DATA RECONCILIATION IN BIOPROCESS DEVELOPMENT

by

John Joseph Prior, Jr.

B.S., Chemical Engineering
University of Connecticut (1986)

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1997

© Massachusetts Institute of Technology 1997. All rights reserved.

Author .................................................................

Department of Chemical Engineering

January 21, 1997

Certified by ..............................................................

Charles L. Cooney
Professor of Chemical and Biochemical Engineering
Thesis Supervisor

Accepted by .............................................................

Robert E. Cohen
Chairman, Departmental Committee on Graduate Students
DATA RECONCILIATION IN BIOPROCESS DEVELOPMENT

by

John Joseph Prior, Jr.

Submitted to the Department of Chemical Engineering on January 21, 1997, in partial fulfillment of the requirements for the degree of Doctor of Science

Abstract

To interpret and reduce bioprocess variability, one needs to assess the accuracy and consistency of the collected data. Faulty data, which can result from human error, miscalibration, drift, and equipment failure, are difficult to distinguish from the inherent noise in time-varying, nonlinear bioprocesses. This research evaluates and improves the methods for using data reconciliation to detect inconsistent data and models and to identify potential faults in bioprocesses. It defines and organizes the necessary assumptions, models, and correlations into a data reconciliation (DR) framework to assist in process diagnosis. Characteristic problems observed in studies of industrial and lab-scale data were used to drive modeling and methodology work.

We developed an algorithm for observability and redundancy classification required for nonlinear reconciliation, and defined sub-classes reflecting the robustness of variables to measurement loss. The algorithm identifies subsets of indistinguishable faults thereby improving computation speed and diagnostic clarity. We found that ensuring the quality of off-gas related measurements, which contribute to a large portion of real-time variables, is especially important for diagnosis and control. We established additional redundancy using models for mass transfer and nutrient yields.

We derived models and heuristics summarizing the effects of accuracy, stoichiometry, component compositions, and process dynamics on the calculated respiratory quotient (RQ) and transfer quotient (TQ). Bicarbonate dynamics cause the RQ to differ from the TQ in processes operated near or above neutral pH levels. RQ accuracy is determined by off-gas measurement accuracy when the process pH is low and by pH measurement accuracy when the pH is neutral or alkaline. We modeled these phenomena in order to optimize the RQ accuracy and to detect incorrect measurements of pH, gas flow, and cell growth rate.

In experiments adding pulses of 0.1 g L⁻¹ of $^{13}$C-labeled glucose and acetate to growing cultures, we found mass spectrometry could measure the resulting fraction of $^{13}$CO₂ in the exit gas stream with a fairly low noise level ($\sigma_y=0.004$). In experiments
growing *A. vinelandii*, we optimized operating conditions in order to achieve a measurement noise reduction sufficient to measure the rate of gaseous nitrogen uptake (NUR). We observed a maximum rate of 2 mmol L\(^{-1}\) h\(^{-1}\) and a specific rate of 1.2 mmol g\(^{-1}\) h\(^{-1}\). When compared with traditional nitrogenase assays, the measurement has a relative standard deviation of 15-25 percent.

Thesis Supervisor: Charles L. Cooney
Title: Professor of Chemical and Biochemical Engineering
Acknowledgments

First, I would like to thank my thesis advisor, Charles Cooney, for providing me with a broad and rich educational experience and for his world-class patience, optimism, and consistency. My thanks also go to the members of my thesis committee: Gregory Stephanopoulos, George Stephanopoulos, and Daniel I.C. Wang. Together they provided a varied set of personal and professional strengths that I will always aspire to approach.

Amylin Pharmaceuticals (San Diego, CA), International Biosynthetics (Kingstree, SC), and Gist Brocades (Delft, The Netherlands) provided the bulk of funding and process data for this research. Additional support was provided by the National Science Foundation through the Engineering Research Center Initiative to the MIT Biotechnology Process Engineering Center (cooperative agreement ECD8A-03014), and by fellowships from Pfizer (Groton, CT) and the BIG EAST Athletic Conference. Thanks also go to the Alberta Research Council (Edmonton, Canada), Repligen (Dedham, MA), BASF (Worcester, MA), Perkin Elmer (Norwalk, CT), Alfa-Laval (Sweden), and Gensym (Cambridge, MA) for their collaboration and equipment donations. I’m grateful for the efforts of my colleagues in these organizations; particularly to Jim Hettenhaus, Theresa Olson, Jennifer Reichert, John Tedesco, Chris Lynch, Lee Dahm, Grame Macaloney, Mike Best, Jan Feijen, Jac Dekker, and Robbie Rogers. Jim’s vision set the stage for my work and his ongoing encouragement and mentoring helped me greatly.

Thanks also to my MIT collaborators and UROP students for their efforts in generating and analyzing a large portion of the data used in this work. Kwong Son and Roy Magnuson collaborated on several small-scale fermentations. Stefan Winkler spent several months with us collaborating on experiments while somehow managing to keep the bioreactor operational. Patti Christie and William Orme-Johnson collaborated on the NUR monitoring work along with Robert Murray whose exemplary, single-handed operation of the MIT pilot plant yielded large amounts high-quality data for study.
Other contributors included Cathryn Shaw and Kristala Jones (\textsuperscript{13}CER), Jeff Galvin (data analysis), Sean Campbell and Dave Hopkins (computer interfacing), and Trang Truong and Steve Wallace (media analysis). The readability of the thesis was greatly improved by the thoughtful review and comments of Dave Hopkins, Paul Bucciglil, Paula Maute, and Jack and Kathy Prior.

Looking back, I owe much of my early successes in life to the faith, patience, and encouragement of my parents as well as my many athletic coaches, notably James McKee, John Misili, and the late Robert Kennedy. I owe the \textit{Ph(ilo)sophy} part of my education to the friendship of my fellow basement-dwellers Sergio Sanchez, G. K. Raju, Robert Murray, David Hopkins, Noriya Yokota, Takahiro Suzuki, and Tadashi Kashihara. G. K. in particular was a great friend and mentor through my entire stay at the institute and he contributed greatly to my work, education, and perspective. I've also valued the friendship of my classmates Brian Kelly, Dan Lasko, David Chang, and Rahul Singvi and many lab-mates including Jim McMillian, Christine Moore, Jean-Francois Hamel, Eric Scharin, John Konz, Chris Dowd, Steffen Ernst, Aleks Engel, Greg O'Connor, and Per Lindel. Orna Gardini, Sonia Foster, Linda Mousseau, Elaine Auferio-Peters, and Janet Fischer were a great help over the years and made MIT a warmer place. Family members Paul, Joe, Nanny, Gram, Gampy, Pat, and Karen were an ongoing source of morale-boosting internet correspondence. The many members of \textit{La Casa} (Spanish house), where Lisa and I spent several years as graduate residents, played a special part of my family's experience at the institute.

