 series of sequencing machines ushered in a new era in sequencing, the so-called “Next-Generation Sequencing” (NGS) era, where the possibility of sequencing a human genome for less than $1000 was now viable. Table 3.1 shows the current performance parameters of high-throughput sequencing technologies currently available or in development. Within a year of the first 454 publication, a paper was published for the first time using this technology to study microbial ecology [97]. In only the last 7 years, the field of microbial ecology has seen rapid transformation due to this technology, with human microbiome studies playing a leading role. Sanger sequencing has gradually been replaced in the lab for the survey of microbial diversity from novel environments, with the first 454 run at Mines completed in 2009 [98].

However, while the cost per base pair of these new technologies is highly attractive, the handling of this new kind of data has posed serious problems for the field, the first being the extremely large amounts of data generated per run. In a multi-well Sanger sequencing format, each well provided one read of about 650 base pairs. In contrast, NGS runs can provide on the order of millions of reads per run with >1000 megabases of
total DNA per run (Table 3.1). Several NGS systems operate by using a plate or slide composed of millions of microscopic wells or attachment sites. Each well or site contains its own sequencing reaction and is monitored with a high sensitivity camera or detection device able to read the sequence at each site in a parallel manner. Several review papers describe the workings of currently available NGS technologies [102-103]. However, by 2007 it became clear that while a lot of data was generated, this technology (and all subsequent technologies on market thus far) suffered from shorter amplicon sizes and higher per base error rates than Sanger sequencing [101]. NGS technology was primarily designed for genome sequencing where accuracy relied on consensus of repeated sequencing of the same reads over and over in a single genome.

Due to the vast amount of generated data and the elevated occurrence of errors, many new bioinformatics technologies had to be developed for microbial ecology studies. These data management ‘bioinformatics pipelines’ are a series of programs run consecutively to quality control and process this stream of information into analyzed data. Since this process is much more automated, hand-curation (as was often done for Sanger sequencing data) is rarely possible. Instead, statistical algorithms and database management systems are used to hold, process, and correlate sequences to real world conditions. This kind of analysis uses a number of interdisciplinary skills, such as computer programming, data management, biochemistry, ecology, and multivariate statistics. As analyses become more refined, a best practice is slowly emerging of how to overcome errors and biases, how to relate sequences to each other and the environment, and how to formulate testable hypotheses based on initial surveys into the field. This chapter is a survey of that journey so far in the Spear Lab working with the 454 GS FLX platform. Some techniques are not developed by the author, and are referenced as such, though the protocols might not be published.
<table>
<thead>
<tr>
<th>Instrument</th>
<th>Run time*</th>
<th>Millions of reads/run</th>
<th>Bases/readb</th>
<th>Yield Mb/run</th>
<th>Reagent cost/runc</th>
<th>Reagent cost/Mb</th>
<th>Minimum unit cost (% run)d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3730xl (capillary)</td>
<td>2 h</td>
<td>0.000096</td>
<td>650</td>
<td>0.06</td>
<td>$96</td>
<td>$1500</td>
<td>$6 (1%)</td>
</tr>
<tr>
<td>Ion Torrent - '314' chip</td>
<td>2 h</td>
<td>0.10</td>
<td>100</td>
<td>&gt;10</td>
<td>$500</td>
<td>&lt;$50</td>
<td>~$750 (100%)</td>
</tr>
<tr>
<td>454 GS Jr. Titanium</td>
<td>10 h</td>
<td>0.10</td>
<td>400</td>
<td>50</td>
<td>$1100</td>
<td>$22</td>
<td>$1500 (100%)</td>
</tr>
<tr>
<td>Starlight*</td>
<td>+</td>
<td>~0.01</td>
<td>&gt;1000</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>0.5-2 h</td>
<td>0.01</td>
<td>860-1100</td>
<td>5-10</td>
<td>$1100-900</td>
<td>$11-180</td>
<td>+</td>
</tr>
<tr>
<td>454 FLX Titanium</td>
<td>10 h</td>
<td>1</td>
<td>400</td>
<td>500</td>
<td>$6200</td>
<td>$12.4</td>
<td>$2000 (10%)</td>
</tr>
<tr>
<td>454 FLX*</td>
<td>18-20 h</td>
<td>1</td>
<td>700</td>
<td>900</td>
<td>$6200</td>
<td>$7</td>
<td>$2000 (10%)</td>
</tr>
<tr>
<td>Ion Torrent - '316' chip*</td>
<td>2 h</td>
<td>1</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>$750</td>
<td>&lt;$7.5</td>
<td>~$1000 (100%)</td>
</tr>
<tr>
<td>Helicos*</td>
<td>N/A</td>
<td>800</td>
<td>35</td>
<td>28 000</td>
<td>N/A</td>
<td>NA</td>
<td>$1100 (2%)</td>
</tr>
<tr>
<td>Ion Torrent - '318' chip*</td>
<td>2 h</td>
<td>4-8</td>
<td>&gt;1000</td>
<td>&gt;1000</td>
<td>~$925</td>
<td>~$0.93</td>
<td>~$1200 (100%)</td>
</tr>
<tr>
<td>Illumina MiSeq*</td>
<td>26 h</td>
<td>3.4</td>
<td>150 + 150</td>
<td>1020</td>
<td>$750</td>
<td>$0.74</td>
<td>~$1000 (100%)</td>
</tr>
<tr>
<td>Illumina iScanSQ</td>
<td>8 days</td>
<td>250</td>
<td>100 + 100</td>
<td>50 000</td>
<td>$10 220</td>
<td>$0.20</td>
<td>$3000 (14%)</td>
</tr>
<tr>
<td>Illumina GAIIx</td>
<td>14 days</td>
<td>320</td>
<td>150 + 150</td>
<td>96 000</td>
<td>$11 524</td>
<td>$0.12</td>
<td>$3200 (14%)</td>
</tr>
<tr>
<td>SOLID-4</td>
<td>12 days</td>
<td>&gt;840c</td>
<td>50 + 35</td>
<td>71 400</td>
<td>$8128</td>
<td>&lt;$0.11</td>
<td>$2500 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 1000</td>
<td>8 days</td>
<td>500</td>
<td>100 + 100</td>
<td>100 000</td>
<td>$10 220</td>
<td>$0.10</td>
<td>$3000 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 2000</td>
<td>8 days</td>
<td>1000</td>
<td>100 + 100</td>
<td>200 000</td>
<td>$20 120</td>
<td>$0.10</td>
<td>$3000 (6%)</td>
</tr>
<tr>
<td>SOLID-5500 (P1)*</td>
<td>8 days</td>
<td>&gt;700f</td>
<td>75 + 35</td>
<td>77 000</td>
<td>$6101</td>
<td>&lt;$0.08</td>
<td>$2000 (12%)</td>
</tr>
<tr>
<td>SOLID-5500d (450)*</td>
<td>8 days</td>
<td>&gt;1410f</td>
<td>75 + 35</td>
<td>155 100</td>
<td>$10 503</td>
<td>&lt;$0.07</td>
<td>$2000 (12%)</td>
</tr>
<tr>
<td>Illumina HiSeq 2000 - v3*</td>
<td>10 days</td>
<td>≤3000</td>
<td>100 + 100</td>
<td>≤600 000</td>
<td>$23 470</td>
<td>≥$0.04</td>
<td>~$3500 (6%)</td>
</tr>
</tbody>
</table>

*a* Instrument time for maximum read length.

*b* Average length for high-quality reads >200 bases (mode is higher); typical maximum for reads ≤150 bases (most reads reach this length).

*c* Includes all stages of sample preparation for a single sample (i.e. library preparation through sequencing; capillary = sequencing only).

*d* Typical full cost (i.e. including labour, service contract, etc.) of the smallest generally available unit of purchase at an academic core laboratory provider for the longest available read (and percentage of reads relative to a full run, rounded to the nearest whole percentage).

*e* Upgrade of the FLX instrument, due out summer 2011.

*f* Instruments and reagents are no longer sold; services are available for any organism.

*g* Mappable reads [number of raw high-quality reads (as reported for all other platforms)] is higher.

*h* More reads are obtained than is needed from any single sample within most experiments, but the value illustrates the costs.

*i* Announced TruSeq v3 reagents & software, reads and yield are half for HiSeq1000.

*j* Information based on company sources alone (independent data not yet available).

+k* Data not yet available.

+/- Indicates a likely value based on unpublished information available in March 2011 (i.e. author speculation).
3.3 Differences between Sanger Sequencing and Pyrosequencing

For Sanger sequencing the standard practice is to prepare 96-well or 384-well format plates where each well contains a single dideoxynucleotide (ddNTP)-terminator amplified PCR product for capillary (Sanger) sequencing [102]. In the beginning, the DNA is extracted, a PCR is completed on a specific portion of a gene, and the PCR product is purified by gel electrophoresis spin column filtration. The PCR must be designed with an extended elongation time for a poly-adenosine tail to be attached to the end of every amplicon. This amplified product is still a mixture of genes from different organisms and must be separated by clone selection. PCR amplicons are inserted into a plasmid with a thymine overhang and taken up by electro-competent *E. coli* cells by electroporation. Each plasmid confers antibiotic resistance to the cells, and successful clones are screened with antibiotic plate cultures. These clones are picked and grown in a multi-well format, each well now containing an individual cloned gene from the PCR. Several steps are completed to lyse, amplify, and purify the DNA from the clones. The final reaction is a dye-terminator PCR. This is a standard PCR of the product, with a minute amount of ddNTPs attached with fluorescent tags. Each unique base (A,T,G,C) has a different fluorescent labeled tag. When randomly incorporated into the elongation of a sequence, these ddNTPs will suspend replication at that base pair and leave a shorter sequence. The result is a PCR product that is comprised of strands of DNA of increasing sequential length, with each end base terminated by one of four fluorescent tags representing the identity of that base. This product is stabilized and loaded to a sequencer, which uses capillary electrophoresis to elute the shortest sequence first, and reads the color of the end fluorescent tag, followed by each sequential tag until the sequence is completely read, in a multiple-channel format. A detailed description of this process is described in standard texts such as Lehninger's Principles of Biochemistry or Brock Biology of Microorganisms [106-107].

The process of pyrosequencing relies on a set of fundamentally different highly-parallel sequencing principles. A 454 pyrosequencing machine is able to multiplex every
Figure 1 | Sample preparation. a, Genomic DNA is isolated, fragmentated, ligated to adapters and separated into single strands (top left). Fragments are bound to beads under conditions that favour one fragment per bead, the beads are captured in the droplets of a PCR-reaction-mixture-in-oil emulsion and PCR amplification occurs within each droplet, resulting in beads each carrying ten million copies of a unique DNA template (top right). The emulsion is broken, the DNA strands are denatured, and beads carrying single-stranded DNA clones are deposited into wells of a fibre-optic slide (bottom right). Smaller beads carrying immobilized enzymes required for pyrophosphate sequencing are deposited into each well (bottom left). b, Microscope photograph of emulsion showing droplets containing a bead and empty droplets. The thin arrow points to a 28-μm bead; the thick arrow points to an approximately 100-μm droplet. c, Scanning electron micrograph of a portion of a fibre-optic slide, showing fibre-optic cladding and wells before bead deposition.

Figure 3.1 The functioning of 454 pyrotag sequencing, reproduced from [96]

single sequence it generates at the same time. At the heart of this 454 technology are two technologies, emulsion PCR (emPCR) and picoliter well sequencing (Figure 3.1). In emPCR, a single DNA fragment is amplified by PCR onto a bead suspended in a water bubble in an oil immersion in a manner such that only one template and one bead are in
each bubble. DNA fragments to be sequenced are then amplified with specific 454 primers for bead attachment. The result is these beads contain only a single DNA strand, and the DNA is amplified at multiple sites across the bead. The next new technology is the use of glass slides that contain several hundred thousand picoliter wells. These slides are made from a bundle of fiber optic cable that is cut cross-sectionally into a thin plate shape, and sealed on the bottom end of the slide. The microscopic beads, each now covered with identical DNA amplicons, are placed into each well. Washes containing a single type of dNTP are passed over the slide and if incorporation occurs a natural product of incorporation, pyrophosphate, is transformed into ATP by ATP sulfurylase and is consumed by luciferase to produce a fluorescent emission. A camera captures each successive wash of different dNTPs (e.g. A, then T, then G, C, A, T, G, C, etc.). The order of incorporation in the camera image for each well determines a single unique sequence. When this is done for the entire plate, hundreds of thousands of sequences are determined at once.

3.4 Preparation of Samples for PyroPCR

As pyrosequencing requires no amplicon separation cloning is completely eliminated from the process and only a single PCR (a pyroPCR) is required to prepare amplicons for sequencing for phylogeny studies. However, this pyroPCR is modified with unique primer regions and shorter amplicons. The shorter amplicons in particular are a major disadvantage of pyrosequencing technology. This is improving as new equipment is developed, but new primers must be devised and tested for the 400 bp maximum length (Table 3.1) and the selection of a proper region is critical [108-109]. Also, spanning regions with multiple sequences and realigning them is not possible so full-length sequences cannot be constructed from this type of data. However, the advent of pyrosequencing tag technology has allowed a single run to span multiple samples and PCRs. The major task for sequencing preparation is then the construction of lengthened primers with barcodes, validation of sample PCR, and then normalization of PCR product across samples so that the sequences returned are equally distributed across samples.
454 Barcodes and Primers

5'-GCCTTGCCAGCCCGCTCAG TC AGAGTTTGATCCTGGCTCAG-3'
  
454 B Primer   Linker   8F Primer
  
5'-GCCTCCCTCGCCCAATCAG AACCAACC CA TGCTGCCTCCCGTAGGAGT-3''
  
454 A Primer   8-Letter Unique Barcode   338R Primer

Final Amplicon Sequence

A  Barcode Linker  + 338R  16S  + 8F  B

Figure 3.2. Primer composition for 454 GS FLX Platform. 454 Roche primers designed to interact in emPCR are shown in blue. A unique barcode identifying the source sample (one per sample) is in orange, and the conventional primers shown in red for the 16S SSU rRNA gene.

To interface with the emPCR primers, 454 Roche has set the forward and reverse end primers with their A and B primer set. The primers for the GS FLX platform are shown below (Figure 3.2). The A primer is usually the sequence initiation primer (though support for B primer sequencing now exists), and is followed by the barcode and then a user determined primer. Two linker primers with minimal known homology to known sequences are added between the user-specified primer and the 454 specified primers in to reduce binding efficiency problems with the 454 primers. The specific example setup shown below and the 8 digit error-correcting barcodes were designed by Hamady and colleagues [10] and first used in the Spear lab by Sahl and colleagues [98]. This design is used for the study in the next chapter.

New barcoded primers must be individually tested before PCR, usually by monitoring a qPCR reaction in real time to verify product and was first used in the Spear Lab by
Jason Sahl. This method also allows for the monitoring of PCR bias by cutting off the program at a peak cycle number where amplification rapidly drops off.

Previous work has noted the difficulty of accurately normalizing minute amounts of amplicon sequence for pooling [107]. Manufacture instructions from Roche recommend quantifying the DNA by picoGreen or another type of fluorescent quantification. With the advent of Agilent microfluidics technology, DNA quantification by Agilent Bioanalyzer has come into its own. The Bioanalyzer couples microfluidic gel electrophoresis with fluorescent quantification of DNA to produce electropherograms that can accurately quantify DNA using small quantities of sample. In Figure 3.3A, this method has been shown to produce tighter clustering in histograms of sequences per sample recovered over multiple 454 runs over conventional Invitrogen DNA normalization plates used previously in the lab. However, it must be cautioned that differences in PCR efficiencies and sample constituents (such as contaminants) will inevitably cause differences in sequences per sample returned. The next step is to estimate the desired sequencing depth per sample. As pyrosequencing returns a consistent number of sequences per run, the number of samples to be included in a run directly determines the number of sequences per sample. Rather than selecting a mean sequences per sample value based on dividing expected yield by the number of samples, the sequences per sample value should be determined by a minimum cutoff of 1-2 sigma (sigma= 400 sequences in Figure 3.3B) below the desired amount to achieve a minimum sequence count per sample.

3.5 Bioinformatics

The information generated by 454 pyrosequencing is encoded into proprietary flowgram files that are typically then decoded to derive base calls and quality scores. Since Sogin et al. (2006), a number of performance and analysis changes have occurred and there is no doubt that further enhancements will take place.
Figure 3.3. Results of sequence distribution across samples for two runs. The top run (A) shows the difference between Bioanalyzer normalization and Invitrogen SeQual Prep plate normalization. The bottom (B) shows a run with only Bioanalyzer normalization.
3.5.1 Flowgram Clustering and Denoising

A major factor affecting quality of base calls has been the occurrence of homopolymers in amplified DNA sequences. Since a sequence of homopolymers are incorporated by the same wash step, the maximum fluorescence intensity is directly proportional to the expected number of homopolymers. Unfortunately, homopolymers can elicit a wide range of fluorescence intensities, making determining the exact homopolymer length difficult in flowgram data. This is one of the major problems that plagued the first studies in this field, and it led to potential over estimation of total (and in particular rare) organisms [108]. Early analysis rule of thumb was that it was often better to cluster sequence data [109] or discard singletons to try to mask some of the error and eliminate others through filtering out ambiguous bases, excessively long or short sequences, and any sequence with high homopolymer count [110]. Recently, flowgram clustering analysis was proposed as a way to remove noise from the data. Since then the term has been referred to as ‘denoising’ with several current variations of denoising algorithms available to use [111–113]. The idea behind this concept was that many identical sequences would be produced in a number of wells. Flowgrams that fit the same homopolymer distributions could be clustered and a consensus sequence for the cluster could be determined based on a characteristic distribution of homopolymer intensities. This has been used to greatly reduce the observed sequence variation. Of course, because of the consensus nature of this process, outlier singletons still pass through the analysis whether or not they are error prone.

3.5.2 Chimera Checking

Another important issue with PCR-based techniques is chimera formation. Chimera formation is not entirely understood but hypothesized to be mis-amplified PCR products that arise when one replicating strand is interrupted. This partially replicated product itself then becomes the primer of a new chain reaction, either wholly or partially annealing to a different but closely related strand of DNA [114–116]. The result is a hybrid amplicon with one end resembling one parent sequence, and the other resembling a second parent sequence called a bi-mera. Due to the homologous nature of the 16S SSU rRNA gene, interrupted amplicons have a higher chance of re-annealing to a similar
complementary DNA strand, so significant portions of a sequencing run can be chimeric [117]. Chimera detection techniques revolve around two ideas. The first is that the chimera is most likely to be split at a point where significant variation in gene sequence exists. In PCR terms, a strand of partially amplified DNA will tend to anneal to closely related segments of a nearby DNA template and resume replication at that point. Thus, chimeras should tend to be hybrids of two (or more) similar parent sequences with a notable breakeven point demarking the interruption of replication. Several programs try to detect positional variations in sequence, mutation rates, or phylogeny by using sequences within the same run as hypothetical parents, or parents from a database (Pintail, Mallard, Bellerophon, Chimera check) [118–120]. Two obvious problems emerge. The first is that database coverage can influence chimera detection; the more sequences in the database the more potential sequences that can serve as parents. The second potential problem is that a sequence hybrid can potentially still be novel diversity that would be removed from analysis as a false positive chimera. The second idea for detecting chimeras is that they are thermodynamically minor products compared to fully replicated sequences and should be amplified less favorably. A pitfall for this strategy is that sequences that resemble other sequences in the same PCR but are at lower abundances can also be viewed erroneously as chimeric [121]. Additionally, chimera reduction strategies include increased elongation times, changes in polymerase type, and reduced cycle number during PCR [125-126].

Detecting chimeras in Sanger sequencing data was already a computationally and manpower intensive process. These programs would be impractical for high-throughput sequencing data sets. Therefore, several groups have used the two ideas stated previously to develop new automated rapid tools for chimera detection in high-throughput runs. One of the first was Chimera Slayer [124], which utilizes a reference parent matching strategy and an evolutionary modeling strategy. An early re-interpretation of this program was implemented in mothur [125]. The 8F-338R sequence set used in the study of the next chapter was taken for analysis and all breakeven points were numbered to E. coli 16S basepair positions and then displayed graphically with 16S rRNA secondary structure of an E. coli reference (Figure 3.4). Numerous breakeven points were observed at the
beginning of the second variable region, and a lack of breakeven points was observed in variable region arms of the 16S SSU rRNA gene. Two possible interpretations are suggested by this data. The first is that while interruption of PCR can happen at any point in replication, only interruptions that happen at the end of a conserved region are most likely to re-anneal onto similar DNA strands and to form chimeras. The second is similar; interrupted products that are interrupted in a conserved region of DNA are more likely to re-anneal onto similar DNA strands to form chimeras, but since the region is conserved, the breakeven point is not detectable by software until the replication proceeds to a more variable portion of the gene. Both of these interpretations are directional. More interruptions occurring at transitions from conserved to variable regions will form chimeras, and thus more breakeven points will be detected at those transition points. This interpretation also suggests that low breakeven regions before a high breakeven region are conserved zones, and low breakeven regions after a high breakeven region are variable zones. This interpretation requires further testing, but would require a very specific sequencing dataset. The dataset tested in this work comprises only a short fraction of the 16S gene. Full-length sequences arising from the same PCR would be needed to test other 16S variable regions to see how well the interpretation holds. Also, since this dataset was produced by a version of mothur that had too many positive detections as reported by the mothur developers, a number of these must be false positives. In order to build a dataset large enough for statistical power, a very large sequencing dataset must be used to produce a similar amount of information from accurate chimera detection algorithms.

The next step was to validate several different chimera-checking programs and see how much overlap existed. This would inform about potential false positive rates of various methods. This is shown below in Table 3.2 for the dataset used in the next chapter. ChimeraSlayer and UChime [121] are both programs that detect chimeras based on the first strategy of comparing amplicons to parent sequences and detected chimeras in more abundant OTUs (Operational Taxonomic Units); interestingly, they show little overlap with each other. UChime in de novo mode had the most detected chimeras, with most in minor abundance. Overall, the results from this analysis show that chimera
Figure 3.4 Chimera break even points along V1 and V2 regions of the 16S gene using mothur implementation of Chimera Slayer. Colors represent 5-base moving averages of break even points detected in chimeric sequences using the pyrotag dataset of the next chapter. V1 and V2 designations are from Baker et al. 2003 [126]. V2.1 and V2.2 designations are from Reau et al. 1988 [127] and are listed as V2 and V3 in that publication, but V2.1 and V2.2 here to preserve the traditional 9 variable region designations. The 16S secondary structure is derived from E. coli and determined from http://www.rna.ccbb.utexas.edu/. Black squares represent unanalyzed bases.

detection rarely identifies the same set of sequences as chimeric. The silver lining of this is that most chimeric sequences are low-abundance sequences and should not factor significantly in large-scale trends of the microbial community as long as no conclusions are made of any one sequence. UChime in de novo mode also compares the abundance of amplicons with the abundance of potential chimera parents and selects chimeras
Table 3.2. Chimeras detected using three different detection algorithms. OTUs detected as chimeras are listed in each row sorted by abundance. The number of sequences detected for each OTU and the fraction of these sequences in the entire study are shown, followed by the algorithm that detected a chimera and the amount of chimera-detection overlap between algorithms.