My daughter Kayla, a 3 year member of the Toddler Research Opportunities Program, has added immeasurably to my life and education as she as patiently waited for me to finish my 'pieces'. Finally, words cannot convey my sense of gratitude and love for my wife Lisa, who ventured to Cambridge with me 9 years ago in our lime-green Toyota to begin this journey and our marriage together. Her support, love, and sacrifice made completion of this thesis possible and I dedicate it to her.
Contents

1 INTRODUCTION .................................................. 15
  1.1 Thesis Organization ........................................ 15
  1.2 Problem Description ......................................... 15
  1.3 Project Objectives ........................................... 16
  1.4 Background .................................................. 17
    1.4.1 The Role of Data Reconciliation ......................... 17
    1.4.2 Respiratory Quotient Interpretation ...................... 21
    1.4.3 Monitoring of Nitrogen Uptake .......................... 21
    1.4.4 Monitoring of $^{13}$CO$_2$ Evolution .................... 23
  1.5 Research Methodology ....................................... 24
  1.6 Thesis Contributions ....................................... 26

2 DATA RECONCILIATION METHODOLOGY ............................ 27
  2.1 System Representation ...................................... 27
    2.1.1 Network Balances ...................................... 30
    2.1.2 Reaction Representation ................................ 31
    2.1.3 Nutrient Yields ....................................... 34
  2.2 Parameter Estimation ...................................... 35
    2.2.1 Linear Models ........................................ 36
    2.2.2 Nonlinear Models ...................................... 37
  2.3 Measurement and Uncertainty Representation ................. 38
    2.3.1 Limitations of Linear Representation .................. 40
    2.3.2 Equations to Estimate Uncertainty ...................... 41
  2.4 Variable Classification ................................... 42
    2.4.1 Definitions .......................................... 43
    2.4.2 Classification Method ................................ 43
  2.5 Detecting and Identifying Gross Errors ..................... 45
  2.6 Performance Metrics ....................................... 50
  2.7 Degrees of Freedom in a Bioprocess ........................ 50

3 BIOPROCESS GAS MODELING ....................................... 55
  3.1 Overview of Gas Variables ................................ 55
  3.2 Dissolved Gases ........................................... 57
    3.2.1 Dissolved CO$_2$ ...................................... 58
    3.2.2 Dissolved CO$_2$ Dynamics ................................ 60
  3.3 Gas Exchange Rates ........................................ 61
  3.4 Overall Volumetric Mass Transfer Coefficient ................. 63
  3.5 The Respiratory and Transfer Quotients ....................... 65

7
3.6 Secondary Estimates ........................................ 67
3.6.1 Dissolved O$_2$ Based on OTR and $k_La$ ............. 69
3.6.2 OTR Based on $k_La$ and Dissolved O$_2$ ............. 69
3.6.3 OUR Based on Growth Rate ............................. 72
3.6.4 OUR Based on Heat Evolution .......................... 72
3.6.5 NUR Based on Growth Rate ............................. 72
3.6.6 RQ Based on Stoichiometry ............................. 74
3.6.7 Cell Yield Based on RQ ................................. 78
3.6.8 Heuristics for RQ Interpretation ....................... 79
3.7 Optimization of Off-Gas Calculation Accuracy ......... 81
3.7.1 Optimization of NUR Accuracy ......................... 81
3.7.2 Optimization of RQ Accuracy ......................... 82

4 MATERIALS AND METHODS .............................. 87
4.1 RQ Interpretation Case Study Example .................. 87
4.1.1 Experimental System ................................. 87
4.1.2 Experimental Operation ................................. 89
4.1.3 Gas Analysis .......................................... 89
4.2 Monitoring of $^{13}$CO$_2$ Evolution ...................... 89
4.2.1 Sampling and Sample Treatment ....................... 90
4.2.2 Cell Density ......................................... 91
4.2.3 Analysis of Nutrients and Waste Products ........... 91
4.3 NUR Monitoring Experiments ............................. 91
4.3.1 Acetylene Reduction Assays .......................... 95
4.3.2 Reactor Monitoring ...................................... 96

5 RESULTS AND DISCUSSION ............................. 99
5.1 Elemental Balance Data Reconciliation .................. 99
5.2 Measurement Consistency Analysis ....................... 105
5.2.1 Use of RQ in Consistency Analysis .................. 105
5.2.2 Use of $k_La$ in Consistency Analysis ............... 109
5.2.3 Use of Nutrient Analysis in Consistency Analysis .. 110
5.3 Measurement of NUR ...................................... 110
5.3.1 Baseline Measurement Noise .......................... 113
5.3.2 Gas Feed Optimization ................................ 113
5.3.3 Observed Cell Growth ................................ 116
5.3.4 NUR Measurement .................................... 117
5.3.5 Assay Correlations with NUR ......................... 117
5.3.6 Systematic Errors .................................... 120
5.4 Monitoring of $^{13}$CO$_2$ Evolution ..................... 120
5.4.1 $^{13}$CER Profile and Label Balance .................. 121
5.4.2 Monitoring Acetate Concentration and Consumption .... 121
5.4.3 $^{13}$CER Dynamics .................................... 125
5.5 Nonlinear Data Reconciliation ........................... 125
5.5.1 Experimental Background .............................. 127
5.5.2 Observability and Redundancy Classification ........................................ 127
5.5.3 Detection of Gross Errors ................................................................. 131
5.5.4 Significance ......................................................................................... 133

6 CONCLUSIONS AND RECOMMENDATIONS ................................................. 135
6.1 Conclusions ............................................................................................ 135
6.1.1 Nonlinear Data Reconciliation ............................................................ 135
6.1.2 Nonlinear Variable Classification ......................................................... 136
6.1.3 Respiratory Quotient .......................................................................... 136
6.1.4 Monitoring of $^{13}$CO$_2$ Evolution ...................................................... 137
6.1.5 NUR monitoring ................................................................................... 137
6.2 Recommendations for Future Research .................................................... 137
6.2.1 Data Analysis Automation .................................................................. 137
6.2.2 Nutrient Medium Design and Analysis ................................................. 138
6.2.3 Monitoring of $^{13}$CO$_2$ Evolution ...................................................... 138

References ..................................................................................................... 139

A Appendix ..................................................................................................... 149
A.1 Nomenclature .......................................................................................... 149
A.2 Media Yields ............................................................................................ 152

B SOURCE CODE .......................................................................................... 157
B.1 RQ, $k_LA$, and NUR Calculations ........................................................... 157
B.1.1 aibabias.m ............................................................................................ 157
B.1.2 aiba.m .................................................................................................. 158
B.1.3 rqvyield.m ............................................................................................ 158
B.1.4 rqcac.m ................................................................................................ 159
B.1.5 yxsgcalc.m ............................................................................................ 159
B.1.6 syxsrrq.m ............................................................................................. 159
B.1.7 rqvyp.m ................................................................................................. 159
B.1.8 yxserr.m ............................................................................................... 161
B.1.9 tqerr.m ................................................................................................ 161
B.1.10 rqerrbic.m ......................................................................................... 162
B.1.11 optaer.m ............................................................................................ 163
B.1.12 bs2sig.m .............................................................................................. 164
B.1.13 nur3g.gms .......................................................................................... 165
B.2 Data Reconciliation Programs .................................................................. 170
B.2.1 analyze.m ............................................................................................. 170
B.2.2 classify.m ............................................................................................. 176
B.2.3 redund.m .............................................................................................. 178
B.2.4 recon.m ................................................................................................. 178
B.2.5 consit.m ................................................................................................. 180
B.2.6 gdet.m .................................................................................................. 180
B.2.7 gaster.m ................................................................................................ 181
# List of Figures

1-1 Pathway view of process information flow ............................................. 18  
1-2 Relationships and levels of abstraction in bioprocess data reconciliation. . 20  
1-3 Project overview. .................................................................................... 25  
2-1 Overview of data reconciliation methodology. ........................................... 28  
2-2 Classification of parameter types. ............................................................. 29  
2-3 Hierarchy of reaction types required to model a bioprocess. .................... 32  
2-4 Variable observability and redundancy classification hierarchy. .............. 44  
2-5 Algorithm for observability and redundancy classification. ..................... 46  
2-6 The $\chi^2$ distribution for various degrees of freedom. ........................... 48  
2-7 Network representation of a typical bioprocess flow sheet. ....................... 51  
2-8 Strategy to manage assumptions used in reconciliation. .......................... 54  
3-1 Overview of gas-related bioprocess variables. ......................................... 56  
3-2 System overview of factors influencing RQ and its measurement. ............... 66  
3-3 Effects of operating conditions on the TQ/RQ ratio ................................. 68  
3-4 Oxygen removal as a function of operating conditions ............................. 71  
3-5 Expected NUR trend for a typical A. vinelandii run. ............................... 73  
3-6 Dependence of RQ on $Y_{XS}$ for selected substrates. ............................ 77  
3-7 Dependence of RQ on $Y_{PS}$ for selected products ................................. 78  
3-8 Dependence of $\sigma_{Y_{XS}}$ on $\sigma_{RQ}$, $Y_{XS}$, and $\gamma_X$ for cell growth on carbohydrate. ....... 79  
3-9 Dependence of $\sigma_{TQ}$ on the stripping rate and gas composition accuracy . 83  
3-10 Dependence of RQ uncertainty on pH .................................................. 85  
3-11 Influence of pH and OUR on the optimum aeration for RQ calculation ... 86  
4-1 Gas handling system for NUR measurement experiments. ....................... 94  
5-1 Schematic representation of industrial enzyme fermentation. ................. 100  
5-2 On-line measurements for Run K890. .................................................. 101  
5-3 $p$-values and suspect measurements for Run K890. ............................... 103  
5-4 The $p$-values for Run K890 after the $NH_3$ flow meter bias was removed . 104  
5-5 Measured and calculated values obtained from the BSROY2 experiment. ... 106  
5-6 Effect of various errors on the RQ for the B. subtilis experiment. .............. 108  
5-7 Calculated RQ trend for the B. subtilis experiment ................................. 109  
5-8 Comparison of measured and $k_{La}$-based DO values ............................. 111  
5-9 Medium nutrient analysis example ......................................................... 112  
5-10 Baseline variability in NUR measurement. ............................................ 114  
5-11 Dependence of NUR uncertainty on Argon concentration. ...................... 115  
5-12 Cell growth profiles observed in NUR experiments. ............................. 116  
5-13 Experimentally measured NUR trends. ................................................. 118
List of Tables

1.1 Process monitoring and analysis objectives ........................................ 16
2.1 Summary of balances and variables associated with network balances ...... 31
2.2 Applications of data reconciliation to bioprocesses .......................... 32
2.3 Data adjustment and parameter estimation research efforts ................. 39
2.4 Progression of variable classification research ................................. 42
2.5 Observability and redundancy classification overview ....................... 45
2.6 Progression of gross error detection and diagnosis research ............. 49
2.7 Description of blocks in bioprocess example .................................. 52
2.8 Summary of bioprocess balances and variables used in bioprocess example 52
2.9 Balances provided by common unit operations .................................. 53

3.1 Henry’s law and equilibrium constants for CO₂ and O₂ ..................... 59
3.2 Fitted coefficients used to predict $k_{La}$ in various reactors .............. 70
3.3 Values for $\gamma$, $RQ_R$, and $RQ_C$ for various compounds ............... 75
3.4 Heuristics for interpreting RQ calculations and trends ..................... 80

4.1 Summary of bioprocesses used in case studies ................................ 88
4.2 Medium for *Bacillus subtilis* experiments .................................... 88
4.3 Medium for ECSON experiments ..................................................... 90
4.4 Run conditions for NUR experiments .............................................. 92
4.5 Fermentation medium used in NUR experiments ............................... 93

5.1 Assumptions used in NUR gas mix optimization .............................. 114
5.2 Bioprocess model relations ......................................................... 129
5.3 Variable classifications for off-gas reconciliation example ................ 129

A.1 Composition of biologically derived nutrients ............................... 153
A.2 Estimates of cell yield on various elements .................................. 154
A.3 Estimates of cell yield on various amino acids ............................... 154
A.4 Estimates of cell yield on various vitamins .................................. 155
Chapter 1

INTRODUCTION

Data collected during the development and operation of bioprocesses often contain discrepancies that can not be attributed to random error. This research focuses on using analytical redundancy to detect these discrepancies and determine their causes.

1.1 Thesis Organization

This chapter describes the challenges of bioprocess data reconciliation and defines our research goals and approach. The role of reconciliation in data processing is explained using an analogy to a metabolic pathway. The chapter concludes with a description of the research methodology and a summary of the major contributions of the work.

Chapter 2 reviews the formulation of generic network balances and the statistical and algorithmic aspects of processing models and data. Chapter 3 develops the models associated with the gas train and describes their use in establishing analytical redundancy and optimizing calculation accuracy. Chapter 4 summarizes the materials and methods relevant to the experimental results in off-gas analysis and data reconciliation, which are reported in Chapter 5. Finally, Chapter 6 summarizes the conclusions and provides suggestions for future research.

1.2 Problem Description

Automated process monitoring systems in use today collect unprecedented amounts of data from research and manufacturing processes. These data are used to accomplish a variety of objectives, some of which are summarized in Table 1.1. Unfortunately, these data are often corrupted by undetected miscalibration, drift, failures, and human error. Use of the data is hindered by fear that the undetected problems will translate to
Table 1.1: Process monitoring and analysis objectives

<table>
<thead>
<tr>
<th>Environmental Optimization</th>
<th>Performance Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Development Medium Design Control Strategy</td>
<td>( \mu, X, Y_{XS} ) ( Q_P, P, Y_{PS} )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale Up Preparation</th>
<th>Variability Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxygen Requirements Sterilization Effects Shear Effects</td>
<td>Set-points Maintained Disturbances Data Consistency</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operational Decisions</th>
<th>Equipment Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Adjustments Induction Timing Harvest Timing</td>
<td>Mass transfer Mixing Fault Detection</td>
</tr>
</tbody>
</table>

faulty process control and decision making. Although these faults are difficult to separate from the inherent noise present in poorly-modeled bioprocesses, process experts often find sufficient redundancy to detect inconsistency and to identify the sources of error. To automate and extend this type of diagnostic approach, these redundancies – which take the form of assumptions, models, statistical correlations, and constraints – must be identified, organized, and evaluated.

### 1.3 Project Objectives

The goal of this research is to evaluate and improve the use of data reconciliation as a framework for bioprocess data analysis. This requires investigating experimental inconsistencies and evaluating models to establish redundancy, particularly in the area of off-gas analysis. The specific objectives of the project were as follows:

1. To identify the models, balances, correlations, and assumptions necessary to analyze data from a diverse set of bioprocesses.

2. To evaluate the use of nonlinear reconciliation as a means to consider a broader and more detailed set of potential faults.
3. To develop a method to classify the observability and redundancy of variables in nonlinear models.

4. To describe the effect of measurement accuracy, process stoichiometry, and operating conditions on respiratory quotient calculations.

5. To investigate the use of $k_{L}a$ correlations in the reconciliation model.

6. To evaluate the feasibility of measuring the rates of nitrogen uptake and $^{13}$CO$_2$ evolution using mass spectrometry.

Meeting the first goal required a data-driven approach to evaluate the methodology and models. The variety and authenticity of industrial and lab-scale data sets used presented new challenges that would not have surfaced if only simulation was used. The second and third objectives were needed to manage the varied information content of bioprocesses. The last three objectives were driven by the need to understand gas-related inconsistencies seen in the lab and industrial data.