<table>
<thead>
<tr>
<th>OTU ID</th>
<th>Fraction of Sequences</th>
<th>Sequences</th>
<th>Chimera Slayer</th>
<th>UChime</th>
<th>UChime De Novo</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1151</td>
<td>0.0462892</td>
<td>2604</td>
<td>X</td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>909</td>
<td>0.0094747</td>
<td>533</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>964</td>
<td>0.0024176</td>
<td>136</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1334</td>
<td>0.0009244</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>349</td>
<td>0.00063933</td>
<td>39</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>712</td>
<td>0.0003555</td>
<td>20</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>880</td>
<td>0.0003355</td>
<td>20</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1277</td>
<td>0.0002844</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>581</td>
<td>0.0002666</td>
<td>15</td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>74</td>
<td>0.0002311</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>376</td>
<td>0.0002133</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>438</td>
<td>0.0001955</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1022</td>
<td>0.0001955</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>772</td>
<td>0.0001778</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>366</td>
<td>0.0001600</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>173</td>
<td>0.0001422</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>739</td>
<td>0.0001067</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>160</td>
<td>0.0001067</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>526</td>
<td>0.0001067</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>51</td>
<td>8.8881E-05</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>8.8881E-05</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1087</td>
<td>7.11048E-05</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1227</td>
<td>7.11048E-05</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>425</td>
<td>7.11048E-05</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>217</td>
<td>5.33286E-05</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1142</td>
<td>5.33286E-05</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>974</td>
<td>5.33286E-05</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>325</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>503</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>825</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>388</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>458</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>542</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>678</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>939</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>953</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1144</td>
<td>3.55524E-05</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>115</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>522</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>690</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>733</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>826</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>840</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1152</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>95</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>359</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>363</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>492</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>556</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>601</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>605</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>631</td>
<td>1.77762E-05</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

45
Table 3.2 Continued

<table>
<thead>
<tr>
<th>OTU ID</th>
<th>Fraction of Sequences</th>
<th>Sequences</th>
<th>Chimera Slayer</th>
<th>UChime</th>
<th>UChime De Novo</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>633</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>654</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>703</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>913</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>956</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>963</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1019</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1024</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1164</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1230</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>1240</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>242</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1078</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1167</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>1261</td>
<td>1.77762E-05</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>1</td>
</tr>
</tbody>
</table>

that tend to be lower abundance hybrids of common sequences in a dataset. Based on this information, and new research from Schloss et al. [128], and input from Charles Pepe-Ranney, UChime in de novo mode was selected for chimera detection in the study presented in the next chapter.

3.5.3 Sequence Alignment, Clustering, Classifications, and Tree-building

Selection of alignment algorithms is a complicated subject that is a prerequisite activity to a number of follow-up analyses. The primary purpose of alignment is to produce a set of data that can be interpreted to measure similarity between sequences (oftentimes as a distance matrix) and is not necessarily a standalone activity. The classification of sequences, the clustering of sequences into OTUs, and the construction of phylogenetic trees can integrate alignment algorithms as well. The two main factors in alignment systems are the algorithms used and the databases used to train alignment datasets (where necessary).

It is easy to think of these technologies as static, but processing datasets increasing sizes in Sanger sequencing runs was already beginning to drive development of informal bioinformatics pipelines. With Sanger sequencing, after alignment hand-curating for misaligned regions was quite common, but this became a herculean task as data increased. Sanger sequencing data was once processed with several different tools, like a multiple sequence alignment program such as MAFFT [129], CLUSTAL [130], or

46
INFERNAL [131]. These programs rely on several strategies for alignment, such as naïve bayesian statistical techniques, base pair matching and scoring techniques, and secondary structure information to produce quality alignments, particularly for tree construction. As database technology progressed, large annotated databases came into use. By using sequences from large databases such as SILVA [132] or greengenes [133] as the basis for alignments and for classification, better alignments due to larger database comparisons, and better annotation of reference sequences due to human curation of databases resulted in higher quality results. Using aligned sequences to make taxonomic assignments was also shown to be superior to older assignment techniques like BLAST [134], [135]. Additionally, web-based features such as in greengenes allowed for remote processing of thousands of sequences at once. For tree building, the use of alignment masks has always been an important feature to remove extraneous variation in datasets to prevent poor branch assignment in phylogenetic trees. Curated databases often included curated alignments with known variable regions of the 16S SSU rRNA gene masked out. LaneMask [136] and LaneMaskPH were two masks applied from these reference alignments for better tree-building, but new database generated averaged masks such as the pos_var_bac mask in the SILVA database include numerous levels of variability masking. Tree-building itself is a well-studied discipline with many possible features to select from. Sanger sequencing tree-construction often employed maximum likelihood, neighbor-joining, and parsimony methods and was the most computationally intensive portion of sequence processing. Lastly, as sequencing studies included more and more sequencing depth, organization of sequences into clustered groups became common, most notably with the DOTUR group of clustering algorithms. Clustering science itself is another wide-ranging statistical field, so several references are suggested but not reviewed here [137–139]. In order to develop microbial profiles of relative abundances for downstream analysis, each sequence must be clustered into a group, typically at a 97% sequence similarity level to represent a ‘species’ (95% genus, 90% family, 80% order). An OTU is essentially a group of sequences clustered at a certain similarity level, though a major challenge for the field is that these arbitrary thresholds and computational definitions may or may not have real life significance [140]. Much work has been done to examine the validity of clustering strategies in microbial ecology and the current
thinking is that abundance sorting and single-linkage pre-clustering with average-linkage clustering accurately represents microbial groups [105].

As the size of datasets grows, alignment tools, classification, and tree-building methods are limited by what is computationally feasible and by what is available for high-throughput data. With pyrosequencing sequences originally as short as 150 base pairs (bp) for first generation machines, there was serious discussion as to how reliable were sequence taxonomic classifications. One notable first attempt was GAST [141], a cluster-computing based consensus strategy modified BLAST algorithm. At the present time, NAST-based aligners (such as in mothur [142], greengenes [143], and PyNAST [144]) have shown the efficiency needed to align large amounts of sequences to increasingly larger databases. The mothur implementation of the NAST aligner was used for the study in the next chapter. The RDP classifier, a naïve Bayesian classifier, has been available as a web-based application running a customized database based on the Bergey’s Manual of Systematic Microbiology [145]. This database has been shown to be overly conservative, but the Bayesian classifier has been shown to be highly accurate and fast enough for high-throughput data [135]. Therefore, the mothur implementation [125] of the classifier trained on a customized curated non-redundant SILVA database was used in the study of the next chapter. Since functional information of microbe lifestyle can only be obtained with low-level classifications, a validation study using two tests of low-level classifications were examined. First, 10% of the reference SILVA 104 non-redundant dataset was randomly sampled and used as a test case and the number of taxa and taxon ranks recovered was determined. This served as a maximum bound of possible recoveries. Next, 10% of the greengenes dataset was randomly resampled and matched with their equivalent SILVA-based taxonomy and the number of matched taxa and taxon ranks recovered. The main difference between the two is that the greengenes database is generated from the greengenes NAST-based aligner and classifier. This served to examine how well the classification in this dissertation compared to the greengenes NAST-based classifier. This information is shown in Figure 3.5. The number of taxonomic ranks of reference database classifications in both subsampled reference datasets rarely went below 6 levels, the rough equivalent of the genus level.
Figure 3.5. Taxonomic recovery of the customized RDP Bayesian classifier trained on a customized Silva 104 NR dataset. Total fraction of sequences is shown on the left axis, and the fraction of classifications present at a taxonomic level is colored, where present. Classifications of the 10% Silva subsample recovered are in black hatch and not recovered are in black. Classifications of the 10% Greengenes subsample recovered are in grey hatch, and not recovered are in grey.

Additionally, recoveries at the genus level of the test dataset were high: nearly 80% of the entire dataset was recoverable, and only 5% of the dataset could not be recovered for the SILVA reference database. As this is the best case, the greengenes subsampled case shows poorer performance, with about 18% of genus level classifications not in agreement with the classification system tested.

Lastly, tree-building, one of the most computationally intensive activities in phylogenetic studies, has been a major bioinformatics bottleneck. FastTree [146] is an “approximate” maximum-likelihood program using a heuristic neighbor-joining algorithm for tree construction that has been shown to be orders of magnitude faster than other algorithms and as accurate to other maximum-likelihood based tree-building
algorithms. Therefore, FastTree was selected for tree construction in the study of the next chapter.

3.6 Microbial Ecology

A major change in sampling design has developed due to the use of pyrotag technology [10] to place numerous samples onto a single sequencing run. At this time, experiment designers can sequence nearly as much as can be sampled, making quantitative summaries of samples and metric-based comparisons between samples increasingly important in sequencing-based studies. This is an area where microbial ecology has finally caught up with macro-ecology and indeed, may supersede it in the coming years in terms of sampling coverage. The concepts of alpha, beta, and gamma diversity were first put forth by Whittaker for the study of plant variation along an elevation gradient in the mountains of the Pacific Northwest [147–149]. Alpha diversity refers to the variation of organisms within a sample site (or collection) and, for Whittaker, was based on the work of Fisher to evaluate biodiversity at a site [148]. The gamma diversity refers to the variation of organisms across an area such as a gradient or patch. Whittaker related the two with a term he called beta diversity with the relation $\beta = \frac{\gamma}{\alpha}$. This relationship was a measure of variation of species number along gradients, but most recent work utilizing beta diversity involves a subset of this, the measurement of ecological distances between two samples [150], something Whittaker had dismissed as being unfeasible due to the under-sampling common at the time [148]. Whittaker also noticed that beta diversity along a gradient will tend to fall off in a logarithmic fashion, now known to be a linear decay relationship with modern distance metrics [148], [151]. Current thinking in the design of microbial ecology experiments involves applying such macro-ecological principles such as gradient and transect analysis to microbial environments to determine the main drivers of selection on microbial ecologies in natural environments in a quantitative manner [152]. Gradient studies seek not to answer the question “Are two samples different?” (as in a Mantel Test) but rather “How different are a series of samples and how do these differences relate to an external gradient (e.g. latitude or elevation)?” This coupled with the application of multivariate statistics can identify
major patterns in metabolism and nutrient cycling in environments by quantitatively tying one to the other.

### 3.6.1 Alpha Diversity and Evenness Metrics

The array of metrics used to examine the total estimated diversity and the evenness of a sample are truly numerous and derive from a number of different applications and theoretical bases. A number of different reviews summarize these metrics and their calculation [153–155]. A few of the most commonly used metrics and the way in which they are used are summarized below. The foundational ecological program, EstimateS, is often cited as the source of a number of these metrics by both mothur and QIIME and the manuals of all three packages are an excellent way to understand the implementation history of the metrics used in this type of work [88], [125], [156].

**Estimators (estimates of total species number)**

Chao [159-160] – Non-parametric estimates of the total number of species based on distributions of singletons and doubletons. When applied to a single sample the variant estimator Chao-1 is used.

Good’s Coverage [156] – The fraction of OTUs that are singletons. This is often listed as an estimate of sequencing depth.

ACE [156] – Abundance Coverage Estimator. Estimates total number of species based on the variation and abundance of rare organisms in the dataset.

CatchAll [158] – A newer estimator developed based on parametric finite mixture models that combine several classes of distributions to calculate a best fit rare abundance profile. This metric was specifically designed for the high-throughput datasets of microbial phylogeny studies.

**Observed Species** – The exact number of species observed.

**Indices (values that represent aggregated properties of diversity)**

Fisher’s alpha [153]– An early metric based on curve-fitting functions of logarithmic distribution but not commonly used now due to sample size bias and lack of universality.
Shannon [153]—Also known as the Shannon-Wiener and is one of the most extensively used indices. Increasing species or increasing evenness will relate with a higher index.

Simpson [153]—Measure of the number of times it would take to select two organisms from a sampling such that they both will belong to the same group. A lower value correlates with higher diversity. Often taken as either the inverse, or the negation from one (Gini-Simpson).

Phylogenetic Distance (PD) [159]—A recently derived phylogenetic based metric defined as the sum of all branch lengths in a tree of a sample or area. This index is the measure of the phylogenetic diversity in a sample and takes into account closely related species.

Rarefaction (subsampling of a sample)

Rarefaction and jackknifing are not metrics per se but rather subsampling techniques used to gauge sampling depth. The concept underlying these techniques are that subsamples of an original sample can be used to examine the incremental diversity loss of each sequence as it is randomly removed from a sample. If every sequence is unique a 1:1 ratio of sequence to OTUs would exist. Though subjective, any suitable deviation from 1:1 to a flat curve would indicate good coverage. A plot of the rarefaction curves of the study in the next chapter, colored based on environment type (Figure 3.6) is shown below. These types of curves can be used to compare sampling efforts of various samples in a relative manner.

3.6.2 Beta Diversity

As with alpha diversity, numerous metrics exist for beta diversity. Historically, the field of information theory has heavily influenced ecological measurement. As high-throughput sequencing has become increasingly common a hotly debated area of research involves trying to understand which beta diversity metrics are the most appropriate for analyzing microbial ecology, particularly along environmental gradients such as pH or
Figure 3.6. Rarefaction curves of sampling effort in the study of the next chapter with environment types clustered based on the sampling legend in Figure 4.1.

distance. Not enough is yet known on how microbial datasets respond to metric-based testing, but several guiding factors can suggest selection of a proper metric for testing.

Microbial datasets often exhibit low sharing and a high degree of uniqueness. If comparisons to external factors such as environmental conditions using multivariate or principal component techniques are desired, a multitude of choices exist. Different metrics represent different comparisons of community membership, community abundance, and phylogenetic relationships. This means careful selection of the right metric is required depending on the type of problem being studied. Several resources exist on this subject and should be consulted during experiment design [137], [139], [150], [154], [160]. A number of concepts that need to be satisfied for a suitable method (triangle inequality axiom, double-zero sensitivity, normalization of factors, ecological basis) as well as some of the metrics considered for this study are described in the text below. Note that a named coefficient, distance, and index may have different meanings.
with the same name, depending on the original application and the derivation from source literature. The same metric can also have different forms depending on if qualitative or quantitative data is used.

Qualitative metrics

These are metrics that consider only community membership, also known as binary, since OTU presence or absence is usually denoted by a 1 or 0 respectively.

Sørensen Similarity Coefficient (Sorensen Index) [137] – Developed from concepts proposed by botanist Thorvald Sørensen, when qualitative data is used for two samples, simplifies to:

\[
S_1 = \frac{2C}{2C + A + B}
\]  (3.1)

where C is the number of shared species, and A and B are the number of unique species in sample 1 and 2 respectively. This is a similarity (S) metric, where increasing values denote increasing similarity (from 0 to 1 generally). It is also called Dice’s coefficient or the complement of the Whittaker’s species turnover index. A quantitative form also exists which is equivalent with the Bray-Curtis metric.

Jaccard [137] – another basic metric that is highly similar to the Sørensen Index and gives equal weighting to sharing and uniqueness with the formula:

\[
S_2 = \frac{C}{C + A + B}
\]  (3.2)

This particular metric however is not typically the metric of choice for diversity computations [163-164]

Quantitative metrics

These are metrics that consider relative abundance of each member, also known as distance metrics.

Euclidian [137] – The generalized multi-dimensional distance based on Pythagoras’ formula. This metric does not satisfy the triangle inequality axiom, that two distances between a point and two other points must be equal or greater than the distance between
the two endpoints (See [137]). It is only included here for historical understanding as a differential distance equation. This is a difference (D) metric, where increasing values denote increasing dissimilarity (from 0 to 1 generally).

\[ D_1 = \sqrt{\sum_{i=1}^{n} (y_{i1} - y_{i2})^2} \] (3.3)

where \( y_{i1} \) is the number of members in the ith species of sample 1 and \( y_{i2} \) is the number of members in the ith species of sample 2, and \( n \) is the total number of species in the entire data set.

Manhattan [137], [162] – Also known as the taxicab or city-block distance, for the shortest distance a taxicab must drive between two points in a city block, and represents the distance between two points in a grid fashion. This serves as the basis for a number of other metrics used, including the Whittaker’s Index of Association, the Bray-Curtis, and the Canberra Metrics. This metric also does not satisfy the triangle inequality axiom.

\[ D_2 = \sum_{i=1}^{n} |y_{i1} - y_{i2}| \] (3.4)

Bray-Curtis [137], [162] – One of the most often used metrics for difference measurement and represents the percentage difference between two samples. This metric does not weight shared zeros. In ecological data, the vast majority of the species examined will not exist at a site, leading to a double zero listing in similarity tables, which can be mis-weighted by certain metrics (termed the ‘double-zero’ problem).

\[ D_3 = \frac{\sum_{i=1}^{n} |y_{i1} - y_{i2}|}{\sum_{i=1}^{n} (y_{i1} + y_{i2})} \] (3.5)

Complement of the Whittaker’s Index of Association [137], [163] – Similar to the Bray-Curtis, but takes the difference of percentages (in other words, the data is normalized first, then subtracted. This metric is based on ecological differences developed by Whittaker and adapted to abundance data and like the Bray-Curtis, satisfies the triangle inequality axiom and the double-zero weighting requirements.

\[ D_4 = 1 - \frac{1}{2} \sum_{i=1}^{n} \left| \frac{y_{i1}}{\sum_{i=1}^{n} y_{i1}} - \frac{y_{i2}}{\sum_{i=1}^{n} y_{i2}} \right| \] (3.6)

Kulczynski – A recently advanced metric based on the work of Faith et al. and Kuczynski et al. [160], [162] (no relation to Kulczynski). This metric has been shown to potentially represent a large amount of variation in model microbial datasets using
multivariate analysis. This metric may be related to the Complement of Whittaker’s based on an assertion from Whittaker’s 1952 paper [147] of $\min(y_{i1}, y_{i2}) = 1 - 0.5 |y_{i1} - y_{i2}|$.

$$D_5 = 1 - \frac{1}{2} \sum_{i=1}^{n} \left[ \frac{\min(y_{i1}, y_{i2})}{\sum_{i=1}^{n} y_{i1}} - \frac{\min(y_{i1}, y_{i2})}{\sum_{i=1}^{n} y_{i2}} \right]$$  \hspace{1cm} (3.7)

Morisita-Horn [154] – One of the most widely used metrics in ecology. However, this measure has recently been shown to perform poorly at showing variations in multivariate analysis using ordination plots and standard clustering techniques [163-164].

$$D_6 = \frac{\sum_{i=1}^{n} y_{i1} y_{i2}}{\left(\frac{\sum_{i=1}^{n} (y_{i1})^2}{\sum_{i=1}^{n} y_{i1}} + \frac{\sum_{i=1}^{n} (y_{i2})^2}{\sum_{i=1}^{n} y_{i2}}\right) \left(\sum_{i=1}^{n} y_{i1} \sum_{i=1}^{n} y_{i2}\right)}$$  \hspace{1cm} (3.8)

 Phylogenetic metrics

These are metrics that consider phylogenetic distance between members.

UniFrac [164] – A widely used metric designed for differentiating microbial communities. This metric measures the normalized branch length of all unique branches of a tree containing all sampled species combined and was specifically formulated to study sequencing diversity.

3.7 Multivariate Statistics

A major new toolkit for the microbial ecologist is the application of multi-dimensional statistical techniques to pyrotag datasets. Like their univariate cousins, these methods can detect components of variation relating to differing factors of external data (metadata). A brief summary follows, though it is highly recommended for the reader to consult statistical references [137], [139], [150], [165] and linear algebra or vector mathematics references. A number of software packages exist for multivariate statistics, such as R, Matlab, and SciPy: As analyses are spread out between different packages, or not yet implemented, a strong programming background (Python, C++) and UNIX (VIM/SED) experience has been beneficial in the analysis stage more so than for pipeline processing, as each study often requires unique analyses or graphics generation and data manipulation to answer specialized hypotheses apart from packaged scripts.
3.7.1 **Principle Component Analysis / Principle Coordinate Analysis (PCA/PCoA)**

These two analyses use the same ordination technique but on differing sets of data. Principle Component Analysis operates on the set of matched environmental variables, such as all pH, NO3, TOC, and other factors (forming an association matrix) over a sample period and tries to answer "Which of these variables are related?" Principle Coordinate Analysis operates on the set of distances between samples (forming a distance matrix), and answers the question "Which of these samples are related?" The eigenvectors and eigenvalues of the data matrix are computed using matrix operations. The eigenvectors are charted in n-dimensional space, in decreasing order of eigenvalue size, and are the ordination coordinates. The number of environmental variables being studied determines the number of eigenvectors, or components. The eigenvectors themselves are meaningless, and ordination plots of eigenvectors do not have a physical quantity associated with them directly. Meaning is drawn from this analysis by the use of contribution plots. The magnitude of each eigenvector component is graphed in bar chart form with each value associated with the environmental factor in the order listed in the data matrix. These loadings represent how each factor varies with the other in this particular dimension. Each dimension can then be characterized as having positive and negative loadings from each factor and conclusions drawn from the context of factor associations (e.g. transmembrane pressure varies inversely to season in a membrane bioreactor along the first dimension, suggesting that the first dimension is represented by the seasonal changes on membrane energy consumption). Additionally, the proportion of each eigenvalue to the sum total of eigenvalues represents the fraction of variation explained by any dimension and is often denoted on the axes of ordination plots. Though two or three dimensions are often charted, statistical testing must be done to check for which dimensions are most significant. Oftentimes a rule of thumb of that each eigenvalue $\lambda > 1$ is significant. However, this can be overly conservative so several different models are proposed with better refinement, such as the 'broken stick' test. By comparison, ordination plots of PCoAs are essentially variations in distance between samples. In some cases, it is useful to overlay a bi-plot representing variable data onto an ordination data. Variable data is transformed by singular value decomposition using the
original eigenvalue to form a matrix of vectors representing each component in the ordination plot and can be thought of overlaying the contribution plots onto the main ordination plot. In this way, loadings from each sample can be compared with variables in the same ordination plot.

These two analysis are primarily used in this study, however, several other types of analysis are commonly used as well are summarized below [165]:

Non-Metric Multi-dimensional scaling (NMDS)

A method similar to PCA/PCoA, but preserving the order of objects rather than distances and can be thought of as multi-dimensional least squares fitting of ordered objects. Results from NMDS have been compared as being similar to PCoA analysis.

Canonical Correspondence Analysis (CCA)

Correspondence Analysis attempts to maximize the correspondence between two sets of data in an ordination space, rather than maximize variation in a single dataset in an ordination. This is done by randomly and iteratively scoring abundances and samples until steady convergence, and then applying decomposition of chi-squared distances (similar to PCA) to lower dimension ordination. This can tease out environmental factors related to species abundance if the latter has a normal distribution to the former.

Canonical Redundancy Analysis (RDA)

This technique can be thought of as an extension of PCA where the components are constrained to be linear combinations of environmental variables, though the analysis question is more similar to CCA. Distance based RDA (db-RDA) operates on a distance matrix rather than an association matrix and has been used to identify trends in microbial recovery from gut antibiotic dosing [166].

UPGMA Clustergrams

A clustering and visualization technique rather than a multivariate analysis, UPGMA can be considered as an alternative to multivariate ordination. UPGMA clustergrams
represent the variation in a dataset in its entirety, rather than the major components, and can be used for datasets with smaller numbers of samples to illustrate relatedness and variation by a tree topology. There is still debate whether ordination plots remove variation of interest in ecological studies, so this may be a suitable compromise solution.

3.7.2 Co-abundance

Testing the relatedness in time of two sample groups has been rarely done, though increasingly scientists are turning to the problem due to interest in stability and persistence of microbial communities. Highlighting temporally significant species can be thought of as an orthogonal operation of analysis of sample relatedness. Rather than seeing if few samples containing numerous species are related, we are examining if numerous species are related by a few samples. This also involves characterization of species time trends to understand what is normal association. A related field of co-occurrence has been long established which relates to the study of co-occurring organisms on independent islands [167–169]. The C-score (checkerboard statistic score) identifies the likelihood that the distribution of organisms could arise randomly. This can be used to study the influence of competition and non-random forces on species presence in independent biomes. However, this cannot be applied to time-course studies that involve related environments over a time period. Additionally, the problem studied for wastewater ecology is not related to how much are variations non-random variations. Therefore, the study of distances between OTUs due to variation in abundance over time will be referred to as co-abundance analysis, or ‘how much do organisms occur at the same point and how similar are their relative abundances.’

Some researchers have attempted to correlate microbes in time using non-sequencing techniques such as ARISA. The most related study to the problem of co-abundance involved designing a lag-delayed Pearson’s correlation coefficient to detect organisms that not only trend together but also lag one another, called the Local Similarity Analysis (LSA) relationship [170]. However, based on work by Wells et al. [171], it has been shown that turnover in wastewater treatment for most organisms last only a week, and a few weeks at most, whereas most of the samples taken for the study of the next chapter
were months apart, making LSA correlations potentially spurious. When distance matrices of correlations between species in time are computed, this data can be translated into co-varying graphs showing networks of co-abundant organisms [172]. However, a major problem in these network graphs is that when complex high-throughput data is portrayed the networks are too complex to be easily understood and contain too many uncharacterized features to be fully displayed. Instead, for the study of the next chapter UPGMA clustering of relatedness is characterized using the Bray-Curtis metric to study how organisms self-organize in microbial systems. A number of tests were performed to understand how much perturbation and error might influence such results. The problem of under-sampling was especially considered in a test case that resulted in highlighting a major property of such datasets. Since the vast majority of microbes only bloom once and disappear (at least at the 97% similarity level), most organisms correlate only with other organisms who appear or bloom at the same maximal time point. Only a small number of sequences correlate at multiple time points. This is illustrated below in Figure 3.7 with sample rarefaction. Randomly subsampled distributions can show how sensitive factors are to undersampling by exhibiting different behavior at smaller subsamples.

Bray-Curtis distances were calculated between all OTUs in the study, while each sample was individually and increasingly subsampled without replacement (jackknifing). All Bray-curtis distances of every OTU were averaged into a single value at each subsample. These averages were averaged again over 100 trials at the same subsampling depth. These average values were then graphed over all subsampling trials in Figure 3.7A. In most cases, Bray-Curtis distances increased as fewer sequences were subsampled, indicating that similarity decreased as samples were subsampled. This can be thought of as: as samples were subsampled, a larger fraction of shared OTUs was progressively removed as opposed to unique OTUs. However, as other samples were increasingly subsampled, averaged OTU Bray-Curtis distances increased, indicating the opposing phenomenon; more unique OTUs were removed rather than shared OTUs. Figure 3.7B shows the similarity between samples as a measure of the UniFrac distance between samples. The amount of shared OTUs removed with subsampling is influenced by the amount of sharing samples had with each other, meaning that some samples that have
Figure 3.7. Effect of OTU sharing and sample similarity on OTU distances. The behavior of subsampling of each sample is shown in A and trends with the level of similarity of each sample, represented in B as the averaged un-weighted UniFrac score between all samples, listed in increasing order of dissimilarity.

Increased similarity with other samples will be sensitive to under sampling, as more sharing is lost. This means that samples with increased similarity with other samples will influence OTU co-abundance distances disproportionately than samples with less similarity. Given that microbial turnover is not constant and that increased similarity assumes lower turnover rates when sampled at a uniform time interval, this implies that samples taken when turnover rates are lower will influence co-abundance relationships more than samples taken when turnover rates are high.

In summary, modern bioinformatics methods have rapidly progressed to address some of the first generation issues in NGS analysis, but numerous areas of research still remain. Additionally, as new platforms become commercially available, new challenges and an even greater amount of data will require new solutions, so the future of bioinformatics is only just beginning.
CHAPTER 4
MICROBIAL TURNOVER, PERTURBATION, AND SELECTION ACROSS A
BREWERY WASTEWATER TREATMENT WORKS AND PILOT BIOREACTOR

Author's Note: This chapter presents the work of a paper submitted to *The ISME Journal*. Authors: Jackson Lee, Andrew Logan, Seth Terry, and John Spear. Co-authors operated the reactor over the study period, collected samples, and provided operational details.

4.1 Abstract

Wastewater treatment plants are designed with distinct unit processes to select for specific biological functions. Pyrotag sequencing has enabled tracking of these microbial systems to a sequencing depth previously unseen. In this study we investigate microbial diversity and dynamics across a full-scale brewery wastewater treatment plant and a parallel pilot bioreactor during start up, stable operation, and post-disturbance. Sequencing yielded many known wastewater-associated phyla consistent with previous research, though with great novelty in-phyla. Classical indicators of alpha and beta diversity identified variations in community structure due to both startup and disruption events in a pilot reactor. The major shifts due to the disruption related to shuffled abundances of several clades of diazotrophic bacteria in Alpha-, and Betaproteobacteria. Moreover, the most abundant organisms in the pilot reactor stood apart as persistent across time and had high maximum relative abundances in the system, which bears resemblance to trends seen in microbial biogeography. Common and cosmopolitan organisms were also prone to be from influent sources. However, the most significant influence on diversity across the treatment plant was the type of unit process. These results show that unit processes in a treatment works are highly selective with rapid turnover and that disruptions to a pilot-scale reactor can permanently alter both the dominant and background consortia. Our results indicated that community membership does not return to initial conditions after disturbance, despite stable treatment performance. This has implications for the concept of core microbiomes in wastewater treatment.
4.2 Introduction

The rapid increase in sequencing throughput with decreased cost of next-generation sequencing technology has greatly expanded the detection coverage of microbial diversity in environmental samples and has the potential to explain the functioning and variation of species in environmental engineering applications such as wastewater technology, though a lack of theoretical development in microbial ecology hampers efforts. To predict, control and enhance current bioreactor designs for new uses, engineers may advantageously apply ecological theory to inform the underlying framework that dictates the stability of complex mixing microbial systems [86], [173]. In this study, the microbial diversity of a brewery wastewater treatment works was analyzed using pyrotag sequencing technology, with particular emphasis on documenting how changes in alpha and beta diversity are influenced by reactor startup and disruption. This study examined the operation of a novel pilot-scale aerobic treatment stage modified to produce a sustainable Single Cell Protein (SCP) product as an alternative to fish meal protein in aquaculture feed. SCP from the microbial treatment of food wastes has the potential to become a sustainable animal feed protein source, a source of vitamin B12 and essential amino acids [23] and a way to reduce sludge disposal costs and environmental impacts. Characterization of the relevant microbial community of the reactor system can be used to help understand the fundamentals of reactor performance and suggest avenues of research for biotechnological improvement of SCP quality. Figure 4.1 shows the brewery Wastewater Treatment Plant (WWTP) and pilot reactor research site used in this study (New Belgium Brewing, Ft. Collins, Colorado). The brewery treatment works itself consisted of an influent flow equalization basin with acidogenic conditions, then a methanogenesis system consisting of an Upflow Anaerobic Sludge Blanket (UASB) bioreactor. Next followed an aerobic treatment basin and clarifier for activated sludge growth and settling prior to discharge to the city Publically Owned Treatment Works (POTW). The pilot reactor received wastewater from the acidogenic basin to aerobically treat the wastewater at low Mean Cell Residence Time (MCRT) (< 8 days) with nutrient addition (N as urea, P as phosphoric acid, and a customized micronutrient cocktail) to increase the protein concentration. The pilot reactor remained online for more than one
Figure 4.1. Overview of the study site showing the brewery treatment works (top) and the pilot scale reactor (bottom). Colored squares indicate sampling location and are used to illustrate figures throughout this chapter.

year and consistently produced product for commercial feed trials [174], except for a brewery maintenance shutdown in December of 2008 which stopped wastewater influent for two weeks and increased dissolved oxygen (DO). Samples collected from all unit processes were examined by pyrotag techniques for diversity across the treatment works, and the pilot reactor was specifically examined for dynamics.

4.3 Materials and Methods

Sanger sequencing, pyrosequencing, bioinformatics, metric-based and co-abundance testing were used in this project to categorize sequences and relate them to each other.

4.3.1 Sampling and Sanger Sequencing

Samples were collected over a period of 1 year from throughout the brewery treatment works as well as from the pilot plant and final dried product to track changes in microbial ecology across treatment stages and in time. Sample DNA extraction, Sanger sequencing, and full-length 16S SSU rRNA gene bioinformatic methods were adapted from Sahl, et al., 2010 for 8F and 1492R bacterial primers [175]. Template DNA was amplified with the 8F (5’-AGAGTTTGATCCTGGCTCAG-3’) [136] and 1492R (5’ – ACCTTGGTACCTTT – 3’) [136] primers using Promega PCR Master Mix (Promega
Corp., Madison, WI), using the thermal-cycler program 94°C, 2 min initial denaturation, 30 X at 94°C 30 sec denaturation, 55.5°C, 1 min annealing, 72°C 1:30 min extension, 72°C, 12:00 min final extension and 4°C final hold. PCR product was then Montage gel purified (Millipore, Billerica, MA) and cloned into Invitrogen Top10 electro-competent cells. Clone plates of 96 wells were Sanger sequenced on a MegaBACE 1000 dye-terminating sequencer. Chromatograms were called and quality controlled by XplorSeq [176] and exported for Mallard analysis. Full length chimera-checked (Mallard) [120] 16S sequences were then clustered and aligned in mothur 1.19.0 [125], [142] using the Silva Non-Redundant 104 database and incorporated into the ARB [177] reference tree alignment of the database for “Silva” (ENA-EBI curated) taxonomy determination by parsimony insertion. Nearest neighbor Silva taxonomies were copied onto the novel sequences. A subset of the bacterial domain in the Silva 104 Non-Redundant sequences (pintail_slv > 90, seq_quality_slv > 50, align_quality_slv > 50, ambig_slv = 0) with the augmented Sanger sequences was then exported for downstream bioinformatics processing. Database entries were pruned for all sequences that ran from E. coli position 100 to 8099 and were greater than 300 bp in length by a custom python script. A filtered alignment file composed of the pos_var_bac Silva mask at threshold ‘0’ was created for tree construction also by a custom python script. These sequences were used as the Silva [132] Non-Redundant 104 database customized database used for downstream analysis.

4.3.2 Pyrosequencing & Bioinformatics Pipeline

PCR of the 16S SSU rRNA gene using the bacterial 27F and 338R primers with sequence barcoding [10] were adapted from Sahl et al., (2010). Using QIIME 1.2.0 [88], raw sequencing files containing pyrosequences were denoised (default parameters) [113] into flowgram clusters and quality-controlled based on average quality score (>Q27), read length outlier screening (239-294 bp), ambiguous bases (0) and homopolymer count (<10) [110]. Operational Taxonomic Units (OTUs) of clustered sequences (Uclust [178], 97% similarity, default parameters) were filtered for chimeras using UChime (default parameters, de novo mode) [121]. The remaining OTUs were aligned using the NAST-based aligner and classifier of mothur [142] trained on the customized Silva 104 non-redundant database. Masked pos_var_bac ‘0’ aligned sequences were used with FastTree
[146] (default parameters) to generate a tree of all unique OTUs for use in Fast-UniFrac and phylogeny examination. The Ribosomal Database Project (RDP) classifier implementation in mothur (trained on the augmented Silva 104 database) was used to classify OTUs and determine taxonomy (80% cutoff). Note that taxonomic assignments with number assignments at the end (i.e. azonexus_2) indicate paraphyletic classification in the Silva reference tree system.

4.3.3 Alpha and Beta Diversity Analysis

A number of steps have recently been taken to understand the appropriateness of various pairwise beta diversity measures and multivariate analysis to represent microbial consortia level changes [150], [160]. However, the concept of using pairwise comparisons of similarity or dissimilarity to show succession over time and space is just beginning to be applied to wastewater technology. Several classical metrics of beta diversity, such as the Sørensen (a qualitative sharing metric) and Whittaker’s Index of Association (quantitative relative abundance sharing metric), were chosen to study the pilot plant system. These were originally conceived for the study of plant communities [149], [179] and have since been applied for the study of variations in bacterioplankton communities across distance in the North Pacific [180]. If time is indeed a successional gradient, or a proxy for other environmental gradients, similar distance-decay relationships should be seen when both the Sørensen and Whittaker’s Index of Association metrics are applied. Additionally, based on recent work on best metrics for representing sample dissimilarity, the unweighted UniFrac metric [181] was used.

QIIME was used to compute alpha diversity estimates of the Chao1 and ACE metrics [156]. The CatchAll statistic [158] was also used with default parameters and the best model total species observed and standard error used. QIIME was used to calculate the beta diversity metrics of unweighted and weighted UniFrac (using a whole tree of the entire dataset), Whittaker’s Index of Association $S_w$ (Equation 1) [137] (using a custom python script, not the qualitative Whittaker’s found in QIIME), and the qualitative Sørensen Index $S_r$ (Equation 2). Unweighted Pair Group Method with Arithmetic Mean (UPGMA) clusters (with or without jackknifing) and principal coordinate analysis with
bi-plots of OTU presence were determined using QIIME. Figures were generated with Excel or Matplotlib (http://matplotlib.sourceforge.net/), and edited with open source graphics software (Inkscape, http://www.inkscape.org).

Complement of Whittaker’s Index of Association $S_w$:

$$S_w = 1 - D_9 = 1 - \frac{1}{2} \sum_{i=1}^{n} \frac{y_{i1}}{\sum_{i=1}^{n} y_{i1}} - \frac{y_{i2}}{\sum_{i=1}^{n} y_{i2}}$$

(4.1)

where $y_{i1}$ is the abundance of the $i$th OTU in sample 1, and $n$ is the number of OTUs in the sample.

Sørensen Index $S_S$:

$$S_S = \frac{2C}{A+B}$$

(4.2)

where $A$ and $B$ are the number of OTUs in each sample, and $C$ is the number of shared OTUs.

4.3.4 Co-abundance Analysis

To understand the co-abundance of OTUs in both the pilot plant and final SCP product, OTUs that were present in these two environments on more than two dates were used. OTUs were included if > 2 sequences / OTU and >1 samples / OTU were present, of which 319 out of 1604 OTUs passed this criteria. The Bray-Curtis pairwise distance metric $D_{BC}$ (Equation 3) [137] was computed for each OTU based on the abundances of each OTU across all sample points (n=13). Each OTU was grouped into a function category of “Carbon-related,” “Nitrogen-related,” or “other” based on the metabolic profile of the nearest taxonomic representative found in Bergey’s Index [182] (characteristic of the entire known genus or family level) or related primary literature (of the genus level only). Clades with significant known metabolic variation or unknown classification (at the genus level) were labeled as “unknown or unclassified.” OTUs were additionally compared by online MetaStats (http://metastats.cbcb.umd.edu/detection.html) [183], a program developed to examine differential abundance of elements associated between patients from a control and
treatment category for medical microbiome studies. Abundance information from the series of equalization basin samples (representing the influent to the pilot plant) and the pilot plant mixed liquor samples were used as the pre- and post-treatment cases with a threshold p-value < 0.05 and a difference in relative abundance of +/- 1% absolute magnitude as being significantly "enriched" or "depleted." The cluster dendrogram, function, dates, and MetaStats significance information were then charted using the interactive Tree of Life (iTOL, http://itol.embl.de/). To test cluster variation due to under-sampling, jackknifing was done at subsample increments every 100 sequences over each sample and all resulting Bray-Curtis distances averaged for each subsampling depth for each sample using a custom Python script.

Bray-Curtis Distance Metric $D_{BC}$:

\[
D_{BC} = D_{14} = \frac{\sum_{i=1}^{n}|y_{i1} - y_{i2}|}{\sum_{i=1}^{n}(y_{i1} + y_{i2})}
\]  

(4.3)

where $y_{i1}$ is the abundance of the $i$th time point for the first OTU, and $n$ is the number of total time points in the sample.

Sample abundances were organized with Excel and consisted of OTUs in the pilot plant and final SCP product with > 2 sequences / OTU, of which 641 OTUs passed this criteria.

A major concern in this work was the role of PCR and extraction biases in calculating co-abundance dissimilarity distances. This work does not serve to capture the true compositional nature of the sample but rather relative abundance data is used to understand patterns in OTU distributions undergoing the same DNA extraction, amplification, and analysis conditions. Another concern was misidentification or the inclusion of spurious sequences leading to poor taxonomic identification. Therefore, no conclusions are made on any single OTU and no specific species level taxonomic assignments are made.

4.3.5 Nucleotide Accession Numbers

A total of 809 Sanger sequences were deposited in GenBank under accession numbers JQ072092 - JQ072899. Sanger sequences were named according to the location and date in which they were sampled (NBBEQ0908x44, EQ = Acidogenic basin, ME =
Methanogenic UASB, PI = Pilot Plant, SP = SCP product, AB = Aerobic Basin, OT = Wet Well Outfall, 0908 = MMYY, x44 = sequence number). 454 pyrotags taken from quality score processing were submitted in FASTQ format to MG-RAST (ID 4477674.3).

4.4 Results

After sequence processing, the microbial diversity and the metric-based variation in the patterns of diversity were analyzed.

4.4.1 Microbial Diversity Across a Treatment Plant

Table 4.1 shows the operational conditions, alpha diversity information, and sequencing depth for each sample. Alpha diversity estimates for each sample were bracketed on the low end by the Chao1 estimator, and on the high end by the CatchAll estimate. Samples from the influent and acidogenic basin showed the lowest levels of diversity, with each unit process tending to increase (on average) the alpha diversity downstream. The pilot reactor treatment process increased diversity over the influent acidogenic basin flow, but the majority of diversity in this study was found in both the UASB and the aerobic basin of the brewery wastewater treatment plant. Figure 4.3A shows the clustered ordering of phylum-level diversity from all samples based on the unweighted UniFrac distance, with the same information shown in Principal Coordinate Analysis (PCoA) ordination plot form in (B), which shows that the type of unit process influenced the consortia pattern the most. The exception to this is that pilot reactor mixed liquor samples clustered closely to the corresponding time point of the final SCP product. The four main dominant phyla were the Bacteroidetes, Firmicutes, Actinobacteria, and Proteobacteria (from the alpha, beta, and gamma classes) and are consistent with previous research on wastewater aerobic treatment [184] and process wastewater specifically [90]. The candidate divisions of SR1 and RF3, seen previously in methanogenesis bioreactor surveys [185], [186], comprised 2.8% and 2.3% of sequences respectively from the methanogenesis UASB process. TM7 [187] and BD1-5 [188] accounted for 5.1% and 4.7% of sequences detected in the treatment plant wet well. Though qualitative, high-rise plots have been used to show a large amount of data on a single figure and are useful for
Table 4.1. Alpha diversity estimates, sampling schedule, and operating parameters of each unit process. Note: Operating parameters are approximate.