1.4 Background

1.4.1 The Role of Data Reconciliation

The assessment of data quality is an important (and sometimes neglected) step in the process information pathway (Figure 1-1). This paradigm, which is analogous to a metabolic pathway, illustrates the steps by which data analysis improves process understanding and control. In this context, the role of data consistency analysis can be clarified as it relates to process monitoring and control. The key point of the pathway is the numerous opportunities for data to be lost, either by not being monitored, processed, or used to make decisions. The location of the bottleneck in the pathway depends on the type of process and the stage of the product’s life cycle. Many processes lack sensors that can monitor and control key variables. Alternatively, the control step sometimes lacks the models upon which to build a control strategy. Nevertheless, the bottleneck often lies in the intermediate links whereby data sets are transformed to useful information and consequently, the large effort required to collect good data goes to waste.
Figure 1-1: Pathway view of process information flow.
One way to assess data quality is to use the analytical redundancy provided by process models (Madron et al., 1977; Wang & Stephanopoulos, 1983; van der Heijden et al., 1994b). Typically this is done by modeling cell metabolism using elemental balances (Cooney et al., 1977). A summary of literature in this area is shown in Table 2.2 of Chapter 2.

Researchers often note the complexity of bioprocesses and the limited number of models and measurements available to study them. The large number of variables and limited number of equations and models typically create many degrees of freedom. Many of the raw measurements are either present in multiple variables or absent from the elemental balances equations. As a result, errors in these measurements are either impossible to detect or difficult to isolate. The degrees of freedom can be eliminated in subsets of the model through the use of additional assumptions and less reliable model relations. Figure 1-2 illustrates numerous opportunities for establishing redundancy for use in data reconciliation. At the top level we can develop correlations describing mass transfer, broth properties, and mechanical equipment performance. At the next level we can evaluate balances on carbon, nitrogen, degree of reduction, energy, and charge as well as chemical and vapor-liquid equilibria. Finally, we can model the detailed metabolism of the cell and the kinetics of cell and product formation. Use of additional constraints in the reconciliation model, such as correlations of $k_La$ against operating parameters, can help evaluate the consistency of additional raw measurements. The additional models also provide information to differentiate between alternate causes of inconsistencies. The downside to this is the risk of introducing additional systematic errors that confound detection of the faults of interest.

One specific aspect of process modeling that drew our attention in this work is that of the respiratory quotient (RQ). We saw in this widely used parameter inconsistencies in elemental balance reconciliation that pointed to modeling errors related to RQ. Further background on RQ is described below.
Figure 1-2: Relationships and levels of abstraction in bioprocess data reconciliation.
1.4.2 Respiratory Quotient Interpretation

While consuming substrates for energy and raw material, cells exchange $\text{O}_2$ and $\text{CO}_2$ with the environment. The ratio of $\text{CO}_2$ produced to $\text{O}_2$ consumed, referred to as the respiratory quotient (RQ), can provide an indication of the compositions and amounts of components involved in metabolism. RQ is an attractive metabolic indicator because it derives from relatively accessible and reliable gas composition measurements that can be measured simultaneously by mass spectrometry or obtained individually using gas analyzers. However, three factors limit the usefulness of RQ as a metabolic indicator. First, similarities in the degrees of reduction (Erickson et al., 1978a) of carbohydrate, product, and biomass can limit RQ use in discriminating among the different pathways. Second, if the amount of product the cells produce is materially insignificant, the resulting impact on RQ can be too small to measure. Finally, in contrast to the acidic pH maintained in *Saccharomyces cerevisiae* cultivation (a common research model system), industrially-relevant *Escherichia coli* and mammalian cell lines are cultivated under pH neutral conditions where dissolved $\text{CO}_2$ dynamics are significant. Besides limiting RQ accuracy, dissolved $\text{CO}_2$ fluctuations also create steady-state and transient RQ bias that may inadvertently be interpreted as metabolic events.

Use of RQ in process control (O’Connor et al., 1992) can improve process operation but success requires detection of measurement and equipment failures along with identification of invalid assumptions. This detection task is difficult to perform early in a batch when indirect measurements such as RQ and oxygen uptake rate are relatively uncertain. Interpreting RQ variability requires understanding its sensitivity to measurement uncertainty as well as metabolic and operational changes. This has motivated us to develop a systematic method for evaluating RQ data.

1.4.3 Monitoring of Nitrogen Uptake

The Haber process for ammonia synthesis fixes about $10^8$ tons yr$^{-1}$ of $\text{N}_2$ gas (Jolly, 1964) with a similar amount fixed ecologically (The National Research Council Subcommittee on Ammonia, 1979). In order to study nitrogenase, the enzyme responsible for biochemical $\text{N}_2$ fixation, researchers need to produce the enzyme
efficiently and economically at large scale (Hamel et al., 1990). In our work, *Azotobacter vinelandii* was grown aerobically in nitrogen-free medium to produce the enzyme.

Enzyme production and degradation rates are a function of cultivation conditions. *A. vinelandii* consumes oxygen at an exceptionally high specific rate (Petersen, 1991). This reflects the unique system of respiratory control the organism has developed to protect the enzyme from damage (Yates & Jones, 1974).

The conversion of acetylene to ethylene (Postgate, 1972) by nitrogenase provides the primary means of assaying enzyme activity. These assays are done on an intermittent basis with the results becoming available about 1-2 h later. Although several direct (Bergersen, 1980) and indirect (Turner & Gibson, 1980) methods can quantify nitrogen fixation, the method demonstrated here in Section 5.3 is the first to monitor enzyme activity on-line and *in situ*. The delicate balance between nitrogenase synthesis and degradation requires controlled cultivation conditions and careful selection of harvest time. Overall process development and particularly the selection of harvest timing would be more efficient and effective if the enzyme activity could be measured continuously.

The stoichiometry of the nitrogen fixation reaction (Orme-Johnson, 1985) is generally considered to be:

\[
N_2 + 8H^+ + 16ATP + 8e^- \rightarrow 2NH_3 + H_2 + 16ADP + 16P_i
\]

Measuring the amount of *H₂* produced by this reaction could provide an on-line nitrogenase assay; unfortunately the *H₂* gas is scavenged by reactions occurring within the cell (Robson & Postgate, 1980). We expect the enzymatic activity of nitrogenase acting upon *N₂* to be one-third of its activity on acetylene because acetylene reduction requires the net transfer of two electrons to the substrate while *N₂* reduction to *NH₃* requires six electrons. This is based upon the assumption that electron and substrate transport do not limit the reaction rate (Postgate, 1972). Our goal was to test the hypothesis that enzyme activity can be monitored with off-gas analysis of *N₂* consumption.

22
Due to the low rate of N₂ consumption and its high concentration in air, measuring N₂ uptake rate is a challenging task. Calculation of gas exchange rates are based on the flow and compositions of the inlet and outlet gas streams. Nitrogen, which typically serves as the reference for calculating outlet gas flow, is inappropriate here because it is not inert. Atmospheric argon would be a logical substitute but the low Ar concentration (1%) combined with the high gas flow rate hinders accurate measurement of the minute nitrogen uptake (< 2 mmol L⁻¹ h⁻¹). Sisler and ZoBell (1951) first noted changes in the N₂-to-Ar ratio as a qualitative indicator of nitrogen fixation when, examining hydrogen utilizing bacteria, they found the unanticipated N₂ fixation reduced the gas volume of vials that were intended to be inert controls.