<table>
<thead>
<tr>
<th>Environment Type</th>
<th>Volume, MCRT</th>
<th>Date, Day #</th>
<th># pyrosequences</th>
<th>Observed OTUs</th>
<th>Chao1</th>
<th>ACE1</th>
<th>Estimated Total Species</th>
<th>Standard Error (SE)</th>
<th>% Coverage</th>
<th>min-max estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment Plant Influent</td>
<td>800 m³/day, 8,000-10,000 mg/L COD, ~35 °C</td>
<td>1/30/08</td>
<td>527</td>
<td>52</td>
<td>68.15</td>
<td>75.99</td>
<td>113.2</td>
<td>42.40</td>
<td>46-76</td>
<td></td>
</tr>
<tr>
<td>Plant Influent Sample</td>
<td>-</td>
<td>3/19/08</td>
<td>821</td>
<td>38</td>
<td>57.43</td>
<td>61.48</td>
<td>70.7</td>
<td>13.70</td>
<td>54-66</td>
<td></td>
</tr>
<tr>
<td>Port, pH 10-12</td>
<td>-</td>
<td>4/30/09</td>
<td>722</td>
<td>8</td>
<td>9.50</td>
<td>11.87</td>
<td>14.2</td>
<td>9.50</td>
<td>56-84</td>
<td></td>
</tr>
<tr>
<td>Acidogenic Basin</td>
<td>600, -</td>
<td>1/30/08</td>
<td>1238</td>
<td>38</td>
<td>54.50</td>
<td>47.92</td>
<td>51.8</td>
<td>6.00</td>
<td>70-79</td>
<td></td>
</tr>
<tr>
<td>Mixed Liquor, pH 5-6</td>
<td>-</td>
<td>3/19/08</td>
<td>3170</td>
<td>85</td>
<td>147.33</td>
<td>134.50</td>
<td>189.7</td>
<td>32.60</td>
<td>45-63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>3/19/09</td>
<td>3457</td>
<td>66</td>
<td>70.71</td>
<td>78.45</td>
<td>98.4</td>
<td>18.40</td>
<td>67-93</td>
<td></td>
</tr>
<tr>
<td>Methanogenesis UASB</td>
<td>800, -</td>
<td>4/30/09</td>
<td>952</td>
<td>41</td>
<td>54.00</td>
<td>54.88</td>
<td>70.8</td>
<td>21.90</td>
<td>58-76</td>
<td></td>
</tr>
<tr>
<td>Outfall</td>
<td>-</td>
<td>1/30/08</td>
<td>1063</td>
<td>128</td>
<td>215.00</td>
<td>205.91</td>
<td>248.9</td>
<td>31.10</td>
<td>51-62</td>
<td></td>
</tr>
<tr>
<td>Aerobic Basin</td>
<td>3820, 16</td>
<td>3/19/08</td>
<td>2242</td>
<td>220</td>
<td>310.29</td>
<td>313.82</td>
<td>372.9</td>
<td>30.70</td>
<td>59-71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>4/30/09</td>
<td>867</td>
<td>131</td>
<td>227.00</td>
<td>233.22</td>
<td>405.6</td>
<td>150.40</td>
<td>32-58</td>
<td></td>
</tr>
<tr>
<td>WWTP Wet Well</td>
<td>-</td>
<td>4/30/09</td>
<td>1185</td>
<td>194</td>
<td>404.91</td>
<td>365.39</td>
<td>434.9</td>
<td>51.10</td>
<td>45-53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>4/30/09</td>
<td>792</td>
<td>155</td>
<td>237.13</td>
<td>270.08</td>
<td>360.7</td>
<td>59.80</td>
<td>43-65</td>
<td></td>
</tr>
<tr>
<td>Effluent</td>
<td>-</td>
<td>4/30/09</td>
<td>1310</td>
<td>211</td>
<td>332.87</td>
<td>357.35</td>
<td>454.7</td>
<td>49.40</td>
<td>46-63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1/30/08</td>
<td>1284</td>
<td>138</td>
<td>234.60</td>
<td>255.96</td>
<td>345.5</td>
<td>58.50</td>
<td>40-59</td>
<td></td>
</tr>
<tr>
<td>Pilot Plant Flow Rate: &lt;4.75 m³/day</td>
<td>38, -</td>
<td>3/19/08</td>
<td>2459</td>
<td>315</td>
<td>519.31</td>
<td>499.90</td>
<td>589.2</td>
<td>41.70</td>
<td>53-63</td>
<td></td>
</tr>
<tr>
<td>Pilot Tank Sample Port</td>
<td>38, -</td>
<td>4/30/09</td>
<td>1080</td>
<td>176</td>
<td>264.67</td>
<td>278.45</td>
<td>392.8</td>
<td>62.20</td>
<td>45-66</td>
<td></td>
</tr>
<tr>
<td>SCP Product</td>
<td>-</td>
<td>03/19/08, 0</td>
<td>4098</td>
<td>88</td>
<td>123.10</td>
<td>116.61</td>
<td>170.8</td>
<td>50.10</td>
<td>52-75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>04/08/08, 20</td>
<td>1347</td>
<td>127</td>
<td>202.64</td>
<td>250.43</td>
<td>324.6</td>
<td>56.70</td>
<td>39-63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>09/10/08, 175</td>
<td>4892</td>
<td>197</td>
<td>276.44</td>
<td>281.69</td>
<td>336.5</td>
<td>28.20</td>
<td>59-71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>01/06/09, 293</td>
<td>2850</td>
<td>193</td>
<td>275.83</td>
<td>284.22</td>
<td>389.3</td>
<td>69.80</td>
<td>50-70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>02/16/09, 334</td>
<td>3014</td>
<td>163</td>
<td>218.31</td>
<td>236.79</td>
<td>306.6</td>
<td>40.30</td>
<td>53-75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>03/19/09, 365</td>
<td>1925</td>
<td>151</td>
<td>215.17</td>
<td>218.02</td>
<td>241.0</td>
<td>19.00</td>
<td>63-70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>04/30/09, 407</td>
<td>1401</td>
<td>115</td>
<td>193.71</td>
<td>222.52</td>
<td>348.9</td>
<td>104.90</td>
<td>33-59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>05/07/08, 49</td>
<td>3275</td>
<td>92</td>
<td>117.09</td>
<td>114.73</td>
<td>145.5</td>
<td>30.20</td>
<td>63-80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>09/10/08, 175</td>
<td>3593</td>
<td>123</td>
<td>176.71</td>
<td>181.82</td>
<td>223.3</td>
<td>30.50</td>
<td>55-70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>01/06/09, 293</td>
<td>2126</td>
<td>171</td>
<td>232.11</td>
<td>226.89</td>
<td>267.4</td>
<td>18.90</td>
<td>64-75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>02/16/09, 334</td>
<td>1314</td>
<td>103</td>
<td>146.00</td>
<td>162.05</td>
<td>196.2</td>
<td>25.10</td>
<td>57-71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>03/19/09, 365</td>
<td>2414</td>
<td>147</td>
<td>217.38</td>
<td>229.96</td>
<td>264.8</td>
<td>25.30</td>
<td>56-68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>04/30/09, 407</td>
<td>1230</td>
<td>94</td>
<td>155.88</td>
<td>161.59</td>
<td>202.4</td>
<td>42.80</td>
<td>46-60</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.2. High-rise figure of treatment works microbial diversity. OTUs from the entire dataset are ordered first alphabetically by taxonomy (down to taxonomic depth 6 (~genus)) and then by rank abundance on the x-axis. Samples are ordered by environment (color based on Figure 1) and then by date on the y-axis. The relative abundance of each OTU in each sample is shown on the z-axis.
Figure 4.3. Phylum level distribution clustered by pairwise UniFrac unweighted distance (A). Samples were clustered by UPGMA with jackknife support of unweighted UniFrac distance and shown as a dendrogram with clades common to a unit process colored (A, top). Stacked bar graphs of phyla distributions (A, bottom) are sorted by total dataset rank abundance starting with the most common from the bottom. The PCoA of unweighted UniFrac of the same dataset is shown in (B).
Figure 4.3B Continued

quickly drawing attention to the inherit patterns that can be distinguished from pyrotag datasets [12], [13]. Figure 4.2 shows that differential enrichment of several types of organisms occurred over successive unit processes primarily in the phyla Bacteroidetes and Proteobacteria, and that much of the classifiable rare biodiversity tended to be in the Firmicutes. In addition, approximately 14% of all OTUs remained unclassified beyond the domain level (6.0% of pilot plant OTUs, 5.2% of methanogenesis OTUs, and 4.0% of aerobic basin OTUs). Moreover, from the wet well 20% of pyrotag sequences do not have classifications deeper than the phylum level. Several full-length sequences form a distinct genus related to Arcicella in the Cytophagaceae (Figure 4.4).

4.4.2 Temporal Dynamics of a Pilot Reactor

Alpha and beta diversity estimation derived from classical macro-ecology (particularly plant ecology) has been used to describe microbial systems [189–191]. In the pilot
Figure 4.4. A novel genus related to genus Arcicella (shown in grey). Full-length aligned sequences (with variable region masking) were added into the Silva 104 non-redundant reference tree by parsimony insertion.
reactor, when three common metrics of alpha diversity (Chao1, ACE, and CatchAll) are taken together, clear trends are visible for both the startup period and disruption (Figure 4.5). Startup entailed the inoculation of the pilot reactor for several weeks with material from the methanogenic reactor, followed by feeding from the acidogenic basin directly. This led to an increase of microbial diversity over the period of several months as enrichment occurred. Around Day 280, a shutdown event occurred which resulted in a disruption of all pilot reactor activities and a loss of alpha diversity without subsequent recovery of that alpha diversity by the end of the study. Concurrently, SCP crude protein values also decreased in this interval though the relationship is not known.

When tracked over time, startup and disruption within the pilot plant produce the same pattern of shifts with the beta diversity metrics studied (Figure 4.6, A, B, C), but do not follow a linear gradient as would be expected with steady succession or turnover (Figure 4.6D). Instead, the pilot reactor community begins similar to the influent environment at startup and then shifts first to one community composition, and then after disruption, shifts to a second community that is more similar to the original influent community than the preceding community. The only time-decay relationship is seen briefly after the disruption event (Figure 4.6D). Examination of bi-plots of the most common family-level classifications of OTUs also indicates organism enrichment occurring due to disruption. The combination of alpha and beta diversity results shows high species turnover and rapid growth rates in the bioreactor during startup, which resulted in many species but a smaller proportion of sharing between time points. However, after the influent was switched to an influent source with more restricted diversity, a subsequent disruption removed much of the alpha diversity and resulted in a more stable but less diverse set of surviving organisms.

4.4.3 Co-abundance of OTUs Across Time

While it is useful to compare environments over time, another approach is the examination of co-abundance of OTUs across time for the purpose of understanding OTU contributions to overall diversity variation. The variations of OTU abundances in the pilot reactor and SCP product samples over time were used to generate Bray-Curtis
Figure 4.5. Alpha diversity dynamics of reactor startup and disruption. Pilot plant mixed liquor samples are shown in black, and SCP samples shown in white over time with three different alpha diversity estimators (Chao1, ACE, CatchAll)

distances between OTUs. The UPGMA cluster map of these OTUs is shown in Figure 4.7. The majority of organisms cluster based around three factors. The dominant clustering factor is the date at which the OTU had its maximum abundance (Figure 4.7A), followed by the magnitude of maximum abundance (Figure 4.7D), and finally, deep clusters (Figure 4.7B, horizontal lines) are related to the relative abundance in samples before and after the disruption event (Figure 4.7C). Particularly, the deep shift in the reactor seen in beta diversity plots is reproduced as three deep clusters: 1. Very abundant outliers (6.2% of OTUs, 67.6% of sequences) (shown in Figure 4.7B); 2. OTUs that appear more before disruption (magenta line); and 3. OTUs that appear more after disruption (teal line). Additionally, the outlier cluster was also clustered into before- and after-disruption clusters. When the MetaStats statistic was applied to these OTUs with relative abundances seen from all influent measurements (Figure 4.7B), a number of abundant OTUs were shown to be enriched by the pilot reactor but not all, indicating that
Figure 4.6. Microbial turnover across a reactor during startup and disruption. PCoA plots of unweighted UniFrac (A), Whittaker’s Index of Association (B), Sørensen Index (C), and distance-decay of Whittaker’s Index of Association (D). PCoA plots (A-C) are included with influent samples as a baseline (white circle) and pilot plant mixed liquor (black) and SCP (grey) time courses. PCoA bi-plots of the level 4 Silva 104 non-redundant database taxonomy of the most abundant OTUs are denoted. Distance-decay pairwise comparisons (D) are graphed according to the difference in days of time points.
Figure 4.6B Continued
Figure 4.6C Continued
Figure 4.6D Continued
relative abundance was influenced from external immigration into the pilot reactor. The abundant OTUs that are enriched in the pilot reactor consist in large part of OTUs that classify as genera from known diazotrophs within the Proteobacteria and match the family-level classifications shown in the beta diversity bi-plots (Figure 4.6); this suggests replacement of diazotrophs by the disruption event.

Though Figure 4.7 clearly shows that abundant organisms stand out as a distinct cluster, it is unclear if organisms that are abundant necessarily remain abundant over time, especially given the concepts of logarithmic growth and viral action [195-196] which suggest that rapid variation and boom-and-bust cycles should be the norm in time-course studies. Figure 4.8 shows the relationship between maximum abundance and the temporal distribution of each OTU. While minimum dispersion has no bound based on abundance, maximum dispersion appears to be related to some upper limit based on logarithmic abundance. Additionally, when the same differentially enriched or depleted OTUs seen in Figure 4.7 are labeled in Figure 4.8, enriched and depleted OTUs separate clearly along a dispersion gradient. This clustering of enriched and depleted OTUs may relate to the differing metabolic role of the acidogenic basin upstream and mixed-liquor organisms (Figure 4.9). The significantly depleted OTUs are found to be primarily saccharolytic fermentative anaerobes such as Prevotella, whereas three separate types of diazotrophs as well as some fermentative anaerobes are enriched in the mixed liquor.

4.5 Discussion

The discussion portion of the study is as follows and tries to relate variation within treatment plants to current thinking on the functioning and organization of microbial communities.

4.5.1 Insights into Microbial Diversity in Wastewater Treatment Plants

This study reveals that despite having a common source of colonization from a brewery, each unit process in brewery wastewater treatment remains highly selective. Moreover, constant turnover of even the most common species is to be expected as a normal occurrence in wastewater plants and fits with previous work on the dynamics of
Figure 4.7. OTUs in a pilot reactor and SCP product clustered over time. OTUs were UPGMA clustered by co-abundance based on Bray-Curtis distance (A). Deep clusters of outlier OTUs, OTUs before disturbance, and OTUs after disturbance, are marked by the brown, magenta, and teal horizontal lines respectively; clades were colored by the date of maximum abundance and marked if significantly enriched or depleted according to the MetaStats statistic (B). The distribution of OTUs before and after disruption is shown in (C). The maximum abundance of each OTU is shown in (D).
Figure 4.8. Relationship between abundance and temporal dispersion. OTUs are graphed by the maximum relative abundance for each OTU over all time points. Semi-log fit of all data is shown (black line) as is the upper limit of dispersion (dashed line). Circles represent OTUs differentially enriched according to the MetaStat statistic test (grey), the OTUs differentially depleted (white), and those OTUs where no significant differential change was found (black).
Figure 4.9 Phylogenetic tree and sequence abundance of depleted and enriched OTUs. Enriched OTUs (orange) and depleted OTUs (blue) show clustering (A) by phylogeny, and have differential metabolic roles with the Proteobacteria (P) being mostly of root-zone associated diazotrophs, and the Firmicutes (F) and Bacteroidetes (B) being anaerobic fermenters. This results in (B) a differential enrichment of root-zone associated diazotrophs into the final SCP product.
wastewater treatment [194]. The large relative abundance of diazotrophs seen in this study has a parallel in the treatment of certain high-strength wastewaters such as olive, paper and pulp mill wastes [195–198]. In the pilot reactor of our studied system, aeration estimates based on best practice produced dissolved oxygen levels 0.1 – 0.5 mg/L in the pilot reactor. There may be similarities with this environment and the rhizospheric habitat where such organisms autochthonously exist, particularly the abundance of simple organic acids from acidogenesis which in the root-zone is used for plant symbiosis with diazotrophs [199]. The wide presence of diazotrophs indicates a potential target for future metabolic biotechnology exploitation. Lastly, unclassifiable OTUs in this study were found in all environments downstream of the acidogenic basin. Particularly surprising is the number of unclassified organisms found in the methanogenesis UASB and the wet well outfall. The design of the UASB may be one contributor to the diversity, as the sludge granules that are retained in the sludge blanket can be considerably older than indicated by the hydraulic residence time. UASB sludge granules are known to be highly complex physical structures with differing microbial composition based on depth from the surface of the granule [200–203]. Methanogenesis (and the complex structure of UASB sludge granules in particular) still remains an open frontier of biofuels discovery. The wet well receives low settling COD effluent from clarifier operations prior to discharge from the plant and may be an overlooked environment for sampling, likely harboring organisms related to wastewater treatment washout.

4.5.2 Alpha Diversity and Coverage

When examining the role of species richness across disruption (Figure 4.5), it is unclear why alpha diversity did not return to original levels. This may indicate a fundamental loss of diversity without adequate replacement from the lower diversity acidogenic basin, or a much slower period of recovery beyond the scope of the experiment, or a permanent change in reactor environmental conditions or influent chemistry. Of particular note is the wide range and error of these alpha diversity estimation models. The use of alpha diversity estimators has limitations for a number of reasons. A central issue still under discussion is the appropriateness of various models to
extrapolate the true rare abundance profile [155]. Another major issue is the assumption that fixed cluster sizes have uniform phylogenetic or taxonomic significance [204]. Despite these potential issues, these estimators provided a consistent relative measure of total species richness, but potentially may not be directly related to system functioning but rather other factors such as residence time.

4.5.3 Normal Variation vs. Disruption

Several recent studies have established a baseline examination of how wastewater community consortia fluctuate in real world systems (though no other high-throughput sequencing projects exist on the effect of disruption in wastewater treatment systems). Wells et al. (2011) [171] showed that after a year of monitoring a full-scale steady aerobic basin, a dynamic succession of microorganisms could be observed with no return to a starting community. Work by Werner et al. (2011) [85] monitoring 9 separate methanogenesis reactors over time showed that a steady but minor succession of changes in community structure occurred punctuated by short bursts of deviation, then returning to an initial point. An analog for disruption comes from gut microbiome sequencing observations on the quasi-stable nature of microbial systems undergoing antibiotic disruption [205] which indicate that microbial communities, once disrupted, have difficulty returning to an original state and may not return at all. The disruption seen in this work is specifically a resource disturbance, rather than a toxic shock from an antibiotic; however, both the study by Dethlefsen et al. (2010) and this work show that significant disruptions can easily switch communities into new community profiles after a period of lowered alpha diversity. Furthermore, such disruptions do not result in a linear distance-decay until after a new equilibrium is reached. With time and availability of DNA sequencing technology, regular analysis of microbial community composition of many reactors will enable engineers to better define the range of healthy microbial succession and the severity of upset consortia turnover.
4.5.4 Common, Cosmopolitan, and Endemic Microorganisms in Wastewater Treatment

A number of studies using different molecular techniques have shown that, like their macro-ecology counterparts, microbes appear to partition into two groups: common (high abundance in a location) and cosmopolitan (widespread across many or all locations), or rare (low abundance in a location) and endemic (present in few or 1 location) whether on a global scale [209-210] or a local distance gradient [180]. A 0.1% - 1% relative abundance rule of thumb (a ‘fuzzy boundary’) is suggested to be the division between biospatially common and rare organisms [191]. In this study a logarithmic trend is observed over a temporal scale and bears resemblance to trends of microbial dynamics in open ocean communities [211-212]. Here, examination of a single site over time reveals a similar pattern of common and cosmopolitan, and rare and endemic OTUs distributed in a logarithmic manner as in biospatial analysis. The two ideas of time and space are related in that species-area and species-time relationships are both phenomena driven by turnover of species along changes in environmental conditions, dispersion, and selection processes (though mechanisms driving each process may not be equivalent) [190], [210].

The distinction of common and endemic organisms appears possibly as an artifact of logarithmic distributions, i.e. more members at low abundance with a few notable outliers at high abundance. There also appears to be a limitation on maximum dispersal related to an organism’s maximum abundance that may be physiological in nature, but more work is needed to understand the exact mechanism of such a limitation. It is reasonable to assume that common and rare organisms have different lifestyles, but given the unusual boom-and-bust spatiotemporal pattern of most species, selecting measures of dispersion and central tendency that have significant ecological or physiological relevance is greatly needed.

4.5.5 Sources of Immigration in a Pilot Reactor

Significantly, from Figure 4.8, the source of common OTUs in the pilot reactor could be distinguished based on how cosmopolitan they were. Almost all common cosmopolitan OTUs found in the pilot reactor were present as OTUs depleted from the influent. This is not unusual as each unit process serves as a constant source of inoculum.
to a downstream process. Moreover, depleted and enriched organisms were clustered phylogenetically (Figure 4.9), which indicates a common metabolism at clusters greater than the 97% similarity level and again highlights that 97% clustering may not have functional meaning in ecology. Additionally, common OTUs preferentially enriched in the pilot reactor were almost never cosmopolitan due to the community-altering disruption mid-way through the experiment. It has been suggested that at small size scales, temporal richness is increased, whereas at larger scales temporal richness is reduced [210]. From a reactor operations perspective, smaller reactors may lack the chemical and microbial buffering capacity to withstand the same upsets, resulting in industrial reactor systems having more steady-state behavior than their smaller research counterparts.

4.5.6 The Core Microbiome and the “Rarer” Biosphere of Wastewater

Finally, if the concept of common and cosmopolitan organisms holds for wastewater treatment, this raises issues for the search for core microbiomes. The idea of common, cosmopolitan organisms and core microbiome organisms share great similarity. The core microbiome concept relies on a common set of organisms across an environment type to account for ecosystem functioning [152]. However, seeking core microbiomes from wastewater ecosystems has resulted in a noticeable but extremely reduced set of OTUs [186]. Given the selective and ‘moving target’ nature of wastewater ecology and variable wastewater environments, universal core members (at least at the 97% similarity level) may exist but will be very site and time specific and greatly depend on the rate of microbial turnover and depth of sequencing [211]. A true core microbiome may be more of a statistical measure of how often and how abundant an organism appears rather than a matter of total sharing, and fits with a more nuanced view of a ‘persistent’ core microbiome [212]. From co-abundance analysis, common, widespread, and endemic organisms are all concurrently in flux and a constant rare biosphere background was not observable at this sequencing depth, posing interesting questions of how to find newer meaningful definitions of ‘rare’ organisms relevant to high-throughput sequencing studies.
4.6 Conclusion

In summary, each unit process of the investigated treatment plant behaves as a unique habitat, weakly linked by geography, more strongly linked by time, and always showing great turnover. A pilot-scale research reactor showed much less diversity and much more variation in its operation compared to its large-scale counterpart, suggesting that studying smaller reactors may have fundamentally different microbial succession activity involved when compared to full-scale systems. Startup resulted in a rapid increase and turnover of species, while disruption was marked by a drop in species, increased similarity over time, a switch in species composition, and failure of recovery of the subsequent community to return to the original diversity levels and community composition. In particular, the distribution of organisms over time in a bioreactor resembles the concept of cosmopolitan, common, and endemic microorganisms seen in spatial biodiversity analysis. Finally temporal trends in external immigration could be distinguished from endogenous growth in a bioreactor. This analysis identified diazotrophic growth in high-strength wastewater treatment systems as a possibility for further improvement of SCP protein content.
CHAPTER 5
PROTEINS AND PROTEOMICS OF A SUSTAINABLE FEED PRODUCED FROM
BREWERY WASTEWATER TREATMENT

Author’s Note: This chapter presents the work of a paper in preparation to *Biotechnology and Bioengineering*. Authors: Jackson Lee, Andrew Logan, Seth Terry, and John Spear. Co-authors provided data on production parameters over time on SCP production and produced SCP material.

5.1 Abstract
Crashing global fish stocks have led to a demand for sustainable aquaculture, but this effort is diluted by the necessity of using large amounts of pelagic fish meal to feed farmed fish. In this interdisciplinary study, protein characterization techniques from several fields (including metaproteomics) are used to identify product constituents and to fractionate a bacterial protein fish meal replacement called Single Cell Protein (SCP) produced from brewery wastewater. In previous work, a pilot bioreactor producing SCP showed that two groups of microbes were dominant in the microbial community, saccharolytic fermenters and diazotrophic Proteobacteria. This study attempts to identify proteins and metabolisms that can be promoted for increased SCP production by these organisms. Combined with organic acid analysis, the results here identified the role of organic acids as a major substrate for diazotrophs in the bioreactor, and the possible currency of exchange between the two groups of organisms. Amino acid profiles showed steady composition but proteome profiles showed variation in expression over time which indicate that the majority of proteins have similar amino acid content and that amino acid profile perturbation is difficult. Approximately half of 2D-PAGE spots yielded protein matches, with peptide identification showing a conversion of beer-related proteins in the influent to proteins related to aerobic growth or organic acid production in the mixed liquor. Overall, this study shows the role of acidogenic pretreatment in producing a normalized SCP product using the natural response of diazotrophs to organic acids.
5.2 Introduction

As fish consumption surges with a growing global population, the toll on marine fisheries has been severe, with approximately 90 million tonnes of fish taken from the world's oceans in 2010 and many species now considered over-exploited or depleted [25]. Aquaculture has grown at about 8.6% a year [25] to fill this need for more protein, but aquaculture relies heavily on marine fish meal for feed, with about 0.63 pounds of wild caught fish to produce one pound of farmed fish on average [27]. This is a major 'sticking point' for the rationale of using aquaculture to sustainably address global protein needs [1], as up to 70% of feeds for carnivorous fish and 50% for shrimp can be comprised of fish meal and fish oil [28]. This has led researchers to explore numerous other potential sustainable sources, such as SCP, to replace fish meal in aquaculture feeds. SCP refers to microbial biomass produced for use as a protein supplement for animal consumption, typically made from dried wastewater aerobic sludge and is seen as one of the more economical, faster-growing, and easy to produce alternatives [8], [28], [30]. Early research of SCP production found that heavy metals and poor product preparation tended to limit usability [45], but recent developments focusing on food-processing wastewater have indicated better production and feed performance, especially for brewery wastewater facilities that have high levels of carbon (as chemical oxygen demand (COD)) (Chapter 2). Past work in SCP manufacture has indicated that increasing amounts of nitrogen and phosphorus as well as reduced SRT and increased F/M ratio can increase the protein content in SCP [23], presumably by increasing the amount of rapidly dividing young cells and increased expression of growth related enzyme systems. However, the proteins associated with this kind of novel feed product and the metabolisms that encourage protein production can be highly dependent on the wastewater substrate used. Previously, we have presented results from a pyrotag sequencing study of the microbial consortia of a pilot scale aerobic bioreactor treating brewery wastes which indicated that differential enrichment of rhizospheric diazotrophic bacteria occurred as the primary microbial treatment response (Chapter 4). Figure 5.1 shows this trend in more detail, with 3 distinct taxonomic groups of root-zone associated free-living diazotrophs alternating in time. As a fraction of all sequences, this group of organisms formed the most abundant enrichment of organisms by the pilot reactor over
Figure 5.1. Distribution of diazotrophic sequences as a fraction of all sequences over the study period in the pilot reactor.

the study period. Another major group of organisms commonly found in the pilot bioreactor were saccharolytic fermenters such as Prevotella that migrated from the upstream acidogenic basin. Together, these two groups accounted for the majority of sequences isolated from the pilot bioreactor. A major concern was that since the pilot bioreactor showed signs of anaerobic and aerobic activity, it is not clear as to what kind metabolic activity accounts for the majority of proteins actually in the final product as diazotrophic organisms generally have shown the ability for nitrogen-fixation, denitrification, and aerobic respiration [182]. Some of the common features of model organisms from these genera are shown in Table 5.1. Affinity for organic acid substrates, ammonium as nitrogen source, elevated temperatures, and neutral pH are common attributes of detected organisms.
Table 5.1. Common characteristics of organic acid utilizing diazotrophs and acidogenic bacteria detected in the pilot bioreactor. Substrate utilization denoted by: (-) no utilization, (x) at least one substrate by one species, (xx) multiple substrates with multiple species, (xxx) most substrates with most species, (n.r.) not reported.

<table>
<thead>
<tr>
<th>Metabolism</th>
<th>Azonexus</th>
<th>Telmatospirillum</th>
<th>Mesorhizobium</th>
<th>Azospirillum</th>
<th>Prevotella</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strict Aerobe</td>
<td>Anaerobic, Microaerophilic</td>
<td>Strict Aerobe</td>
<td>Facultative Aerobe</td>
<td>Saccharolytic Fermentation</td>
</tr>
<tr>
<td>Associated Environment</td>
<td>Grass Rhizosphere and Wastewater</td>
<td>Bogs, Fens</td>
<td>Root Nodules</td>
<td>Grass Rhizosphere</td>
<td>Oral Cavity</td>
</tr>
<tr>
<td>Plant Associated</td>
<td>Root-associated</td>
<td>n.r.</td>
<td>Root Nodules</td>
<td>Root-Associated</td>
<td>No</td>
</tr>
<tr>
<td>N-Fixation</td>
<td>Microaerobic</td>
<td>Microaerobic</td>
<td>Nodules Only</td>
<td>Microaerobic</td>
<td>No</td>
</tr>
<tr>
<td>N-Fixation in Plants</td>
<td>Sometimes</td>
<td>n.r.</td>
<td>Always</td>
<td>Sometimes</td>
<td>No</td>
</tr>
<tr>
<td>Denitrification</td>
<td>x</td>
<td>n.r.</td>
<td>n.r.</td>
<td>xx</td>
<td>No</td>
</tr>
<tr>
<td>Carbon Substrates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xxx</td>
</tr>
<tr>
<td>Saccharides</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>xxx</td>
</tr>
<tr>
<td>Organic Acids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xxx</td>
</tr>
<tr>
<td>Alcohols</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>xx</td>
</tr>
<tr>
<td>Nitrogen Source</td>
<td>Ammonium</td>
<td>n.r.</td>
<td>Ammonium, Nitrate, Urea, Amino Acids</td>
<td>Ammonium, Nitrate, Nitrite, Amino Acids</td>
<td>Peptides</td>
</tr>
<tr>
<td>Maximal Temperature Range</td>
<td>20-40</td>
<td>4-30</td>
<td>4-42</td>
<td>33-41</td>
<td>25-45</td>
</tr>
<tr>
<td>pH Range</td>
<td>neutral</td>
<td>4-7</td>
<td>4-10</td>
<td>5.5-7.5</td>
<td>4.5-7.5</td>
</tr>
<tr>
<td>Additional Sources</td>
<td>[216–219]</td>
<td>[220]</td>
<td>N/A</td>
<td>[221-222]</td>
<td>N/A</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>Betaproteo</td>
<td>Alphaproteo</td>
<td>Alphaproteo</td>
<td>Alphaproteo</td>
<td>Bacteroidetes</td>
</tr>
</tbody>
</table>
The flexible lifestyle of root-zone diazotrophs means that the ability to understand changes in protein content due to environmental conditions and the ability to replicate this technique to other sites is largely undeveloped, particularly given the "black box" nature of conventional wastewater treatment. The advent of 'omics' technology allows for unique analysis of environmental systems, especially metaproteomics, the study of proteins isolated directly from the environment [213], [214]. The use of metaproteomics in this type of study is two-fold. The environmental proteome characterizes the proteins in the wastewater mix being used as feed, but also represents the dominant metabolic pathways that can be an indicator as to how to increase productivity by encouraging these pathways.

5.3 Proteins in the Wastewater, Feed, and Biotechnology Sectors

Standard methods have existed for decades in the fields of biochemistry, feed science, and wastewater engineering for the measurement of proteins, though their crossover application to SCP produced from brewery wastewater has not been straightforward.

5.3.1 Feed Industry Practice

In the feed industry the value of a protein source often is directly related to the amount of protein by standardized industry protein quantification techniques. Because of the wide range of proteins, characterization and enumeration of proteins can quickly become a separation and quantification task requiring great analytical chemistry expertise and funding. Therefore, the primary focus of the industry has been to develop robust methods for assaying a wide array of feed sources and types, namely the crude protein assay. Crude protein measurements are based on an assay of Kjeldahl nitrogen content with a conversion factor (6.25x) to estimate the amount of protein from the nitrogen content measurement [49]. However, these measurements can be interfered with by non-protein nitrogen sources (at the heart of the melamine tainted pet food and baby’s milk scandals from China [215]) and there is evidence that non-animal sources of protein have differing protein to nitrogen ratios [216]. In addition to crude protein measurements, amino acid
Table 5.2. Different types of fish meal replacement alternatives and their limiting amino acids (LAA, right column). Reproduced from [28]

<table>
<thead>
<tr>
<th>Feedstuff</th>
<th>THR</th>
<th>VAL</th>
<th>MET</th>
<th>CYS</th>
<th>ISO</th>
<th>LEU</th>
<th>PHE</th>
<th>TYR</th>
<th>LYS</th>
<th>HIS</th>
<th>ARG</th>
<th>TRY</th>
<th>LAA&lt;sup&gt;4&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow lupin</td>
<td>66</td>
<td>81</td>
<td>20</td>
<td>126</td>
<td>117</td>
<td>125</td>
<td>85</td>
<td>94</td>
<td>64</td>
<td>117</td>
<td>192</td>
<td>135</td>
<td>Met</td>
</tr>
<tr>
<td>Haricot bean</td>
<td>80</td>
<td>103</td>
<td>43</td>
<td>67</td>
<td>120</td>
<td>121</td>
<td>118</td>
<td>83</td>
<td>92</td>
<td>127</td>
<td>104</td>
<td>129</td>
<td>Met</td>
</tr>
<tr>
<td>Cottonseed</td>
<td>65</td>
<td>102</td>
<td>52</td>
<td>118</td>
<td>92</td>
<td>94</td>
<td>122</td>
<td>89</td>
<td>52</td>
<td>117</td>
<td>205</td>
<td>141</td>
<td>Lys</td>
</tr>
<tr>
<td>Sunflower</td>
<td>65</td>
<td>124</td>
<td>83</td>
<td>137</td>
<td>115</td>
<td>104</td>
<td>109</td>
<td>91</td>
<td>42</td>
<td>119</td>
<td>159</td>
<td>165</td>
<td>Lys</td>
</tr>
<tr>
<td>Sesame</td>
<td>58</td>
<td>98</td>
<td>109</td>
<td>148</td>
<td>91</td>
<td>105</td>
<td>86</td>
<td>114</td>
<td>33</td>
<td>114</td>
<td>211</td>
<td>153</td>
<td>Lys</td>
</tr>
<tr>
<td>Coconuts</td>
<td>65</td>
<td>114</td>
<td>61</td>
<td>96</td>
<td>115</td>
<td>112</td>
<td>95</td>
<td>92</td>
<td>37</td>
<td>81</td>
<td>217</td>
<td>123</td>
<td>Lys</td>
</tr>
<tr>
<td>Groundnut</td>
<td>55</td>
<td>99</td>
<td>39</td>
<td>133</td>
<td>117</td>
<td>100</td>
<td>107</td>
<td>117</td>
<td>53</td>
<td>100</td>
<td>196</td>
<td>141</td>
<td>Met</td>
</tr>
<tr>
<td>Rapeseed</td>
<td>93</td>
<td>118</td>
<td>83</td>
<td>70</td>
<td>113</td>
<td>116</td>
<td>94</td>
<td>77</td>
<td>74</td>
<td>131</td>
<td>112</td>
<td>159</td>
<td>CyS</td>
</tr>
<tr>
<td>Soybean</td>
<td>74</td>
<td>101</td>
<td>46</td>
<td>130</td>
<td>128</td>
<td>115</td>
<td>105</td>
<td>97</td>
<td>76</td>
<td>106</td>
<td>123</td>
<td>176</td>
<td>Met</td>
</tr>
<tr>
<td>Corn gluten meal</td>
<td>61</td>
<td>105</td>
<td>88</td>
<td>113</td>
<td>137</td>
<td>244</td>
<td>133</td>
<td>71</td>
<td>22</td>
<td>92</td>
<td>55</td>
<td>56</td>
<td>Lys</td>
</tr>
<tr>
<td>Pea seed</td>
<td>83</td>
<td>124</td>
<td>49</td>
<td>66</td>
<td>143</td>
<td>124</td>
<td>124</td>
<td>-</td>
<td>86</td>
<td>127</td>
<td>113</td>
<td>122</td>
<td>Met</td>
</tr>
<tr>
<td>Potato protein</td>
<td>89</td>
<td>125</td>
<td>63</td>
<td>96</td>
<td>128</td>
<td>120</td>
<td>112</td>
<td>149</td>
<td>74</td>
<td>73</td>
<td>73</td>
<td>118</td>
<td>Met</td>
</tr>
<tr>
<td>Leaf protein</td>
<td>84</td>
<td>127</td>
<td>57</td>
<td>56</td>
<td>112</td>
<td>120</td>
<td>122</td>
<td>129</td>
<td>71</td>
<td>90</td>
<td>96</td>
<td>141</td>
<td>CyS</td>
</tr>
<tr>
<td>Algal SCP (Spirulina)</td>
<td>87</td>
<td>136</td>
<td>52</td>
<td>30</td>
<td>159</td>
<td>118</td>
<td>105</td>
<td>123</td>
<td>55</td>
<td>75</td>
<td>111</td>
<td>165</td>
<td>CyS</td>
</tr>
<tr>
<td>Yeast SCP (Bakers)</td>
<td>93</td>
<td>116</td>
<td>63</td>
<td>85</td>
<td>139</td>
<td>112</td>
<td>91</td>
<td>108</td>
<td>86</td>
<td>106</td>
<td>89</td>
<td>141</td>
<td>Met</td>
</tr>
<tr>
<td>Yeast SCP (Torula)</td>
<td>94</td>
<td>118</td>
<td>54</td>
<td>81</td>
<td>144</td>
<td>98</td>
<td>137</td>
<td>117</td>
<td>84</td>
<td>104</td>
<td>86</td>
<td>118</td>
<td>Met</td>
</tr>
<tr>
<td>Bacterial SCP</td>
<td>97</td>
<td>134</td>
<td>89</td>
<td>59</td>
<td>115</td>
<td>107</td>
<td>115</td>
<td>138</td>
<td>71</td>
<td>83</td>
<td>84</td>
<td>118</td>
<td>CyS</td>
</tr>
<tr>
<td>Meat meal</td>
<td>73</td>
<td>124</td>
<td>61</td>
<td>113</td>
<td>109</td>
<td>111</td>
<td>89</td>
<td>69</td>
<td>90</td>
<td>94</td>
<td>145</td>
<td>94</td>
<td>Met</td>
</tr>
<tr>
<td>Meat &amp; bone meal</td>
<td>77</td>
<td>128</td>
<td>59</td>
<td>89</td>
<td>109</td>
<td>113</td>
<td>88</td>
<td>60</td>
<td>86</td>
<td>100</td>
<td>150</td>
<td>88</td>
<td>Met</td>
</tr>
<tr>
<td>Blood meal</td>
<td>69</td>
<td>158</td>
<td>33</td>
<td>52</td>
<td>24</td>
<td>162</td>
<td>124</td>
<td>69</td>
<td>89</td>
<td>214</td>
<td>62</td>
<td>123</td>
<td>Ils</td>
</tr>
<tr>
<td>Liver meal</td>
<td>76</td>
<td>135</td>
<td>72</td>
<td>89</td>
<td>105</td>
<td>121</td>
<td>109</td>
<td>106</td>
<td>71</td>
<td>98</td>
<td>105</td>
<td>153</td>
<td>Lys</td>
</tr>
<tr>
<td>Poultry byproduct meal</td>
<td>67</td>
<td>125</td>
<td>81</td>
<td>141</td>
<td>132</td>
<td>123</td>
<td>80</td>
<td>60</td>
<td>71</td>
<td>87</td>
<td>134</td>
<td>112</td>
<td>Tyr</td>
</tr>
<tr>
<td>Feather meal</td>
<td>91</td>
<td>164</td>
<td>24</td>
<td>289</td>
<td>131</td>
<td>124</td>
<td>78</td>
<td>86</td>
<td>33</td>
<td>50</td>
<td>147</td>
<td>76</td>
<td>Met</td>
</tr>
<tr>
<td>Earthworm meal</td>
<td>107</td>
<td>99</td>
<td>106</td>
<td>52</td>
<td>112</td>
<td>124</td>
<td>84</td>
<td>108</td>
<td>79</td>
<td>125</td>
<td>98</td>
<td>82</td>
<td>CyS</td>
</tr>
</tbody>
</table>

<sup>4</sup>Chemical scores based on comparison with the mean essential amino acid requirements of rainbow trout and common carp (Ogino, 1980). Mean EAA requirement (expressed as % of total EAA) being: threonine 10.6, valine 9.5, methionine 5.4, cystine 2.7, isoleucine 7.5, leucine 13.5, phenylalanine 9.5, tyrosine 6.5, lysine 16.8, histidine 4.8, arginine 11.6 and tryptophan 1.7. Adapted from Tacon (1992). Numbers marked in italics represent limiting essential amino acids (present below 30% mean fish requirement).

analysis is also an industry standard used to determine amino acid ratios of feeds. Modern feeds are commonly formulated based on standardized nutrient needs of each species and age class. Due to the lack of certain essential amino acids in alternative protein sources (methionine, cysteine (Table 5.2)) careful monitoring of these amino acids is needed for feed development. Therefore, methods for high-pressure liquid chromatography (HPLC) techniques with reverse-phase separation and colorimetric or fluorometric derivitization have been used to quantify amino acids in feeds [49]. Such amino acid techniques are rarely applied to wastewater sludge however, except for early
work on SCP sludge which showed that the amino acid profile was comparable to fish meal [23].

5.3.2 Laboratory Detection of Protein

In laboratory settings, extraction of proteins from pure cultures and tissues is quite common, often for further downstream analysis such as SDS-PAGE. Colorimetric estimation of protein content is then often compared with a standard such as bovine serum albumin (BSA) to give results of ‘true’ protein as BSA. A notable array of techniques exists for a variety of applications, but most of these methods involve colorimetric and fluorometric modifications of active groups (often amine groups or peptide bonds). Therefore, they suffer from chemical interferences without proper purification steps. Furthermore, as these are laboratory assays, they have been developed with laboratory interferences in mind, usually from pure culture cell disruption products such as surfactants, polyphenols, and simple carbohydrates. These protein assays have been applied with mixed results in the far more complex wastewater and natural water settings. Some studies do not report interference (particularly after fractionation) [217] or find that interference is negligible [218], while others report difficulty in getting reproducible results [219]. Several studies have attempted to correlate interferences from substances such as humic acids with control standards [230-231]. Others have noted that humic acids and proteins can commonly become bound together and are difficult to separate [232-233]. Particularly fluorometric quantification methods suffer from interference from aromatic species (such as humics), while only being sensitive to amino acid groups with aromaticity (namely, tryptophan, phenylalanine, and tyrosine). What these disparate results likely indicate is that wastewater originating from different sources cannot be considered a single group and that interfering compounds are highly specific for each waste stream [224]. More recently, it has become common to characterize the dissolved organic matter constituent of wastewater and natural water using fluorospectroscopic excitation-emission matrix (EEM) spectra. Though not quantitative, the fluorometric signature of different carbon constituents such as proteins, fulvic-like, and humic-like substances can be determined by charting their fluorescent excitation and
emission wavelengths [17], [225]. A number of these methods were tested for this study and their result is shown in Table A1.

5.3.3 Fractionation of Proteins

As a result of this heterogeneity, two approaches to examining wastewater proteins can be taken. The first is to attempt to measure wastewater proteins with a method not specifically sensitive for proteins, such as by native fluorescence or spectroscopy. The second approach is to fractionate with separation chemistry techniques while facing recovery losses. Using analytical chemistry methods, fractions can be separated by HPLC and collected for different types of quantification. Much work has been done in the field of melanoidin research from food analytical chemistry, and the extraction of extracellular polymeric substances (EPS) from wastewater. Melanoids are water-soluble polymers of covalently linked amino acid and carbohydrate fractions that are primarily color and flavor molecules found in roasted food products such as malt and coffee [226] and are notoriously difficult to treat in rum distillery wastes [42]. They also bear remarkable similarity to natural humic substances [227]. Their structure is still debated and fractions of melanoids separate only after significant column chromatography [228]. Ultracentrifugation techniques are possible with melanoids [229] but require access to ultra-centrifuges. Dialysis is another possibility for fractionation [240-241]. From studies of EPS, cation exchange has also been used extensively to fractionate proteins from wastewater sludges followed by high-pressure size exclusion chromatography or ultrafiltration [217], [221] which allows for the study of proteins isolated from the extracellular matrix.

Another path to protein separation has been developed through biochemistry techniques of isolating protein fractions from environmental sources such as soils and natural waters. A number of cleanup steps can be used such as bead-beating and sonication to break apart organic matter and extraction solvents such as phenol–chloroform to remove organics, TCA precipitation to remove humic acids and nucleic acids, Methanol–Chloroform extraction to remove salts, and centrifuge filters or dialysis filters with molecular weight cutoff limits to remove large molecules. Total recovery
efficiency is likely to be low with these methods as multiple steps are often required [14]. Once purified, gel electrophoresis techniques such as SDS-PAGE and 2D-PAGE have been shown to adequately separate a wide variety of proteins. These chromatographic techniques fractionate proteins based on molecular weight (SDS-PAGE), or isoelectric point and molecular weight (2D-PAGE). 2D-PAGE has been used for years on pure culture and tissue samples to study differential protein expression in biochemical studies. Finally, from the rapidly expanding field of proteomics, environmental or metaproteomes are becoming common [214] though a major issue is the fact that few extraction protocols exist for the fractionation of proteins from environmental samples for these types of studies [213], and a lack of sensitivity at high levels of microbial diversity [214].

5.3.4 Identification of Proteins

After fractionation, mixtures of proteins isolated from a number of different environmental sources can be identified by mass spectroscopy techniques based on predicted mass fragment profiles. Though powerful and specific, a major weakness of mass spectra identification of peptides is the need for carefully constructed predicted protein spectra from genomic sequencing sources, a resource notably absent for environmental samples. Given that these wastewater systems are highly customized complex mixtures, in this paper two approaches are taken. First a holistic examination of the wastewater carbon fraction is presented. Next, fractionation specifically for wastewater proteins is done followed by analysis of proteins by proteomic techniques.

5.4 Materials and Methods

The materials and methods were wide ranging and incorporated disparate analysis from the feed industry, wastewater engineering, and molecular biology areas.

5.4.1 Overview of Production and Study Site

Figure 5.2 shows the brewery Wastewater Treatment Plant (WWTP) and pilot reactor research site used in this study (New Belgium Brewing, Ft. Collins, Colorado). Separate from the main aerobic basin, the pilot reactor (Oberon FMR) received wastewater
Table 5.3. Standard industry protein analysis techniques

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude Protein Combustion</td>
<td>AOAC 990.03</td>
<td>Nitrogen freed by combustion in pure oxygen is measured by thermal conductivity detection and converted to equivalent protein by numerical factor</td>
</tr>
<tr>
<td>Method</td>
<td>[49]</td>
<td></td>
</tr>
<tr>
<td>Crude Protein Cu Kjeldahl</td>
<td>AOAC 984.13 / SM</td>
<td>Using CuSO4 as catalyst, convert all N to NH3 in H2SO4 and measure by colorimetric technique and convert to equivalent protein by numerical factor</td>
</tr>
<tr>
<td>Method</td>
<td>4500-Norg B [49],[232]</td>
<td></td>
</tr>
<tr>
<td>Amino Acid HPLC Method</td>
<td>AOAC 994.12</td>
<td>Digest sample in trypsin and separate by HPLC with ninhydrin derivatization and subsequent colorimetric detection of eluent based on an amino acid standard suite.</td>
</tr>
</tbody>
</table>

Figure 5.2. Overview of the study site showing the brewery treatment works (top) and the pilot scale reactor (bottom). Colored squares indicate sampling location and are used to illustrate figures throughout this chapter.
from the acidogenic basin and attempted to aerobically treat the wastewater at low Mean Cell Residence Time (MCRT) (< 8 days) with nutrient addition (N as urea, P as phosphoric acid, and a customized micronutrient cocktail) to increase the protein concentration. The pilot reactor remained online for more than one year and consistently producing product for commercial feed trials [174], except for a 2-week brewery shutdown in December of 2008 (Day 280 of the study period) which disrupted production operations.