The inadequacy of air as a feed gas required us to design a gas mixture that would maximize the NUR signal-to-noise ratio while allowing control of dissolved O₂ and N₂ levels within desired ranges. The fraction of N₂ that the cells strip from the gas stream, ordinarily less than 0.2%, increases if agitation instead of aeration is used to increase oxygen transfer. Doubling the inlet O₂ concentration and halving the N₂ concentration further improved the accuracy of the NUR calculation. The consequent reduction in the gas flow required to meet the oxygen demand increases N₂ removal five-fold from 0.2 percent to 1 percent. Still, this is a small quantity compared to the amount of O₂ removed, which can reach as high as 75% in high density E. coli cultures.

The capability of the Perkin Elmer MGA 1600 mass spectrometer to measure H₂ and Ar concentrations allowed us to calculate the NUR and H₂ evolution rates (HER). An on-line optical density sensor enabled us to continuously calculate the specific enzyme activity, which was used to select the harvest time.

1.4.4 Monitoring of ¹³CO₂ Evolution

Off-gas mass spectrometers monitor the consumption and production of gases (typically CO₂ and O₂) by cells growing in a reactor. One measure of the accuracy of a mass spectrometer is its unit resolution level. This is defined as the mass-to-charge ratio (m/e) at which signal overlap exceeds 10% of the peak height. After an equipment upgrade increased our unit resolution capability from 40 to 60, we decided to test the
feasibility of obtaining separate measurements of the concentrations of $^{13}$CO$_2$ (m/e 45) and $^{12}$CO$_2$.

Measuring the relative evolution rates of $^{12}$CO$_2$ and $^{13}$CO$_2$ from a bioreactor provides us with a novel means to monitor variables such as substrate concentration, substrate utilization kinetics, and metabolic pathway fluxes. Several researchers have measured $^{14}$CO$_2$ by radio-respirometry (Wang, 1963; Zagallo & Wang, 1967; Kelleher & Bryan, 1985; Chu et al., 1985; Bruinenberg et al., 1986; Schukat & Janke, 1988; Dunstan et al., 1990), but unfortunately, their measurements are made off-line and require special handling for radioactive materials. Although mass spectrometry monitoring of $^{13}$CO$_2$ has seen recent use in human studies of substrate consumption (Gautier et al., 1993) and CO$_2$ wash-out (Zanconato et al., 1992), we are unaware of it being applied to fermentation or cell culture.

1.5 Research Methodology

The overall flow of the project is shown in Figure 1-3. To identify the primary problems in bioreactor data analysis, we studied approximately 140 batches drawn from eight processes described in Table 4.1 of Chapter 4. The reactors, seven types in all, each had unique monitoring and control capabilities. We conducted 27 experiments at laboratory-scale to perform real-time analysis and directly investigate inconsistent data. The two large data sets were provided by collaborators working in process development and manufacturing settings. This mix provided realistic scenarios to model the problems faced in research and industry.

In examining data from each case study, a series of common questions emerged:

1. Given the process measurements available, what relationships (see Figure 1-2) allow us to check the consistency of measurements or estimate important variables?

2. What simplifying assumptions need to be made to increase the set of applicable relationships?

3. Can potential causes of data inconsistency be identified using data reconciliation?

4. Is the aggregated data (e.g., batch mass balances) consistent?
Figure 1-3: Project overview.
5. Can the source of the inconsistency be isolated in time by analyzing the data on a cumulative basis and over given time-periods?

6. Can the hypothesized cause(s) be confirmed so that we can evaluate the methodology?

As we attempted to answer these questions, we were able to evaluate the reliability of the models and their utility in process diagnosis.

1.6 Thesis Contributions

- We were the first to explore the use of nonlinear data reconciliation in analysis of bioprocess data. We found this allowed a broader and more detailed set of variables to be analyzed than in previous efforts.

- We extended an algorithm for variable classification to enable the use of nonlinear models and also defined additional classes to characterize redundancy in greater detail.

- We developed novel strategies to measure and use nitrogen uptake rate (NUR) and $^{13}$CO$_2$ evolution rate.

- We examined the calculation and estimation of the respiratory quotient (RQ) to improve its accuracy and interpretation.

- We demonstrated the use of additional redundancy gained from $k_{L}a$ correlations in data reconciliation and diagnosis of dissolved oxygen probe faults.
Chapter 2

DATA RECONCILIATION

METHODOLOGY

Kuehn and Davidson introduced the use of data reconciliation (DR) for process data analysis to the chemical engineering field in 1961. Reviews of the ongoing development of the DR methodology include those by Hlavacek (1977), Mah et al., (1981; 1985; 1990), and Crowe (1996). The goals of data reconciliation are: (1) to assess whether collected data are statistically consistent with the model of the process, (2) to calculate optimal measurement adjustments in order to obtain a consistent and more accurate set of data, (3) to obtain the best estimate of unmeasured quantities, and (4) to detect and identify measurement and process faults. The main steps in the method are model formulation, variable classification, parameter estimation, and gross error handling. This chapter describes these steps, illustrated in Figure 2-1, and reviews the literature relating to the topic.

2.1 System Representation

The first step in data reconciliation is to assemble a set of mathematical relationships that describe the process and its measurements. These relationships typically consist of equality constraints approximating the relationships among the real and lumped parameters of the chemical process. Examples include conservation balances, empirical correlations, and physical property relations. This section describes the models used in data reconciliation and describes the variables and relations associated with a typical bioprocess.

When a set of over-determined model equations, $f$, are evaluated with a set of parameters, $z$, the resulting residuals, $\epsilon$, are attributable to errors in the structure of the models ($\epsilon_s$) and in the parameters ($\epsilon_z$).
Figure 2-1: Overview of data reconciliation methodology.
There are several categories of gross errors that can occur. Particular measurements or assumptions can have a bias (miscalibration) or a variance different from that assumed in the model. Process leaks, neglected process dynamics, or unidentified reaction participants are examples of structural modeling errors. Models selected for use in data reconciliation are typically limited to those with negligible error in structure:

\[
f(\tilde{z}) = f(z + \delta_z) = \epsilon_z + \epsilon_z \tag{2.1}
\]

\[
f(\hat{z}) = f(\hat{z} + \hat{\delta}) = 0 \tag{2.2}
\]

If necessary, the modeling error can be represented as additional variables with known covariance as in the approach taken by Romagnoli (1983).

As shown in Figure 2-2, model parameters can be subdivided into six classes depending on if they are fixed or time-varying and to what extent their values are known. The assignment of parameters to particular classes depends on the amount of information available and whether a linear or nonlinear model is chosen. Certain and uncertain parameters can be further classified as redundant or (just) measured and similarly, unknown parameters can be classified as observable or unobservable.

The models used in data reconciliation can be divided into:

1. Balances (mass, component, energy, charge)
2. Reactions (lumped, detailed, stoichiometric, elemental balanced, metabolic, yield)

3. Empirical constraints (vapor-liquid equilibrium, reaction kinetics, transport, physical properties, equipment)

2.1.1 Network Balances

Consider a generic non-steady-state process network where $J$ streams convey $C$ components (chemical species) among a set of $K$ unit operations. This could represent a refinery or a bioprocess model.

Each unit must satisfy one energy balance and a set of $C$ component balances. This leads to $K(C+1)$ equations, with $K(C+1)$ accumulations and $J(C+1)$ flows comprising the unknowns. While Duham's Theorem (Prigogine & Defay, 1954) states $C+2$ parameters are required to specify a system, here we enumerate only $C+1$ balances since the remaining balance requires process-specific information (e.g., pressure drops). Next, we allow each unit the option of containing $R$ stoichiometric reactions and one lumped reaction constrained by $E$ elemental balances. The result is a total of $K(C+E+1)$ balances with $((K + J)(C + 1) + KC)$ variables. These balances can be organized into block matrix form using an approach similar to that of Crowe (1983):

\[
\begin{align*}
\begin{array}{ccc}
\text{input-output} & \text{reaction} & \text{accumulation} \\
\widehat{Bu} & - S^T \xi - \phi & = \dot{\omega} \\
\text{DT} & \phi & = 0 \\
Ah^* & = \dot{h}^u
\end{array}
\end{align*}
\] (2.3)

In practice, only a subset of these species are present in a given unit and its connecting streams. Consequently, many of the above equations take the trivial form $0=0$. The dimensions of the balances after the removal of the trivial equations are determined by defining and combining the sub-totals for each of the units ($k \in 1..K$):

- $R_k = \#$ of defined reactions in unit $k$
- $E_k = \#$ of elemental balances in unit $k$
- $C_k^u = \#$ of components present in unit $k$ and its connecting streams
- $C_k^{ue} = \#$ of elementally-balanced components in unit $k$
- $C_k^s = \#$ of components present in stream $j$
Table 2.1: Summary of balances and variables associated with network balances

<table>
<thead>
<tr>
<th>Balances</th>
<th>type</th>
<th>number</th>
<th>Unknowns</th>
<th>type</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>energy</td>
<td></td>
<td>enthalpy flows</td>
<td></td>
<td>J</td>
</tr>
<tr>
<td></td>
<td>component</td>
<td>( \sum_k C_k^u )</td>
<td>enthalpy accumulations</td>
<td>( K )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>elemental</td>
<td>( \sum_k E_k )</td>
<td>component flows</td>
<td>( \sum_j C_j^s )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>component accumulations</td>
<td>( \sum_k C_k^{au} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>reaction rates</td>
<td>( \sum_k R_k )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>consumption rates</td>
<td>( \sum_k C_k^{ue} )</td>
<td></td>
</tr>
</tbody>
</table>

\[ \sum_k R_k = \text{sum of defined reactions} \]
\[ \sum_k E_k = \text{sum of elemental balances} \]
\[ \sum_k C_k^u = \text{sum of components totals} \]
\[ \sum_k C_k^{ue} = \text{sum of elementally-balanced components} \]
\[ \sum_j C_j^s = \text{components in all } J \text{ streams} \]

The component accumulations \( (\omega_k) \) and flows \( (u_j) \) for each unit and stream are stacked into two block vectors \( \omega \) and \( u \). Matrix \( B \) is derived from the \( K \times J \) flow sheet incidence matrix, \( A \), by replacing each entry with a corresponding \((C \times C)\) identity or null matrix and then deleting the individual columns and rows for the components not present in the associated streams and units. The resulting matrix has \( \sum_k C_k^u \) rows and \( \sum_j C_j^s \) columns. The block matrices \( S \) and \( D \) are built from the corresponding reaction stoichiometry \( (S_k) \) and component composition \( (D_k) \) matrix of each unit. The matrix \( T \) is constructed to select the subset of \( \sum_k C_k^{ue} \) reacting components for use in the elemental balances. Table 2.1 summarizes the equations and unknowns associated with this representation. This provides a means for quickly identifying the dimensionality of the system.

2.1.2 Reaction Representation

Figure 2-3 provides a hierarchical view of the types of representations used to model the reactions occurring in a bioprocess. They are described here to show where nonlinear data reconciliation is necessary and to describe our work in media analysis. The primary distinction is between reactions characterized as at equilibrium \( (e.g., \text{ions in the media}) \) and those characterized by their rate or extent.
Table 2.2: Applications of data reconciliation to bioprocesses

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977</td>
<td>Madron et al.</td>
<td>Applied DR to bioprocess data</td>
</tr>
<tr>
<td>1978a</td>
<td>Erickson et al.</td>
<td>Used degree of reduction in mass balancing</td>
</tr>
<tr>
<td>1978b</td>
<td>Erickson et al.</td>
<td>Used degree of reduction to estimate product</td>
</tr>
<tr>
<td>1979</td>
<td>Madron</td>
<td>Discussed observability and redundancy concepts in fermentation elemental balance</td>
</tr>
<tr>
<td>1980</td>
<td>Dekok and Roels</td>
<td>Added heat balance to Madron (1977)</td>
</tr>
<tr>
<td>1988</td>
<td>Tsai and Lee</td>
<td>Used metabolic reactions in DR</td>
</tr>
<tr>
<td>1989</td>
<td>Vallino and Stephanopoulos</td>
<td>Used Metabolic network, inequality constraints, DR applied to lysine production</td>
</tr>
<tr>
<td>1989</td>
<td>Liao</td>
<td>Proposed empirical reaction subspace; industrial application</td>
</tr>
<tr>
<td>1994</td>
<td>van der Heijden et al.</td>
<td>Applied matrix projection classification</td>
</tr>
<tr>
<td>1994</td>
<td>van der Heijden</td>
<td>Used Vector comparison for GEI</td>
</tr>
<tr>
<td>1994</td>
<td>van der Heijden</td>
<td>Proposed method for multi-sample DR (accumulate error, project primary to rate errors)</td>
</tr>
<tr>
<td>1996</td>
<td>Noorman et al.</td>
<td>Applied van der Heijden methods to more complex example</td>
</tr>
</tbody>
</table>

Figure 2-3: Hierarchy of reaction types required to model a bioprocess.
The multitude of reactions that make up cell metabolism can be modeled by several representations. One representation is the use of a single reaction with variable stoichiometric coefficients calculated from the measured data and elemental balance constraints (Cooney et al., 1977; Wang & Stephanopoulos, 1983). A second option is to replace this representation with a set of reactions with fixed stoichiometric coefficients but unknown relative rates. Researchers use three types of formulations to represent fixed-coefficient reactions. One type is a set of lumped, analytically derived reactions representing respiration, cell mass production, and product formation (Grosz et al., 1984). The second type replaces these lumped reactions with a detailed set of metabolic reactions (Tsai & Lee, 1988; Vallino & Stephanopoulos, 1989). The third type uses empirically derived reactions to represent statistical regularities found in historical data (Liao, 1989).

The first and second type often have equivalent degrees of freedom. The additional equations present in the second type are offset by an equal number of new unknowns. Although the metabolic network representation introduces intermediate metabolites, most are eliminated using pseudo steady-state assumptions. The remainder of the intermediate metabolites can either be estimated or treated as pseudo-measurements. That is, we assign a value and uncertainty and allow the reconciliation algorithm to test the validity of the assumed value in the same way it tests the measured values. The directionality or availability of particular pathways can provide constraints absent from the elemental balance representation. For example, the requirement of concomitant CO₂ consumption during metabolism of ethanol (Calvin cycle) is missing from the elemental balance constraints. Liao's approach, which is based on singular-value decomposition, sometimes generates fewer empirical reactions than the analytical representation; this reflects physical constraints implicit in the data but missing from the analytical model.

The next layer in the hierarchy (Figure 2-3) is the additional consideration of varying component compositions and coefficients. For example, the nitrogen content of a cell varies from strain to strain and over the course of a culture. The resulting mismatch between the model and the data either must be assigned to the rate of consumption of the nitrogen source or to an adjustment of the composition estimate.
The latter approach requires a nonlinear reconciliation.