5.4.2 Crude Protein and Amino Acid Profiles

Crude protein was measured on a LECO FP 528 (LECO, St. Joseph, MI) with EDTA and lysine standards by a commercial testing company (Industrial Labs, Golden, CO). For amino acid analysis, SCP samples were hydrolyzed and derivatized, and then separated on a HPLC. Dried SCP was hydrolyzed to free amino acids (AOAC method 994.12) [49] with 6 N HCl solution (containing 1% phenol) was added and mixed well and incubated at 110 °C for 24 hr. L-Norleucine was added as an internal standard. The supernatant was combined with methanol/water/TEA (2.2:1; v/v) and mixed well. The sample was vacuum-dried derivatized by PITC [233] [methanol, PITC, TEA, water (7:1:1:1; v/v)] and then vacuum-dried. The derivatized amino acids were separated with a Beckman System Gold HPLC equipped with a 166-diode array detector and a 508 autosampler (Fullerton, CA). Elution buffer A: 50 mM sodium acetate (pH 5.45); buffer B: 70% acetonitrile and 32 mM sodium phosphate (pH 6.10). A linear gradient was used where buffer B was increased from 7 to 60% over 30 min, then decreased from 60 to 7% over 10 min. The derivatized samples were redissolved in 1.5 mL of a buffer solution (buffer A/buffer B 93:7) that was used as the initial mobile phase for HPLC on a C18 reverse-phase amino acid column (5 μm, 250 X 4.6 mm; Supelco, Sigma Inc, St. Louis, MO). Derivatives were detected by UV absorbance at 254 nm. Because tryptophan and cystine/cysteine are destroyed by acid hydrolysis, alkaline hydrolysis and performic acid oxidation of the protein samples were conducted for analysis of tryptophan and cystine/cysteine content, respectively, according to AOAC method 988.15 and method 985.28 [49]. The derivatization and HPLC separation procedures for alkaline hydrolyzed
samples were the same as for acid hydrolyzed samples. Amino acid profiles were run by the Oceanic Institute (Waimanalo, HI).

5.4.3 Excitation-Emission Spectra

Fluorescence excitation emission matrix (EEM) spectra were recorded for each sample using a Horiba JobinYvonFluoroMax 4 fluorometer (Edison, NJ, USA). Scans were conducted using excitation from 240 to 450 nm at 5 nm steps, emission from 290 to 580 nm at 4 nm steps, 2 nm bandwidth and 0.25 s integration time. The intensity of all EEM spectra was normalized by dividing the integrated intensity area of the Raman water line at 350 nm excitation from the 370 to 450 nm emission range. Data analysis included a correction due to water-scattering by subtracting the signal of a blank MilliQ water sample analyzed under the same conditions. Three-dimensional contour plots were created by plotting fluorescence intensity as a function of emission (x-axis) and excitation (y-axis) wavelengths.

5.4.4 Fractionation of Proteins for 2D-PAGE

Cleanup of samples is critical in order to produce proper 2D-PAGE images. A major concern in the analysis of beer proteins is the presence of polyphenols which influence haze and taste in beer [39]. In brewery wastewater fractions similar substances are noted from tank bottom wastes [234]. In the brewing industry, such materials are precipitated by Polyvinylpyrrolidone (PVPP), an inert hydroscopic branched polymer [235], and have been used as a cleanup step for 2D-PAGE proteomic analysis of beer [236]. Such polyphenols and are thought to also be common in plants and a concern for plant protein extraction. Therefore, PVPP has been used to remove polyphenols during protein extraction from plant tissues [237] and to purify DNA from humic acid in wastewater [238]. The exact mechanism of PVPP binding is not well understood, but is thought to be a hydrogen bonding effect with polyphenols, and tannins in particular [239]. Based on the similarity in structure of other phenol and quinone-containing compounds such as humic acids and melanoidins from brewery wastes, a PVPP-based protein extraction protocol was developed to produce clearer 2D-PAGE gels with significantly less smearing. 0.14 g of wet cell mass spun at 10,000 RCF, 10 min or 0.03 g of dried SCP
product was added to 1.5 ml denaturing buffer (7 M Urea, 2 M Thiourea, 4% CHAPS, 10 mM Tris, 1 mM EDTA, 50 mM Dithiothreitol (DTT), and 1 tablet Complete protease inhibitor cocktail, mini (per 10 ml) (Roche, Basel, Switzerland)) lysed by sonication (Qsonica, Newton, CT) with 6 second pulses and then incubated for 20 min at room temperature. 400 ul spun supernatant (13,000 RCF, 10 min) was taken for TCA precipitation. 88 ul (20% (v/v) final) TCA was added and briefly vortexed and incubated on ice for 1 hour. The resulting solution was washed in 800 ul cold (-20 °C) acetone, and incubated at -20 °C for 30 min. The samples were centrifuged (13,000 RCF, 10 min) and the supernatant discarded. This was repeated once more with incubation at -20 °C for 10 min. The resulting pellet was allowed to air dry in a clean BSL 2 cabinet. The pellet was resuspended in 3 ml PVPP treatment buffer (50 mM DTT, 10 mM Tris, 1 mM EDTA, 1 tablet Complete protease inhibitor cocktail, mini (per 10 ml)) and 0.05 g PVPP and stirred overnight at room temperature. The solution was then filtered on filter paper (VWR, Qualitative 413) using baked glassware. The filtrate was collected and precipitated with equal volume methanol, and 1/4 volume chloroform. After vortexing, the solution was centrifuged (13,000 RCF, 10 min) and a protein layer became visible at the interface. The top layer was extracted and the remaining volume mixed with another aliquot of methanol and centrifuged (13,000 RCF, 10 min) to pellet the precipitate. The pellet was again allowed to dry and resuspended in 150ul DeStreak Rehydration Solution (GE Healthcare, Buckinghamshire, UK). IPG strips (7 cm, 3–10 pH) were rehydrated overnight using this solution in an IPGbox (GE Healthcare). Isoelectric focusing (300 V, 30 min, gradient 1000 V, 30 min, gradient 5000 V, 1:20 hr, 5000 V, 25 min) was done on an Etta IPGphor II (GE Healthcare), followed by incubation in reducing buffer 15 min (500 ul NuPAGE Sample Reducing Agent, 4.5 ml 1X NuPAGE LDS Sample Buffer (Invitrogen, Carlsbad, CA)), and alkylating buffer 15 min (116 mg Iodoacetamide, 5 ml 1X NuPAGE LDS Sample Buffer). The second dimension was run on a Noyex 4-20% Tris-Glycine ZOOM Gel (Invitrogen) on an XCell SureLock Mini-Cell (Invitrogen) at 125 V for 90 min, and stained with Coomassie Blue (PageBlue, Fermentas Thermo Fisher Scientific, Waltham, MA) and imaged.
5.4.5 LC-MS/MS Peptide Digest Preparation

All plasticware used for peptide digestion were washed 3x with a solution of 50% Acetonitrile, 5% Formic acid (1 hour) with brief vortexing. All reagents were LC grade where available. For peptide identification the largest spots seen in gels were excised and treated with 200 ul wash solution (50% (v/v) methanol, 45% milliQ water, 5% acetic acid) overnight at room temperature. A second wash was done for 2 hours. The gel pieces were dehydrated with 200 ul acetonitrile for 5 min and supernatant removed. The pieces were then washed with 200 ul 100 mM ammonium bicarbonate for 10 min and supernatant removed, and again dehydrated with acetonitrile. The pieces were dried in a BSL 2 cabinet and then rehydrated with 30 ul ice-cold trypsin solution (1 ml 50 mM ammonium bicarbonate, 20 ug sequencing grade trypsin (Promega, Fitchburg, WI)) for 10 min on ice and the supernatant removed. 5 ul 50 mM ammonium bicarbonate was added, and the pieces were then digested overnight at 37 °C. The digested peptides were extracted with first 30 ul 50 mM ammonium bicarbonate, (10 min, occasional vortexing) and the supernatant added to a separate tube. Then 30 ul of extraction buffer (50% (v/v) Acetonitrile, 45% milliQ water, 5% Formic Acid) and the supernatant added to the new tube. This step was repeated once more, and the total collected supernatant dried down by vacuum evaporation for downstream LC-MS/MS. Samples were reconstituted in LC loading buffer and separated by a C18 reverse phase column prior to MS/MS on an Agilent 6520 QTOF with Chipcube interface. Peptide LC-MS/MS and data interrogation was completed by the Colorado State University proteomics core facility (Ft. Collins, CO) using SpectrumMill (Agilent).

5.4.6 Peptide Identification

Detected MS/MS spectra were matched for identification to the NCBI-nr-microbe (a database of non-redundant protein-coding amino acid sequences) and Swiss-Prot [240] databases. Unlike other metaproteome studies that attempted to identify minute fractions of a metaproteome based on a metagenome, this study attempts to identify only the most common fraction of proteins and identify ways of promoting their production. Each protein spot can contain numerous variations of the same protein arising from a multitude of organisms. Due to the selectivity of MS/MS spectra, even proteins closely matched
with reference proteins might not produce strong spectrum matching. Therefore, a consensus approach was taken where low scoring peptides matching general classes of proteins are grouped together. Individual peptides that were found common to most of these protein matches in each group were compared with the Clusters of Orthologous Groups (COG) database [241]. The advantage of COG genes for homology analysis is that they are categorized into several thousand different functional groups or families across a number of genomes spread over the three domains of life. First, each unknown peptide was identified with its BLAST COG family of closest match. Next, all genes in that COG family were combined with their respective identified peptides and aligned using Clustal Omega [242]. Two criteria were evaluated to see how well individual peptides served to properly identify proteins. First, if a protein was present in a sample, it was likely a mixture of a number of different homologous proteins of the same function from a variety of microbes. The more unique spectra of the same protein family that were detected, the more likely the protein was in the sample. This was reported as the fraction of spectra of a protein family to all spectra detected for each spot. The second criteria assumed that peptides identified without a metagenome are likely of conserved domains within similar proteins. To test this, the percent conservation of the most common amino acid for each alignment position was calculated from alignment files of all COG families that had an identified peptide matched to it. Each identified peptide was then assigned an average peptide conservation (APC) score by averaging the conservation score of each alignment position of that unknown peptide from the COG family alignment file. As a comparison metric, the average conservation scores of a random subsample (n=1000) of hypothetical peptides of the same length were then taken from the same COG protein family and shown as a boxplot distribution of average peptide conservation scores. Any single protein match that has a SpectrumMill MS/MS search score greater than 90 (usually represented by more than 4 quality peptide matches to the same protein) was considered a strong match and exempted from the analysis.

5.4.7 Organic Acid Analysis

An ion exclusion organic acid column Animex HPX-87H (BioRad, Hercules, CA) with a pre-column filter (BioRad) was used with an Agilent 1100 HPLC for organic acid
analysis. Samples were spun (10,000 RCF, 5 min) and 0.45 um syringe-filtered before use. Standards for formic, acetic, lactic, propionic, succinic, butyric, isobutyric, valeric, isovaleric, 2-methylbutyric, and citric acid were run at three intervals in replicates of three with an injection volume of 50 um. Running buffer consisted of 0.04 N Phosphoric Acid, 0.60 ml / min, 40 C, 35 min. DAD signal frequency was 210 nm, reference 360 nm.

5.5 Results

The results are split into two parts, holistic analysis results of organic acids and other bulk properties, and peptide fractionation and identification results.

5.5.1 Results from Holistic Analysis

Monitoring of organic acid content shows that the primary behavior of this pilot bioreactor was a switch from acetic acid consumption to acetic acid production, as well as the consumption of butyric, formic, and valeric acids over the study period (Figure 5.3). The influent acid content changed over time, with increasing butyric and propionic acid replacing falling acetic acid levels. Finally, acid levels on Day 365 show a distinct turnover, with several different organic acids either being produced or consumed for a short time and then resumption of normal activity. Figure 5.4 shows measured values of crude protein, total amino acid profile, and production rate over the study period. Crude protein content varied between 58-68% of total dry weight over the course of the study. Total amino acid profiles show rough, if not total agreement with crude protein results, at an amino acid to nitrogen ratio for SCP material of approximately 5.7 +/- 0.7 to total amino acids to nitrogen, slightly less than the protein to nitrogen ratio of 6.25 for animal tissues. Using EEMs (Figure 5.5) protein fluorescence is strongly observed in the 250-300 nm excitation, 300-375 nm emission range typical of microbial products [243]. Additionally, humic-like and melanoidin-like substances are observed in the higher regions (300-400 nm excitation, 400-500 nm emission) [225], [244] at a much lower intensity.
Figure 5.3. Organic acid composition of the Oberon pilot plant (solid lines) and acidogenic basin (dashed lines) by mass basis. Conductivity of mixed liquor is shown in the last panel.
Figure 5.4. Crude protein and total amino acid over the study period (top) and production (bottom), with 3-day moving averages.

Figure 5.5. Excitation-emission matrix of pilot bioreactor mixed liquor.
Figure 5.6. Amino acid (AA) composition over time, total amino acid basis.

5.5.2 Amino Acid Profiles

Analysis of overall amino acid variation shows fairly steady composition (Figure 5.6), but an increase in glutamate/glutamine content at one data point (Day 208). A Principle Component Analysis (PCA) ordination plot of un-scaled amino acid values (B) shows that samples clustered based on analysis date, which indicates that instrument variation or sample preparation were the major factors in variation, except for at one time point (Day 208) which corresponds to the glutamate/glutamine spike. This PCA ordination plot also shows a major cluster of aspartate/asparagine, alanine, and leucine as peaking behind the glutamate/glutamine spike.

5.5.3 2D-PAGE

The composition of the amino acids should be determined by the active synthesized proteins used in metabolism of the biomass in the pilot bioreactor. To study this problem, proteomic biochemistry techniques were used to fractionate and identify proteins from wastewater in an effort to identify metabolic function. 2D-PAGE gels of
Figure 5.7. Amino acid variation analysis, showing changes in amino acids from average (A), and a PCA ordination plot of variation (B) with samples in black (number indicating day of sampling) and amino acids in red. Instrument runs (A, B, C) are shown in grey.
three sample days (Day 175, 335, and 365) from the pilot bioreactor were run and dominant spots extracted for peptide identification by LC-MS/MS (Figure 5.8). Little similarity was observed between the three pilot bioreactor samples (A,B,C). Spot distribution and sizes are visibly fewer and showed less evenness in the pilot bioreactor compared to the influent material (E). Few shared spots exist between the influent and pilot bioreactor, indicating very little remains of the influent proteome in bioreactor gel images. SCP material and pilot bioreactor mixed liquor showed great similarity between gels (C and D), with SCP containing smaller spots in the same relative location, but with much wider fan of unresolved low molecular weight peptides or contaminants at the bottom of the gel, possibly from proteolysis during the centrifugation and drying process.

5.5.4 Peptide Identification

Figure 5.9–Figure 5.12 shows the corresponding peptide identification of spots from Figure 5.8. About half of all spots had no detectable proteins, but several of the largest spots resulted in multiple types of proteins being detected. Of the proteins that were detected, most individual proteins were only detectable at the single spectra level. Average detection intensities were much higher than the keratin background indicating that these were not necessarily contaminants. Additionally, protein homology in spots resulted in a high number of repeated or related spectra being grouped in LC-MS/MS post-processing results. When sequences of these spectra matches were matched by BLAST to their likely COG family and compared with random distributions of peptides, they reveal a high level of BLAST homology to their COG family and occur largely in the top or second quartiles of average conserved peptide scores. This showed that many (but not all) of the detected proteins are likely matched into the correct protein family.

When the taxonomic identification of proteins by spectra matching was compared to previous pyrotag results there were mixed results in corroborating taxonomic assignment to 16S SSU rRNA data. Contrary to expectations, no proteins from *Prevotella* were detectable despite a number of genomes for this group of model organisms available in protein matching databases. Peptides from *Megasphaera* sp. and *Azospirillum* sp., two organisms seen as enriched, were detectable at the single spectra level. However, an
Figure 5.8 2D-PAGE gels of fractionated proteins from pilot bioreactor Day 175 (A), Day 335 (B), Day 365 (C), SCP product Day 365 (D), plant influent Day 365 (E). Numbers mark spots selected for sequencing.
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Protein Family</th>
<th>Organism</th>
<th>MS/MS Search Score</th>
<th>Total Peptides</th>
<th>No. Peptides</th>
<th>% of Spot</th>
<th>Mean Spectral Intensity</th>
<th>Sequence</th>
<th>MS/MS Search Score</th>
<th>% SPI</th>
<th>BLAST COG</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Elongation factor Tu</td>
<td>Bradyrhizobium sp.</td>
<td>36.17</td>
<td></td>
<td>4</td>
<td>5.97</td>
<td>8.9E+04</td>
<td>GRLLAGINDGAGQSQDLLK</td>
<td>20.23</td>
<td>83.2</td>
<td>COG0050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glutamine synthetase</td>
<td>Azospirillum brasilense</td>
<td>34.67</td>
<td></td>
<td>6</td>
<td>8.96</td>
<td>5.2E+04</td>
<td>TKGYAHAVNNHVVYK</td>
<td>18.22</td>
<td>82.9</td>
<td>COG0174</td>
<td>75.5</td>
</tr>
<tr>
<td></td>
<td>Elongation factor Tu</td>
<td>Shigella flexneri</td>
<td>22.22</td>
<td></td>
<td>24</td>
<td>50.75</td>
<td>1.4E+05</td>
<td>EKIGVIEVEEGVK</td>
<td>22.22</td>
<td>94.4</td>
<td>COG0050</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Proplonyl-CoA carboxylase</td>
<td>Myxococcus xanthus</td>
<td>21.66</td>
<td></td>
<td>1</td>
<td>1.49</td>
<td>3.3E+04</td>
<td>IRAGGSDDIGVPAGVR</td>
<td>21.66</td>
<td>94.5</td>
<td>COG4799</td>
<td>61.8</td>
</tr>
<tr>
<td></td>
<td>Glutamine synthetase</td>
<td>Actinobacillus pleuro pneumoniae</td>
<td>20.54</td>
<td></td>
<td>21</td>
<td>31.34</td>
<td>1.5E+05</td>
<td>RKYSNTSPNTFANLKR</td>
<td>20.54</td>
<td>91.8</td>
<td>COG0174</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3A</td>
<td>Chemotaxis histidine kinase</td>
<td>Azospirillum brasilense</td>
<td>26.71</td>
<td></td>
<td>2</td>
<td>100.00</td>
<td>1.2E+04</td>
<td>HRLPAVPRQVIEETD FVR NRFFPSDDVLLVR</td>
<td>16.19</td>
<td>81</td>
<td>COG0643</td>
<td>45.7</td>
</tr>
<tr>
<td>3B</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4B</td>
<td>Aldehyde dehydrogenase</td>
<td>Streptomyces avermitilis</td>
<td>23.96</td>
<td></td>
<td>8</td>
<td>100.00</td>
<td>1.5E+04</td>
<td>QRGIDLYSLIK ARGIDLYSLIK NTKQVYHDLMmK KKYGGFAAHAPYIHY CNTWVR</td>
<td>13.59</td>
<td>82.1</td>
<td>COG1012</td>
<td>29.7</td>
</tr>
<tr>
<td>5</td>
<td>Aldehyde dehydrogenase</td>
<td>Deinococcus radiorumans</td>
<td>33.63</td>
<td></td>
<td>60</td>
<td>39.22</td>
<td>2.1E+05</td>
<td>DIKGVIEVEEGVK NTKQVYHDLMmK</td>
<td>20.19</td>
<td>98.4</td>
<td>COG1012</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>Elongation factor Tu</td>
<td>Desulfovibrio vulgaris DP4</td>
<td>24.88</td>
<td></td>
<td>93</td>
<td>60.78</td>
<td>2.0E+04</td>
<td>DIKGVIEVEEGVK NTKQVYHDLMmK</td>
<td>17.7</td>
<td>80.1</td>
<td>COG0050</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Figure 5.9. Peptide identification table for 2D-PAGE gel A. Protein matching software hits for proteins are shown by family and organism (left). Total protein consensus peptides are shown for all peptides detected (middle). Individual common peptide validation statistics are shown for each common peptide detected (right). Average peptide conservation (APC) is shown in the last column, with values matching the red level, and compared with the box plot distribution of randomly sampled peptide APC scores, in percent.
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Protein Family</th>
<th>Organism</th>
<th>MS/MS Search Score</th>
<th>No. Peptides</th>
<th>% of Peptides</th>
<th>Mean Spectral Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Enolase</td>
<td>Saccharomyces cerevisiae</td>
<td>72.13</td>
<td>14</td>
<td>3.46</td>
<td>3.08E+04</td>
</tr>
<tr>
<td></td>
<td>ATP synthase subunit beta</td>
<td>Bradyrhizobium sp.</td>
<td>26.34</td>
<td>130</td>
<td>32.10</td>
<td>2.54E+04</td>
</tr>
<tr>
<td></td>
<td>Elongation factor Tu</td>
<td>Brucella ovis</td>
<td>20.08</td>
<td>185</td>
<td>45.68</td>
<td>3.02E+04</td>
</tr>
<tr>
<td></td>
<td>Succinyl-CoA ligase</td>
<td>Rhodospirillum centenum</td>
<td>53.94</td>
<td>5</td>
<td>1.23</td>
<td>1.67E+05</td>
</tr>
<tr>
<td></td>
<td>Alcohol dehydrogenase</td>
<td>Saccharomyces cerevisiae</td>
<td>47.38</td>
<td>4</td>
<td>0.99</td>
<td>2.18E+04</td>
</tr>
<tr>
<td></td>
<td>Serpin Z.</td>
<td>Hordeum vulgare</td>
<td>47.38</td>
<td>3</td>
<td>0.74</td>
<td>2.09E+04</td>
</tr>
<tr>
<td></td>
<td>ATP synthase subunit alpha</td>
<td>Rhodopseudomonas palustris</td>
<td>47.38</td>
<td>51</td>
<td>12.59</td>
<td>3.99E+04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Protein Family</th>
<th>Organism</th>
<th>MS/MS Search Score</th>
<th>No. Peptides</th>
<th>% of Peptides</th>
<th>Mean Spectral Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ATP synthase subunit alpha</td>
<td>Bradyrhizobium sp.</td>
<td>86.63</td>
<td>31</td>
<td>41.89</td>
<td>1.44E+05</td>
</tr>
<tr>
<td></td>
<td>Aldehyde dehydrogenase</td>
<td>Streptomyces avermitilis</td>
<td>19.85</td>
<td>43</td>
<td>58.11</td>
<td>5.79E+05</td>
</tr>
<tr>
<td></td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aldehyde dehydrogenase</td>
<td>Streptomyces avermitilis</td>
<td>51.43</td>
<td>53</td>
<td>100.00</td>
<td>3.28E+05</td>
</tr>
<tr>
<td></td>
<td>Elongation factor Tu</td>
<td>Desulfovibrio vulgaris DP4</td>
<td>28.34</td>
<td>93</td>
<td>100.00</td>
<td>2.01E+04</td>
</tr>
<tr>
<td></td>
<td>Aldehyde dehydrogenase</td>
<td>Streptomyces avermitilis</td>
<td>27.85</td>
<td>11</td>
<td>100.00</td>
<td>2.75E+04</td>
</tr>
<tr>
<td></td>
<td>Glucosyl transferase</td>
<td>Rhizobium leguminosarum</td>
<td>12.7</td>
<td>1</td>
<td>9.09</td>
<td>2.64E+05</td>
</tr>
<tr>
<td></td>
<td>Glyceraldehyde 3-</td>
<td>Saccharomyces pastorianus</td>
<td>11.99</td>
<td>6</td>
<td>54.55</td>
<td>6.18E+03</td>
</tr>
</tbody>
</table>