2.1.3 Nutrient Yields

The uncertain nitrogen composition mentioned above can be thought of as an uncertain yield of cell mass on nitrogen ($Y_{XN}$). The advantage of using yield instead of composition is that it allows the uncertainty to reflect the substitution of alternative materials, unused substrates, unmeasured substrates, or unrecognized byproducts. When the spectrum of nutrients required for cell growth and product formation is considered, we find that the respective yields on nutrients (other than carbon, nitrogen, and oxygen) are often imprecise. The nature of the biotechnology industry is such that processes are rapidly scaled-up to use different growth media, higher cell densities, and modified operating conditions. In this procedure, imprecise yields can lead to a medium design with an unintended imbalance between nutrient supply and cell demand which limits process performance.

Nutrient yields reported in the literature are determined by measurement of cell composition or by experimental observation of incremental improvements in cell yield after nutrient supplementation (Suzuki et al., 1985). Data compiled by Truong (1993) are summarized in Appendix A.2. Table A.1 lists the composition of several complex nutrients while Tables A.2, A.3, and A.4 list the range of expected yields on various elements, amino acids, and vitamins respectively.

When we use data reconciliation to troubleshoot a chemical process, the model usually assumes that the cells are growing at their nominal rate or are constrained by the rate of addition of the carbon or nitrogen source. When an unidentified nutrient is limiting, the assumption is invalidated. If this possibility is absent from the set of potential faults examined, then the diagnosis will either be inconclusive or incorrect. Our approach to this problem was to canvas the literature to obtain a set of estimates of the upper and lower bounds for cell yield on common nutrients. These yields can be combined with the target cell density for a culture to create a rough estimate of the upper and lower bounds on the concentration needed for each component. We then use these estimates to screen the amount of each nutrient in a particular medium to
ascertain if its concentration falls outside this range. Here $Y_{x_j}^{\text{max}}$ and $Y_{x_j}^{\text{min}}$ represent the maximum and minimum yield of element $j$ on a cell (g·g$^{-1}$) and $C_j^{\text{min}}$ and $C_j^{\text{max}}$ represent the corresponding available concentrations of element $j$ in the culture medium.

$$C_j^{\text{min}} = \frac{X_{\text{target}}}{Y_{x_j}^{\text{max}}}$$  \hspace{1cm} (2.4)

$$C_j^{\text{max}} = \frac{X_{\text{target}}}{Y_{x_j}^{\text{min}}}$$  \hspace{1cm} (2.5)

Based on the above information, we compare the concentrations of each nutrient to that necessary to obtain the desired cell mass. This provides a set of candidates to evaluate as potentially limiting cell growth.

### 2.2 Parameter Estimation

The objective of the parameter estimation step is to estimate the measurement errors and the values of the unmeasured variables. Table 2.3 summarizes the literature in this field. Characteristics that differentiate the various efforts include the model type (i.e., linear or nonlinear, general or specific), measurement structure (complete or partial), and analysis window (steady state or dynamic, snapshot or multi-observation).

The goal is to find the set of parameters ($\mathbf{z}$) that maximize the likelihood function ($L$) for a set of process models ($f$), a set of measurements ($\hat{\mathbf{x}}$), and the measurement variance-covariance ($\Psi$):

$$\max_{\hat{\mathbf{x}}} L((\hat{\mathbf{x}}, \hat{\mathbf{y}})|(\mathbf{x}_m, \Psi)) \text{ subject to } f(\hat{\mathbf{x}}) = 0$$  \hspace{1cm} (2.6)

If the distribution of random errors is multivariate Gaussian, then the optimum solution is equivalent to that provided by weighted least squares:

$$\min_{\hat{\mathbf{x}}} \frac{1}{2} \hat{\mathbf{y}}^T \Psi^{-1} \hat{\mathbf{y}} \text{ subject to } f(\hat{\mathbf{x}}) = 0$$  \hspace{1cm} (2.7)
This is converted to an unconstrained form using Lagrangian multipliers ($\lambda$):

$$
\min_{\hat{x}, \lambda} h(\hat{x}, \lambda) = \frac{1}{2} [\hat{x} - \hat{x}]^T \Psi^{-1} [\hat{x} - \hat{x}] + \lambda^T f(\hat{x})
$$

(2.8)

with the optimum occurring at the stationary point

$$
\frac{dl}{d\hat{x}} = 0 = \Psi^{-1} \hat{x} + J^T \lambda
$$
$$
\frac{dl}{d\lambda} = 0 = f(\hat{x})
$$

(2.9)

where $J$ is the Jacobian of $f$.

### 2.2.1 Linear Models

If the constraints are linear, then the solution can be obtained in one step by solving the above equations, shown here in matrix form:

$$
\begin{bmatrix}
\Psi^{-1} & A^T \\
A & 0
\end{bmatrix}
\begin{bmatrix}
\hat{x} \\
\lambda
\end{bmatrix}
= 
\begin{bmatrix}
\Psi^{-1} \hat{x} \\
0
\end{bmatrix}
$$

(2.10)

If all the variables are measured, the solution to the above equations (Kuchn & Davidson, 1961) is given by

$$
\hat{\delta} = \Psi A^T (A \Psi A^T)^{-1} \hat{\epsilon}
$$

(2.11)

$$
\hat{x} = [I - \Psi A^T (A \Psi A^T)^{-1} A] \hat{x}
$$

(2.12)

Otherwise, the constraint matrix must be partitioned and projected to eliminate the coefficients of the unmeasured variables:

$$
P (A_m x_m + A_c x_c) = P A_m x_m = R x_m = 0
$$

(2.13)

The projection matrix $P$ can be found by Gaussian elimination or by calculating $x_c$ from the pseudo-inverse (Strang, 1988) of $A_c$, and substituting the result into the
original constraint equation (van der Heijden et al., 1994a):

\[
x_c = -A_c^* A_m x_m \tag{2.14}
\]

\[
0 = A_m x_m - A_c A_c^* A_m x_m \tag{2.15}
\]

\[
P = (I - A_c A_c^*) \tag{2.16}
\]

This decouples the measurement adjustment from the calculation of the unknown variables. Matrix \( R \) summarizes the redundant relations among the measured variables and is analogous to Equation 2.12:

\[
\hat{x}_m = [I - \Psi R^T (R \Psi R^T)^{-1} R] \hat{x}_m \tag{2.17}
\]

This approach applies only if the statistical basis is valid; therefore we must detect and identify gross errors in the data before we can use the resulting estimates.

### 2.2.2 Nonlinear Models

Retaining the nonlinear structure of the model is desirable because it facilitates uncertainty assignment and gross error identification. Unfortunately the methods for reconciliation and variable classification lack the flexibility and rigor of those used in the linear case. For nonlinear models an iterative solution is required. Typically the solution is found by repeatedly linearizing the model around the current estimate of the solution until convergence is reached:

\[
f(x) = 0 \tag{2.18}
\]

\[
f(\hat{x}) \approx J(\hat{x} - x) + f(x) \approx J(\hat{x} - x) \tag{2.19}
\]

\[
J_m \Delta x_m + J_c \Delta x_c = -f(\hat{x}) \tag{2.20}
\]

Kuhen and Davidson (1961) used an algorithm based on Newton's method to find the solution:

\[
x^{i+1} = \hat{x}^i - \Psi J(J \Psi J)^{-1} f(\hat{x}^i) \tag{2.21}
\]
Britt and Luecke (1973) tailored Newton's method for simultaneous estimation and reconciliation. Their algorithm, which constitutes the basis for most work in nonlinear reconciliation, was subsequently described and tested by Knepper and Gorman (1980). The algorithm only applies if \( m \geq e > c \). Since many reconciliation models do not meet this criteria, the models must first be partitioned to eliminate dependent variables. For example, analyzing data from three tanks in series with measured first and last streams requires aggregation of the intermediate nodes such that \( (m=2, e=3, c=2) \) becomes \( (m=2, e=2, c=1) \).