Figure 5.10 Peptide identification table for 2D-PAGE gel B
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Protein Family</th>
<th>Organism</th>
<th>MS/MS Search Score</th>
<th>No. Peptides</th>
<th>% of Spectral Intensity</th>
<th>Mean Sequence Score</th>
<th>% SPI</th>
<th>BLAST COG</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.11 Peptide identification table for 2D-PAGE gel C
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Protein Family</th>
<th>Organism</th>
<th>MS/MS Score</th>
<th>% of Spectral Intensity</th>
<th>Sequence</th>
<th>MS/MS Score</th>
<th>% SPI BLAST COG</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Enolase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>167.39</td>
<td>46 88.46 1.09E+05</td>
<td>NKFDLDKGDK</td>
<td>11.08</td>
<td>74.3 KOG2670</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Phosphoglycerate kinase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>48.53</td>
<td>6 11.54 2.05E+04</td>
<td>EKAVPALSYK</td>
<td>12.38</td>
<td>73.6 KOG1967</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YKIPPLAVIR</td>
<td>7.99</td>
<td>61.8 KOG1967</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>Acyl-CoA dehydrogenase</td>
<td><em>Megasphaera elsdenii</em></td>
<td>93.96</td>
<td>6 66.67 6.55E+04</td>
<td>GRGEEGMNENAPVK</td>
<td>20.82</td>
<td>83.2 COG1960</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Serpin Z</td>
<td><em>Hordeum vulgare</em></td>
<td>40.97</td>
<td>3 33.33 3.00E+04</td>
<td>EKADGAGGGDIULV</td>
<td>14.51</td>
<td>77.6 KOG2392</td>
<td>25.9</td>
</tr>
<tr>
<td>3</td>
<td>Elongation factor Tu</td>
<td><em>Azobacteroides pseudotrichonymphae</em></td>
<td>54.36</td>
<td>16 100.00 5.15E+04</td>
<td>KRFMEVGTVVSK</td>
<td>15.78</td>
<td>89.1 COG0050</td>
<td>81.6</td>
</tr>
<tr>
<td>4</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Pyruvate decarboxylase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>193.78</td>
<td>31 100.00 5.77E+04</td>
<td>IRAYGDAAYANLEN</td>
<td>23.07</td>
<td>93.4 KOG1184</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ANGAWR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Enolase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>210.85</td>
<td>45 97.83 1.01E+05</td>
<td>AAAAAMBVLJLVA</td>
<td>13.3</td>
<td>75.1 KOG2670</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GLK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>GRETTELEVTPNBR</td>
<td>18.73</td>
<td>89.6 KOG2670</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Phosphoglycerate kinase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>378.11</td>
<td>74 100.00 1.33E+05</td>
<td>YKIPPLAVIR</td>
<td>17.7</td>
<td>94.2 KOG1367</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>YKIPPLAVIR</td>
<td>15.1</td>
<td>79.7 KOG1367</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>acetyl-CoA acetyltransferase</td>
<td><em>Clostridium tetani</em></td>
<td>21.08</td>
<td>1 100.00 6.19E+04</td>
<td>VKLNFAPVEVPIGAT</td>
<td>21.08</td>
<td>94.9 COG0183</td>
<td>39.2</td>
</tr>
<tr>
<td>12</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Glycerol dehydrogenase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>173.81</td>
<td>59 100.00 2.31E+05</td>
<td>VKalPCNNTCSAN</td>
<td>24.89</td>
<td>96.9 KOG0657</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SVIK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AKAAGTSSIPPIGNSAT</td>
<td>18.04</td>
<td>93.2 KOG0657</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>Glycerol dehydrogenase</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>282.08</td>
<td>75 100.00 4.57E+05</td>
<td>VKALPCNNTCSAN</td>
<td>25.94</td>
<td>98.4 KOG0657</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SVIK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AKAAGTSSIPPIGNSAT</td>
<td>15.72</td>
<td>90.7 KOG0657</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 5.12 Peptide identification table for 2D-PAGE gel E
abundance of Gammaproteobacteria proteins were detected where no
Gammaproteobacteria were found in previous work. A number of Rhizobium sp. and
Rhodospirillum sp. proteins were detected and relate taxonomically only as close as the
order level to organisms in Table 5.1. Similarly, the presence of Gammaproteobacteria
and other clinical or model bacteria such as Brucella sp. indicates a high degree of
sequence database bias (from genome sequencing selection bias of model organisms).
Surprisingly, serpins derived from hops (H. vulgarae), documented as being thermo-
stable in the brewing process [236], were detected in wastewaters at a single spectra
level. Finally, yeasts and yeast proteins, being the most well studied strain in the system,
produced the most reliable results in this study and served as verification for methods
development. Overall, the identification of protein function tended to be reliable, but
identification of organisms from single spectra proved to be inconsistent, though in
certain cases matches were highly reminiscent of results from pyrotag studies of the same
material. The fact that proteins were detectable when public database genomes existed
indicates that in some cases, engineering proteomics without metagenome database
support is possible, particularly for industrial applications. But this strategy fails when
no organisms in a sample have sequenced genome representatives, as in the case of study
day 365.

The identification of protein functions based on consensus peptide detection resulted
in consistent results for shared spots (A4/B5-6, A5/B7, B9/114), though these were never
the dominant spots. These spots were typically of proteins related to growth (elongation
factors and ATP synthase) and organic acid uptake (aldehyde dehydrogenase, and
succinyl-CoA ligase). Identified influent proteins related to fermentation reactions (such
as phosphoglycerate kinase, and glyceraldehyde 3-phosphate dehydrogenase, and
enolase, all involved in glycolysis), indicating the presence active glycolysis or residual
yeasts. Alcohol dehydrogenases, phosphoglycerate kinases, and enolases were also
previously identified from yeast in beer [236] and were sometimes identified in the
bioreactor. At one time point glutamine synthetase was also identified, indicating
glutamine production from free ammonia. At no point was nitrogenase detected, despite
the primary association of diazotrophs with nitrogen fixation.
5.6 Discussion

The discussion relates the amino acid and protein variation seen in the results to what is known about diazotroph metabolism and reactor functioning.

5.6.1 Amino Acid Variation

The only significant amino acid variation was the result of an increase in one kind of diazotroph followed by pulses of glutamate/glutamine, and then aspartate/asparagine, alanine, and lysine. The regulation of protein breakdown and amino acid recovery in bacteria is not well understood, and it is not clear why glutamine assimilation did not lead to increased ratios of other amino acids as well. It is possible that some amino acids are synthesized in such small quantities to begin with that their variation and their regulation is not sensitive to precursor concentration, or that the variation is within the signal noise of the method. For example, this analysis shows that cysteine has a high component of variation, but since it is also the rarest amino acid small changes over time result in magnitudes of difference. Rather, increases in amino acids may vary together and optimal growth conditions increase total biomass rather than alter amino acid profile.

5.6.2 Metabolic Activities Related to SCP Production

Three metabolic processes in three environments tend to dominate proteins in SCP. A large footprint from yeast alcohol fermentation existed in wastes received from the brewery. This brewery waste then underwent acid production in the acidogenic holding basin before being transferred to an aerobic basin. Here, as microbes processed waste beer into cell mass during aerobic treatment, proteins from yeast were replaced by proteins derived from endogenous growth of aerobic diazotrophs on organic acids. A significant amount of acetic acid is formed in the bioreactor, indicating fermentation as a continuing activity as well. Lastly, the level of conductivity in these wastewaters indicates a high salt content and will likely favor salt-tolerant organisms and the production of proteins related to salt tolerance. These factors of organic acids, microaerophilic conditions, salinity and ammonia as nitrogen donor likely are the reasons why diazotrophic organisms were specifically enriched in this systems.
5.6.3 Proteome Variation

The results of this study reveal that over the study period little consistent gel pattern existed for the same bioreactor. This implies that proteomes are highly sensitive to local and temporal factors normally seen in a wastewater treatment plant. Given the industrial real-world nature of the system studied, more work is needed to understand protein variation over time in wastewater systems and what kind of environmental parameters influence such variation. Yet, despite having a proteome turnover, amino acid composition did not vary greatly except at a single time point. This can be interpreted as the proteome extracted was only a minor fraction of total proteins, or that the proteins expressed had a similar amino acid composition. If the latter case is likely, then this suggests that altering protein expression is not a significant driver of amino acid profiles.

5.6.4 Three Strategies for Protein Enhancement

We envisioned that proteins should be differentially enriched by cells by three potential strategies:

1. Fixation activities such as nitrogen and carbon dioxide fixation lead to accumulation of a large number of energy transport and fixation proteins such as dinitrogenase, or RuBisCO.

2. Toxic shock resistance proteins, such as due to antibiotic resistance, oxygen tolerance, or osmolaric regulation result in a number of protectant proteins being up-regulated several orders of magnitude by cells.

3. Growth and anabolic activities in a cell should increase proteins related to assimilation and cell division as well as synthesis proteins such as ribosomes.

In this study, the third case was observed and resulted in a successful protein production strategy. Using this same strategy at other sites, certain strains of microbes such as *Azospirillum* sp. living a non-nitrogen-fixing lifestyle, can be used to produce a high amount of biomass and proteins quickly, but this kind of system does not result in an enrichment in the more valuable limiting essential amino acids.

However, a different kind of SCP could be produced with the first case and using the same root-zone diazotrophic organisms. It is known that nitrogen fixation proteins have
higher sulfur-containing amino acids and results in a higher amount of cysteine and methionine, but tend to consume a large amount of energy in fixing nitrogen gas to ammonia [245]. The same type of pilot system can be examined without nitrogen addition to encourage such growth, sacrificing biomass for increased nitrogenase content in cells and saving nitrogen feedstock costs. As nitrogen fixing wastewater systems already exist in prototype [196], examining nitrogenase content in such systems is one potential avenue of research. Additionally, it has been observed that glutamate also has a role as an osmolyte for *Azospirillum* in saline conditions often seen in the rhizosphere [246]. This idea can be used to forcibly induce *Azospirillum* to accumulate certain protein osmoprotectants by varying salinity in culture. Indeed, in real world systems, a combination of these three factors may be responsible for wastewater protein expression as wastewater effluent conditions vary with industrial production cycles.

5.7 Conclusion

An experimental pilot bioreactor was monitored over time to understand factors influencing amino acid and protein production for an alternative fish meal protein product. As expected, variation in crude protein performance is observed, but well within 50-60% crude protein, and often higher. A steady variation in microbial consortia relating to diazotrophic bacteria resulted in a changing proteome expression profile. For the most part, neither 2D-PAGE gel patterns nor proteins identified stayed constant, except for several proteins related to growth and organic acid assimilation. Despite the noticeably poor database matching of many spots, useful metabolic information could be identified from highly conserved regions of proteins and a consensus analysis developed, but only for protein types and not taxonomic assignment. This work suggests that the major form of protein variation lies not in the amino acid profile expressed by conventional aerobic treatment and that amino acid enhancement strategies might not revolve around organism types but rather metabolism types that encourage a different protein profile expression.
CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Summary of Conclusions for Research Tasks
The major findings from experiments conducted in this dissertation followed by future work are listed in this chapter. The conclusions from research tasks addressed in Chapter 1 are summarized below:

6.1.1 Task 1: Microbial Ecology Survey and Identification of Significant Organisms in Wastewater Treatment
1. Define patterns of alpha diversity measurements that indicate changes in the functioning of the plant
   - Alpha diversity measurements during startup showed increases in alpha diversity but little relationship with functioning is observable.
2. Determine the pattern of beta diversity and co-abundance that indicate succession and disruption of bioreactors.
   - Beta diversity was used to relate the functioning of a pilot bioreactor to it's influent and to change over time. Large swings in the ordination position of a pilot bioreactor corresponded with disruption, small changes in ordination position related to temporal distance-decay steady-state behavior.
3. Determine the most important factors of variation of microbial ecology across a treatment plant.
   - Beta diversity multivariate statistics and clustering plots showed that the type of unit process had the largest influence on the microbial community. Additionally, very little shared relationship between stages was observable, except for the pilot reactor and SCP product.
4. Identify the most commonly enriched organisms in SCP production and their distribution.
   - Co-abundance analysis revealed that large amounts of saccharolytic fermenters passed from the influent to the mixed liquor, and that a number
of different diazotrophic Proteobacteria enriched in the pilot reactor, as well as some saccharolytic fermenters.

- The distribution of organisms in the pilot bioreactor were arranged in a common and cosmopolitan or rare and endemic profile, and organisms commonly exhibited a sharp boom-and-bust cycle over time.

5. Examine common metabolism of enriched organisms.

- Three different groups of diazotrophic bacteria in Alpha and Betaproteobacteria were detected. This suggests that the lifestyle of diazotrophic bacteria had a major role in SCP production in this pilot plant.

6.1.2 Task 2: Protein in Single Cell Protein

1. Examine SCP by traditional colorimetric, fluorometric, HPLC, and feed industry techniques

- Colorimetric and fluorometric techniques have difficulty with interferences without significant sample separation.
- Separation by HPLC can be used to quantify amino acids without needing much pre-treatment and was in agreement with crude protein values. Multivariate analysis of amino acid profile variation showed that a spike in glutamate/glutamine production occurred at one sample time, which corresponded with the presence of glutamine synthetase detected in mixed liquor.
- EEMs showed some contamination from humic-like or melanoidin-like substances, resulting in yellow-brown extracts.


- 2D-PAGE gels of several sample types showed that mixed liquor protein expression varied considerably from the influent protein profile.
- 2D-PAGE gels show that over time mixed liquor protein expression varied as well, but SCP product resembled mixed liquor for the same date.

3. Identify proteins related to SCP production and organisms from which SCP proteins derive.
• Peptide identification was possible despite the low spectra number of the peptide matches. Detected peptides tended to be from highly conserved and homologous and specific protein domains.

• Consistently detected peptides related to growth and organic acid uptake. Detected influent peptides tended to relate to yeast and alcohol fermentation, which is consistent with the organic acid results.

4. Examine common metabolisms from Task 1 to see if there is agreement.

• There was inconsistent agreement between phylogeny determined from proteomes and pyrotag. At least one entire group of dominant organisms did not have any peptides detected from it.

• Peptides detected in this study show that growth, rather than fixation or stress tolerance, was the main contribution to SCP protein composition.

• The use of diazotrophs for protein production in SCP research is suggested.

6.2 Conclusions and Recommendations

In this dissertation molecular biology tools such as DNA and protein sequencing technologies have been applied to the engineering of microbial systems. In particular, high-throughput sequencing and beta diversity analysis were used to track microbial diversity and relate it to bioreactor functioning. Additionally, metaproteomic and protein techniques were used to fractionate and analyze industrially produced SCP to make determinations about the dominant metabolic activities. The results of this work suggest that if the third strategy of growth is used, the focus of research should be on increasing production rates rather than altering amino acid profile. Also, these conclusions suggest that acidogenic pretreatment of wastes can be used to encourage diazotroph growth and is a potential area of future research.

6.3 Future Work

The future work discussed relates to how to make an improved SCP product and what kind of future research is enabled by pyrotag sequencing of bioreactors and microbial engineering systems in general.
6.3.1 Making a Better SCP Product

A major goal of SCP research is to produce a product with higher amino acid content and to characterize the variation in amino acid content. In this work it has been shown that a consortia of diazotrophic bacteria are feeding on organic acids and producing proteins, however, more research needs to be done to see if this result is repeatable, especially for other types of food-processing wastes. Plants will exude organic acids and simple sugars from roots, attracting diazotrophs by chemotaxis [247]. In return root-zone associated diazotrophs must produce fixed ammonia and retain enough for growth while excreting some for host plants [248]. This ability to manipulate ammonia may factor into their success in a bioreactor with high urea addition (which is rapidly converted into ammonia). For example, it is known that Azospirillum amazonense is sensitive to ammonia uptake based on the type of nitrogen used for pure culture and will actually increase glutamine synthetase expression with increased glutamate [249]. This led to monosodium glutamate wastewaters being tested to specifically culture for Azospirillum rugosum for root zone inoculant production [250]. This type of wastewater may serve as an excellent feedstock for a diazotroph based protein product as well.

The ultimate result of this work may be that acidogenic pre-treatment and other steps to liberate readily accessible carbon, combined with known response to certain nitrogen donors, can be used for producing a known microbial composition of SCP. Another possible application would be to tie microbial consortia or metabolic activity to a certain amino acid profile. This work has shown that proteins generated in SCP are a constantly changing mix. However, using the first or second strategy of nitrogen fixation or toxic shock, strong and repeated expression of specific proteins can be encouraged. Some of these proteins, such as nitrogenase, are indeed high in cysteine and methionine, but production rate will likely be sacrificed using such strategies.

6.3.2 Remaining Questions on the Structure of Microbial Communities

Microbial community analysis has matured leaps and bounds in only several years, but if anything this had only led to more issues. Technical problems remain such as the
analysis of chimeric sequences and reduction of the false positive rate. More fundamental problems are posed by patterns in microbial community results. Before theoretical principles can be defined, the fundamental unit of functional taxonomy must be understood. What is the appropriate level of grouping of organisms that represent a cohesive unit of behavior and interaction in an ecosystem? What is the set of advantageous attributes that represent clusters of organisms and what do these clusters mean physiologically or microbiologically? Is it the presence or absence of a gene, a gene system, a set of gene systems, or optimization of one of these that makes groups of organisms more useful over others, and do they relate on the level of an ecotype, a genus, or a family? As this problem is likely field specific, this could lead to the breakup of microbial ecology into subfields studying different aspects of community assembly and function, each obsessed with a different level of taxonomy to represent their problem. Moreover, the use of replicates and understanding how much variation occurs within a bioreactor or biofilm needs to be assessed.

6.3.3 Intentional Microbial Perturbation

As the power of genetic information comes not from its code but its variation, the study of variation in ‘omics’ will become more widespread now that tools for assessing that variation exist. Based on current thinking on the intermediate disturbance hypothesis and the unstable equilibrium of ecosystems, it should be possible to shift communities from one quasi-stable state to another and to hold it there. If abnormal communities can be identified, they can be treated by chemical stimulation or inoculation in a way designed to move their state from one equilibrium to a more desirable one. In wastewater engineering, understanding the normal and abnormal states of variation in constantly changing communities will serve a first step in finally characterizing engineering microbial systems. However, the ability of high-throughput sequencing to provide feedback to intentional perturbations of microbial communities will be its greatest advantage. Feedback allows engineers to modify and improve their manipulations of microbial ecology incrementally despite the lack of applicable theoretical models. This can one day enable the dream of bioaugmentation and targeted inoculation. Previous work in bioaugmentation showed that these manipulations often had little long-term
effect as inoculum organisms rapidly died out [251]. From what is known now about microbial ecology, consortia designed for sustained inoculation can one day be continuously cultured or combined in real time to suit a variety of constantly changing treatment conditions. In environmental engineering, this could be a boon to the study of enhanced biological phosphorus removal, methanogenesis digestion, and sub-surface organic DNAPL bioremediation. Similarly, probiotics, the use of beneficial microbes to influence health by influencing microbial community, is a close analog to this kind of problem. For the first time, the microbial shift due to probiotics can be definitively tracked and the exact mode and longevity of probiotics finally quantified using sequencing techniques combined with bioinformatics and multivariate statistical comparisons.

An example application of tracking microbial ecosystems for wastewater engineering purposes is the monitoring of UASB sludge granule formation. Structurally, sludge granules bear great similarity to microbial mats and have shown a cross-sectional heterogeneity tied to the syntrophic process of methane formation [200-201]. Granulation confers several benefits over conventional treatment that are related to the protection of sensitive methanogens from external system upsets by growing them within the core of sludge granules and provides a surface for acetogenic bacteria to colonize proximate to these methanogens [252]. The UASB process itself allows for higher loading rates for smaller reactor sizes with cost savings [252]. However, not all anaerobic wastes can be formed into granules and the process by which granules form is still debated, with several different proposals relating to nucleation, thermodynamics, and community consortia [252], [253]. By tracking the startup of a UASB, the formation of initial flocs and development and growth of granules can be monitored with a comprehensive pyrotag survey. Pyrotag results can be related to known performance parameters such as substrate concentration, pH, temperature, and granule size distributions. As the survey of UASB microbial diversity is far from complete, especially during the granule formation stages, this study will be significant for it’s level of coverage of rare or undiscovered UASB associated-organisms, and for identifying microorganisms and conditions that specifically initiate and promote granulation.
Lastly, the ability to compare organisms now extends well beyond the 16S gene alone. Using the proteo-genomic approach, varying gradients (time, distance or otherwise) of metagenomes, metatranscriptomes, and metaproteomes can all be compared with environmental and real-time PCR data to make useful conclusions about metabolic and ecosystem functioning. Though at present, the costs of such sequencing and analysis make this type of project impractical for the sample sizes needed for multivariate statistics.
REFERENCES CITED


[198] M. Papadelli, A. Roussis, K. Papadopoulou, A. Venieraki, I. Chatzipavlidis, P. Katinakis, and K. Ballis, “Biochemical and molecular characterization of an Azotobacter vinelandii strain with respect to its ability to grow and fix nitrogen in


APPENDIX

Personal Motivation

Must it not then be acknowledged by an attentive examiner of the histories of mankind, that in every age and in every State in which man has existed, or does now exist [1]that the increase of population is necessarily limited by the means of subsistence, [1]that population does invariably increase when the means of subsistence increase, and, [1]that the superior power of population is repressed, and the actual population kept equal to the means of subsistence, by misery and vice.

-Malthus, An essay on the principle of population (1798)

In rural development for subsistence farmers, there is always the uneasy unspoken question: Why increase available food, just to see it all disappear again? Why develop technologies to help the environment when we will only come up with more ways to spoil it and more people to fill it? Countries once self-sufficient in staple crops from the green revolution now find themselves importing more basic staples for a number of different reasons, most significant of which is the arrival of fuel crops, which have fundamentally changed agrarian economics. The competition of land for food and fuel has been implicated in a number of recent price shocks in cereals and rice.

We are faced with an unusual choice as a species. When explaining basic population ecology to students and friends, many find it amusing that humans have the same equations for growth as a bacterium on a petri dish, or an insect in a forest, or any other sort of animal. The logistic growth model has been used by students and scientists alike for decades and locks all species into an inexorable pattern of growth based on only a few fundamental principles. Our entire fate as a species is written in stone based on the rate of our breeding (r), and how many organisms the earth can hold (K). Like any ordinary bacterium, these two numbers strictly determine on paper the extent of our species and our civilization.
Or is it inexorable?

With human cognition comes free will. Knowing we’re trapped in a bottle we are presented with a choice. Are we r-strategists, rapidly growing to consume any environment, before dying off and repeating this miserable feast-or-famine cycle until the growth dynamics are so unstable we drive ourselves extinct (as predicted by discretized population modeling)? Or are we K-strategists, specialists who slowly fill up a space to equilibrium? Living in the ivory tower we are free to debate slopes on a line, but out there on all the dusty unpaved roads of the world live rural peoples who are impacted by this reality every day. “How many children can I have?” “How many will I have to bury in my lifetime?” “Will harvests be good or bad this year?” “If things get really bad, which children should we choose to live, or choose to send away or stay?” If the purpose of development is to provide rural farming families the food security and basic education needed to free communities from the hand-to-mouth struggle of subsistence agriculture, an increasing array of sustainable hardy technologies are needed to provide food sources for a future fraught with the consequences of climate change and a still increasing population. With this ability to innovatively alter our behavior as a species comes the hope that one day we become what we always assume we are, an organism exceptional from others with the ability to choose more than just the method of our own demise. We can choose to see all our children grow up healthy, educated, and employed. We can choose to feed our families without spoiling coral reefs and rainforests. We can own the land we farm rather than lose it in a bad year. I came back to environmental engineering because after the despair of being a powerless Peace Corps volunteer in the face of widespread rural poverty I wanted to ask what is actually possible in a practical manner. As environmental engineers, we do not fight Malthus so much as we try to enable people with reasonable options. In that sense, we try to build a future free of Malthusian predetermination.

This work begins with my story of the little Tilapia fish. I have a most unusual association with this creature for the past 15 years; and as time goes by I believe much of
the world will too. As a young intern at NASA studying how humans will stay alive on Mars living in completely closed off habitats, the question of how colonists will obtain their protein was raised. Early research has suggested that the tilapia fish (in particular the Nile Tilapia *Oreochromis niloticus*, (Figure A1)), a hardy scavenging feeder with a mild flavor and originating from Africa, could be raised in low-gravity or zero-gravity environments feeding on only greenhouse scraps. Fast forward several years, and I found myself eating that same fish, once or twice per day, as my main protein source (along with the many other Filipinos I lived with) in the Peace Corps. Suddenly the problem was more than academic, and with few other protein options I began to look forward to eating a saucer-sized tan and black striped fish, smacking it down tail and all. Our ability to provide an effective cheap source of protein due to the exponential cultivation of this fish has been so successful that it has displaced much of the other fish traditionally consumed in The Philippines, particularly as traditional fish become rapidly fished to extinction. Now, families will eat traditionally caught fish on special occasions or when they can afford it, and fishermen must go to greater and greater lengths to obtain them. A major BBC series highlighted Filipino fisherman risking death by the bends to obtain fish by compressor diving well beyond traditional human tolerance. It is clear that without a stable source of protein we will again be under Malthus’s thumb.
Figure A2. Tilapia cultivation, in thousands of tonnes per year harvested (FAO)

During this time in The Philippines, I was trained to scuba dive by a man who was a passionate reef conservationist. He offered to teach me and any other volunteer at cost under one condition: that we would advocate fish conservation in our work. I was doing water engineering at the time, so I had no idea how that would happen, but his words proved to be prophetic. Moving forward another five years, I look back now on that promise and wonder if he knew then that I would become an avid diver and engaged in a project of sustainable fish feed development. Though there are numerous challenges with sustainably expanding Tilapia aquaculture, it is already happening at an explosive rate (Figure A2). However, as the Tilapia is a hardy grower with cheap and even fish meal free diets, I still recommend it as the path to a more sustainable seafood future.

And I still love eating Tilapia fish, the Filipino way, fried in palm oil with a little salt.
## Supplemental Figures and Tables

### Table A1. Chemical Protein Methods Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Principle of Operation</th>
<th>Comparison to Crude Protein</th>
<th>Potential Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kjeldahl-N Crude Protein [49], [262]</td>
<td>Combustion of sample with subsequent detection of nitrogen. Nitrogen is multiplied by empirical factor 6.25 based on meat protein studies</td>
<td>Same</td>
<td>6.25 factor is inappropriate. Some studies suggest 4.3 is better for plant and bacterial [226]. This fixes harvest protein %</td>
</tr>
<tr>
<td>Natural Absorbance [14]</td>
<td>Spectroscopy at 280 nm of natural absorbance based primarily on tryptophan indole absorbance</td>
<td>No quantifiable data</td>
<td>Interference by unknown substance (possibly phenol group). Signal is lost in the shoulder of this other analyte</td>
</tr>
<tr>
<td>Bradford/ Ninhydrin Reagent [263]</td>
<td>Absorbance shift due to ionization of protein functional groups at 595 nm / 570 nm, primarily arginine, also basic (His, Lys), and aromatic (Try, Tyr, Phe).</td>
<td>Not used</td>
<td>Known incompatibility with SDS and other surfactants used to lyse cells</td>
</tr>
<tr>
<td>BCA Reagent [264]</td>
<td>Peptide bonds bind with Cu2+ triggering a BCA Cu1+ absorbance product at 562 nm</td>
<td>Measurements are 2x crude protein values based on standards with BSA</td>
<td>Interference by unknown substance (possibly humics). Signal is lost in the shoulder of another analyte</td>
</tr>
<tr>
<td>BCA Reagent + Precipitation Separation [264]</td>
<td>Precipitation reaction is a acid salt 2 step precipitation method by Pierce for the removal of contaminated substances</td>
<td>Data was consistently 15% magnitude less than crude protein until 9/10/08 based on standards with BSA</td>
<td>During precipitation, a brown possibly humic substance was observed on top of extracted white protein layer</td>
</tr>
<tr>
<td>Fluorescamine + Precipitation Separation [265]</td>
<td>Fluorophore is bound to amine group and then excitation at 390 nm and emission at 475 nm</td>
<td>Single point maxima was used as the measurement factor. Result is a factor of 10 lower than crude protein content based on standards with BSA</td>
<td>Possibly mixed humics that precipitated with protein are interfering or quenching fluorescence. Fluorescamine used with pure extracted humic acid reagent produced a 1:1 signal strength at the detection EE.</td>
</tr>
<tr>
<td>Natural Fluorescence [235]</td>
<td>Natural fluorescence of NOM detected in a raster scan from excitation 240 - 450 nm, emission from 290 - 580 nm of bulk fluid. Tryptophan containing compounds fluoresce strongly at 280 nm excitation, 330 nm emission</td>
<td>Signal peaks were integrated over 2d range 240-300 nm excitation, 290 - 400 nm emission. Result is a factor of 10 lower than crude protein content based on standards with BSA. Whole EEMs showed that humics remained in the precipitated product.</td>
<td>Possibly mixed humics that precipitated with protein are interfering, scattering, or quenching fluorescence. Humic acid reagent had zero protein signal.</td>
</tr>
<tr>
<td>HPLC Amino Acid Analysis [49]</td>
<td>Digestion of proteins with ninhydrin IC or LC separation and quantification by absorbance at 570 nm</td>
<td>Single data point matches crude protein data point, but other data from Oberon not conclusive</td>
<td>Expensive (600 USD / analysis). Possible measurement errors.</td>
</tr>
<tr>
<td>Spot No.</td>
<td>Search DB</td>
<td>Protein Family</td>
<td>MW</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>----------------</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>NCBInr-elongation factor Tu</td>
<td>44</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>glutamine synthetase</td>
<td>52</td>
<td>5.2</td>
</tr>
<tr>
<td></td>
<td>elongation factor Tu prolyl-CoA carboxylase</td>
<td>43</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>glutamine synthetase</td>
<td>57</td>
<td>6.5</td>
</tr>
<tr>
<td>2</td>
<td>NCBInr-chemotaxis</td>
<td>54</td>
<td>5.1</td>
</tr>
<tr>
<td>3A</td>
<td>μ histidine kinase</td>
<td>120</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spot No.</td>
<td>Search DB</td>
<td>Protein Family</td>
<td>MW</td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>----------------</td>
<td>-----</td>
</tr>
<tr>
<td>3B</td>
<td>N.D.</td>
<td>N.D.</td>
<td></td>
</tr>
<tr>
<td>4A</td>
<td>N.D.</td>
<td>N.D.</td>
<td></td>
</tr>
<tr>
<td>4B</td>
<td>N.D.</td>
<td>Swiss-</td>
<td>4680</td>
</tr>
<tr>
<td>5 Prot</td>
<td></td>
<td>Aldehyde-NCBInr-</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Elongation factor Tu</td>
<td>Desulfovibrio vulgaris subsp. vulgaris (strain DP4)</td>
<td>ABM27470.1</td>
</tr>
<tr>
<td>4B</td>
<td></td>
<td>Aldehyde-NCBInr-</td>
<td>55</td>
</tr>
<tr>
<td>5 Prot</td>
<td></td>
<td>Aldehyde-NCBInr-</td>
<td>56</td>
</tr>
</tbody>
</table>

Table A2. Full Peptide Identification Tables Continued
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Search No.</th>
<th>Protein Family</th>
<th>MW (kDa)</th>
<th>pI</th>
<th>Organism</th>
<th>NCBI No.</th>
<th>MS/MS Search Score</th>
<th>Peptide #</th>
<th>Mean Spectral Intensity</th>
<th>Sequence</th>
<th>MS/MS Search Score</th>
<th>% BLAST COG Category</th>
<th>Peptide #</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47</td>
<td>Swiss-Prot Enolase</td>
<td>6.2</td>
<td>5288c</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508)</td>
<td>CAA97283.1</td>
<td>72.13</td>
<td>14</td>
<td>3.46</td>
<td>3.08E+04</td>
<td>20.08</td>
<td>84</td>
<td>COG0148 G</td>
<td>B1.1</td>
</tr>
<tr>
<td>1</td>
<td>51</td>
<td>ATP Synthase Beta Subunit</td>
<td>5.1</td>
<td>BAA-1182</td>
<td>Bradyrhizobium sp. (strain BTAI1 / ATCC 63290 / NCTC 12742)</td>
<td>ABQ32694.1</td>
<td>26.34</td>
<td>130</td>
<td>32.10</td>
<td>2.54E+04</td>
<td>12.91</td>
<td>79</td>
<td>COG0148 G</td>
<td>B1.2</td>
</tr>
<tr>
<td>1</td>
<td>43</td>
<td>Elongation Factor TU</td>
<td>5.3</td>
<td>10512</td>
<td>Brucella ovis (strain ATCC 25840 / ATCC 63290 / NCTC 12742)</td>
<td>ABQ60947.1</td>
<td>20.08</td>
<td>185</td>
<td>45.68</td>
<td>3.02E+04</td>
<td>17.22</td>
<td>88</td>
<td>COG0055 C</td>
<td>B1.3</td>
</tr>
<tr>
<td>1</td>
<td>43</td>
<td>Succinyl-CoA Ligase</td>
<td>5.4</td>
<td>ATCC 51521 / SW</td>
<td>Rhodopsirillum centenum (strain ATCC 25840 / ATCC 63290 / NCTC 12742)</td>
<td>ACI99574.1</td>
<td>53.94</td>
<td>5</td>
<td>1.23</td>
<td>1.67E+05</td>
<td>11.75</td>
<td>70</td>
<td>COG0045 C</td>
<td>B1.7</td>
</tr>
<tr>
<td>1</td>
<td>41</td>
<td>Alcohol dehydrogenase</td>
<td>5.8</td>
<td>S288c</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508)</td>
<td>CAA29410.1</td>
<td>47.38</td>
<td>4</td>
<td>0.99</td>
<td>2.18E+04</td>
<td>19.32</td>
<td>91</td>
<td>COG1454 C</td>
<td>B1.8</td>
</tr>
<tr>
<td>1</td>
<td>43</td>
<td>Serpin</td>
<td>5.7</td>
<td>Hordeum vulgare</td>
<td>Rhodopsseudomonas palustris (strain BisA53)</td>
<td>CAA36015.1</td>
<td>47.38</td>
<td>3</td>
<td>0.74</td>
<td>2.09E+04</td>
<td>8.28</td>
<td>64</td>
<td>KOG2392 V</td>
<td>B1.9</td>
</tr>
<tr>
<td>1</td>
<td>55</td>
<td>ATP Synthase Subunit Alpha</td>
<td>5.9</td>
<td></td>
<td>Hordeum vulgare</td>
<td>ABJ04241.1</td>
<td>47.38</td>
<td>51</td>
<td>12.59</td>
<td>3.99E+04</td>
<td>13.68</td>
<td>77</td>
<td>COG0056 C</td>
<td>B1.1</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>Serpin</td>
<td>5.5</td>
<td>Hordeum vulgare</td>
<td>CAA64599.1</td>
<td>47.38</td>
<td>2</td>
<td>0.49</td>
<td>1.36E+04</td>
<td>10.95</td>
<td>67</td>
<td>KOG2392 V</td>
<td>B1.1</td>
<td></td>
</tr>
</tbody>
</table>

**Table A2. Full Peptide Identification Tables Continued**
Table A2. Full Peptide Identification Tables Continued

<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Search DB</th>
<th>Protein Family</th>
<th>MW</th>
<th>pI</th>
<th>Organism</th>
<th>NCBI No.</th>
<th>MS/MS Search Score</th>
<th>No. Peptides</th>
<th>Mean Spectral Intensity</th>
<th>Sequence</th>
<th>MS/MS Search Score</th>
<th>% BLAST COG Category Peptide # APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(G)KGDIPNGLA DYY(R)</td>
<td></td>
<td>19.7</td>
</tr>
<tr>
<td>4</td>
<td>N.D.</td>
<td>ATP synthase alpha subunit</td>
<td>6.9</td>
<td>5.7</td>
<td>ORS 278</td>
<td>YP_001202610</td>
<td>86.63</td>
<td>31</td>
<td>41.89 1.44E+05</td>
<td>(Q)RGIDLYSL(K)</td>
<td></td>
<td>19.85</td>
</tr>
<tr>
<td>5</td>
<td>NCBlnr-α</td>
<td>aldehyde dehydrogenase</td>
<td>56</td>
<td>5.3</td>
<td>4680</td>
<td>NP_828353.1</td>
<td>19.85</td>
<td>43</td>
<td>58.11 5.79E+05</td>
<td>(N)KTTQGHDL MM(K)</td>
<td></td>
<td>13.84</td>
</tr>
<tr>
<td>6</td>
<td>NCBlnr-α</td>
<td>aldehyde dehydrogenase</td>
<td>56</td>
<td>5.3</td>
<td>4680</td>
<td>NP_828353.1</td>
<td>51.43</td>
<td>53</td>
<td>100.00 3.28E+05</td>
<td>(Q)RGIDLYSL(K)</td>
<td></td>
<td>20.39</td>
</tr>
<tr>
<td>7</td>
<td>Swiss-Prot</td>
<td>Elongation Factor TU</td>
<td>43</td>
<td>5.2</td>
<td>DP4</td>
<td>ABM27470.1</td>
<td>28.34</td>
<td>93</td>
<td>100.00 2.01E+04</td>
<td>(D)KGGGVEEG V(K)</td>
<td></td>
<td>17.78</td>
</tr>
<tr>
<td>8</td>
<td>NCBlnr-α</td>
<td>aldehyde dehydrogenase</td>
<td>56</td>
<td>5.3</td>
<td>4680</td>
<td>NP_828353.1</td>
<td>27.85</td>
<td>11</td>
<td>100.00 2.75E+04</td>
<td>(Q)RGIDLYSL(K)</td>
<td></td>
<td>16.33</td>
</tr>
<tr>
<td>9</td>
<td>NCBlnr-α</td>
<td>aldehyde dehydrogenase</td>
<td>39</td>
<td>9.1</td>
<td>viciea 3841</td>
<td>YP_768790.1</td>
<td>12.7</td>
<td>1</td>
<td>9.09 2.64E+05</td>
<td>(D)RPNTVSL(A)(K)</td>
<td></td>
<td>12.7</td>
</tr>
<tr>
<td>10</td>
<td>Saccharomyces</td>
<td></td>
<td>36</td>
<td>6.7</td>
<td>pastorius</td>
<td>BAHO03720.1</td>
<td>11.99</td>
<td>6</td>
<td>54.55 6.18E+03</td>
<td>(A)KAATEGSSPS1 NGSAT(R)</td>
<td></td>
<td>11.99</td>
</tr>
</tbody>
</table>
Table A2. Full Peptide Identification Tables Continued

<table>
<thead>
<tr>
<th>Spot</th>
<th>Search</th>
<th>Protein Family</th>
<th>MW</th>
<th>pl</th>
<th>Organism</th>
<th>NCBI No.</th>
<th>MS/MS Search Score</th>
<th>No. Peptides</th>
<th>% of Spot</th>
<th>Mean Spectral Intensity</th>
<th>Sequence</th>
<th>MS/MS Search Score</th>
<th>BLAST E Value</th>
<th>COG Category</th>
<th>Peptide #</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Swiss-</td>
<td>Enolase</td>
<td>47</td>
<td>6.2</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>CAA97283.1</td>
<td>167.39</td>
<td>46</td>
<td>88.46</td>
<td>1.09E+05</td>
<td>(N)KFOLDYKGD</td>
<td>11.08</td>
<td>74</td>
<td>KOG2670 G</td>
<td>IN1.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Prot</td>
<td>Phosphoglycerate kinase</td>
<td>45</td>
<td>7.1</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>CAA42329.2</td>
<td>48.53</td>
<td>6</td>
<td>11.54</td>
<td>2.05E+04</td>
<td>(E)KAVPALS(K)</td>
<td>12.38</td>
<td>74</td>
<td>KOG1367 G</td>
<td>IN1.2</td>
<td>70.9</td>
</tr>
<tr>
<td>2</td>
<td>Swiss-</td>
<td>Acyl-CoA dehydrogenase</td>
<td>41</td>
<td>5.1</td>
<td>Megasphaera elsdenii</td>
<td>AAA03594.1</td>
<td>93.96</td>
<td>6</td>
<td>66.67</td>
<td>6.55E+04</td>
<td>(G)KGEELMN</td>
<td>7.99</td>
<td>62</td>
<td></td>
<td>IN1.3</td>
<td>74.4</td>
</tr>
<tr>
<td></td>
<td>Prot</td>
<td>Serpin</td>
<td>43</td>
<td>5.7</td>
<td>Hordeum vulgare Azobacteroides pseudotrichonymphae genomovar</td>
<td>CAA36015.1</td>
<td>40.97</td>
<td>3</td>
<td>33.33</td>
<td>3.00E+04</td>
<td>(E)KADGAGGD</td>
<td>20.82</td>
<td>83</td>
<td>COG1960 I</td>
<td>IN2.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Swiss-</td>
<td>Elongation factor Tu N.D.</td>
<td>44</td>
<td>5.5</td>
<td>CFP2</td>
<td>BAG83276.1</td>
<td>54.36</td>
<td>16</td>
<td>100.00</td>
<td>5.15E+04</td>
<td>(K)RFMEVGT</td>
<td>15.78</td>
<td>89</td>
<td>COG0050 J</td>
<td>IN3.1</td>
<td>81.6</td>
</tr>
<tr>
<td>6</td>
<td>Swiss-</td>
<td>Pyruvate decarboxyylase</td>
<td>62</td>
<td>5.8</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>CAA28380.1</td>
<td>193.78</td>
<td>31</td>
<td>100.00</td>
<td>5.77E+04</td>
<td>(I)RAYGDAAYAANLENANGAW</td>
<td>23.07</td>
<td>93</td>
<td>KOG1184 EH</td>
<td>IN6.1</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Swiss-</td>
<td>Enolase</td>
<td>47</td>
<td>5.7</td>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>AAA88713.1</td>
<td>210.85</td>
<td>45</td>
<td>97.83</td>
<td>1.01E+05</td>
<td>(A)JRAAmsSVGLIANLGl(K)</td>
<td>13.3</td>
<td>75</td>
<td>KOG2670 G</td>
<td>IN7.1</td>
<td>-</td>
</tr>
</tbody>
</table>

162
<table>
<thead>
<tr>
<th>Spot No.</th>
<th>Search DB</th>
<th>Protein Family</th>
<th>MW</th>
<th>pl</th>
<th>Organism</th>
<th>NCBI No.</th>
<th>NCBI Search Score</th>
<th>MS/MS Search %</th>
<th>Mean Spectral Intensity</th>
<th>MS/MS Search %</th>
<th>BLAST COG Category</th>
<th>Peptide #</th>
<th>APC</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>N.D.</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Swiss-Prot Phosphoglycerate kinase</td>
<td>45 7.1 S28Rc</td>
<td>AAA88729.1</td>
<td>378.11</td>
<td>74 100.00 1.33E+05</td>
<td>(Y)KITPLAAV(R)</td>
<td>17.7</td>
<td>94 KOG1367 G</td>
<td>IN10.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Phosphoglycerate kinase acetyl-CoA acetyltransferase</td>
<td>41 5.9 E88</td>
<td>NP_781017.1</td>
<td>21.08</td>
<td>1 100.00 6.19E+04</td>
<td>(Y)KDLNDLQI(K)</td>
<td>15.1</td>
<td>80</td>
<td>IN10.2</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>N.D.</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Swiss-Prot Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>36 8.3 S28Rc</td>
<td>CAA24609.1</td>
<td>173.81</td>
<td>59 100.00 2.31E+05</td>
<td>(Y)KLAPC(V)TSANSVI(K)</td>
<td>24.89</td>
<td>97 KOG0657 G</td>
<td>IN13.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>36 8.3 S28Rc</td>
<td>CAA24609.1</td>
<td>282.08</td>
<td>75 100.00 4.57E+05</td>
<td>(Y)KLAPC(V)TSANSVI(K)</td>
<td>25.94</td>
<td>98 KOG0657 G</td>
<td>IN14.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>36 8.3 S28Rc</td>
<td>CAA24609.1</td>
<td>282.08</td>
<td>75 100.00 4.57E+05</td>
<td>(Y)KLAPC(V)TSANSVI(K)</td>
<td>25.94</td>
<td>98 KOG0657 G</td>
<td>IN14.1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>