Solutions based on general nonlinear programming algorithms such as Sequential Quadratic Programming (SQP) (Powell, 1978) are robust and flexible but computationally costly. To circumvent this, Tjoa and Biegler (1992) tailor SQP for the quadratic objective function and obtain faster convergence properties. Tjoa and Biegler also use a corrupted Gaussian distribution to model both random and systematic errors. This approach performs the reconciliation and gross error detection in one step. Johnston and Kramer (1995) further generalize this strategy.

This research, like most of the data reconciliation efforts, analyzes snapshots of steady-state and time-averaged data. Data from multiple time steps can be handled in a manner analogous to Kalman filtering (Wang & Stephanopoulos, 1983; Stanley & Mah, 1977). Alternately, the data can be processed simultaneously in one batch, as in the work of Rollins and Devanathan (1993), by using the raw mass measurements available at each time step instead of calculated flow rates and accumulations.

### 2.3 Measurement and Uncertainty Representation

Another contrast highlighted in the literature is the method that investigators use to derive and represent uncertainty. Most examples use calculated rate data obtained from the literature. They assume the variance-covariance of the rate data is diagonal and often assign an arbitrary relative uncertainty to all the measurements. This approach has two limitations which we have sought to avoid as described below.
<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1961</td>
<td>Kuehn and Davidson</td>
<td>Introduced DR to chemical engineering</td>
</tr>
<tr>
<td>1973</td>
<td>Britt and Luecke</td>
<td>Developed method for nonlinear error-in-variables estimation</td>
</tr>
<tr>
<td>1977</td>
<td>Stanley and Mah</td>
<td>Used temporal and spatial redundancy</td>
</tr>
<tr>
<td>1977</td>
<td>Madron et al.</td>
<td>Included method for $\Psi$ calculation for indirect measurements</td>
</tr>
<tr>
<td>1983</td>
<td>Crowe et al.</td>
<td>Developed matrix projection method for linear reconciliation</td>
</tr>
<tr>
<td>1983</td>
<td>Romagnoli</td>
<td>Represented model noise as well as measurement noise</td>
</tr>
<tr>
<td>1986</td>
<td>Crowe</td>
<td>Adapted matrix projection for specific bilinear terms</td>
</tr>
<tr>
<td>1986</td>
<td>Serth et al.</td>
<td>Used Britt et al., (1973) for reconciliation</td>
</tr>
<tr>
<td>1988</td>
<td>Pai and Fisher</td>
<td>Used Broyden’s method for nonlinear reconciliation</td>
</tr>
<tr>
<td>1988</td>
<td>Simpson et al.</td>
<td>Eliminated dependent variables using graph techniques</td>
</tr>
<tr>
<td>1991</td>
<td>Darouch and Zasadzinki</td>
<td>Extended DR to linear dynamic systems</td>
</tr>
<tr>
<td>1991</td>
<td>Tjoa and Biegler</td>
<td>Used bivariate Gaussian and SQP decomposition for simultaneous reconciliation, estimation, and gross error detection</td>
</tr>
<tr>
<td>1993</td>
<td>Meyer et al.</td>
<td>Used graph-based decomposition</td>
</tr>
<tr>
<td>1993</td>
<td>Rollins and Devanathan</td>
<td>Proposed alternative to Darouch and Zasadzinki (1991) formulation &amp; comparison</td>
</tr>
<tr>
<td>1995</td>
<td>Johnston and Kramer</td>
<td>Extended Tjoa &amp; Biegler, utilized historical data hyperbolic basis function pdf, robust regression, optimal model</td>
</tr>
<tr>
<td>1995</td>
<td>Albuquerque and Biegler</td>
<td>Formalized general nonlinear DAE DR</td>
</tr>
</tbody>
</table>
2.3.1 Limitations of Linear Representation

The first limitation is the inability to trace the fault when a calculated rate is suspect. This problem arises because a rate may be derived from multiple raw measurements taken at two or three time points. For example, a specific sugar consumption rate is calculated from two reactor concentrations, two reactor volumes, two feed-vessel weights, and a feed-vessel concentration. Some raw measurements such as reactor volume and gas flow impact rates in adjacent periods and multiple rates in a given period. The consequent indirect manifestation of measurement faults make it difficult to identify the actual source of the fault. Although Wang and Stephanopoulos (1983) demonstrated that the non-diagonal term can be neglected in the case of small gas flow error, neglecting large errors leads gas flow faults to be assigned to other variables, such as carbon source feed. This is because the two erroneous rate calculations (OUR and CER) inappropriately corroborate each other.

The second limitation of using rate data is the difficulty it creates in estimating the variance-covariance matrix, which represents the uncertainty and correlation of the data. Arbitrary assignment of relative error can be problematic since the reference value can be ambiguous or time varying. Also, while the rates obtained from an exponentially growing culture increase by several orders of magnitude, the accuracy of the physical measurement systems do not change proportionately. When looking at snapshots of data over the course of an exponentially growing culture, we found it more appropriate to model the uncertainty with two components, one proportional to the magnitude of the value and the other representing the lower limit of the absolute accuracy of the measurement. These values are more easily selected from equipment specifications, calibration data, and engineering judgment. The raw measurement uncertainty can be used directly in a nonlinear model formulation. To estimate the accuracy of derived calculations (e.g., metabolic rates), the raw measurement uncertainties must be combined using error propagation (e.g., Eq. 3.50). The rigorous uncertainty calculation is important, for example, in the case of oxygen uptake rate measurements where accuracy depends on the accuracy of the gas flow and concentration differentials which in turn depend on operating conditions.
2.3.2 Equations to Estimate Uncertainty

Errors in measurements can be due to random noise or to sensor bias, drift, or failure. Measurement noise ($\delta_n$) is often assumed to be normally distributed. A similar approach can be used to represent the variation in routine batch-to-batch calibration bias ($\delta_b$):

\[
\delta = \tilde{x} - x = \delta_n + \delta_b \tag{2.22}
\]

\[
E[\delta] = 0 \tag{2.23}
\]

\[
\delta_n \sim N(0, \Psi) \tag{2.24}
\]

\[
\delta_b \sim N(0, \Psi_b) \tag{2.25}
\]

When analyzing a snapshot of data, these two properties are indistinguishable. The measurement error ($\delta$) leads the equality constraints to have residuals ($\epsilon$) which have a covariance $\Phi$:

\[
\epsilon = A\tilde{x} = A\delta \tag{2.26}
\]

\[
\Phi = \text{cov}(\epsilon) = AA^T \tag{2.27}
\]

The process of reconciling data leads to an improvement in measurement accuracy:

\[
\text{cov}(\tilde{x} - x) = \Psi \tag{2.28}
\]

\[
\text{cov}(\tilde{x} - x) = \Psi [I - A^T(A\Psi A^T)^{-1}A] \tag{2.29}
\]

When some variables are not measured, the redundancy matrix, $R$, replaces $A$ and an additional equation calculates the uncertainty of the estimates:

\[
\text{cov}(\tilde{x}_m - x_m) = \Psi(I - R_r^T(R_r\Psi R_r^T)^{-1}R_r\Psi) \tag{2.30}
\]

\[
\text{cov}(\tilde{x}_c - x_c) = A_c^#A_m\Psi A_m^T(A_c^#)^T \tag{2.31}
\]

For nonlinear formulation, Britt and Luecke (1973) provide estimates of the resulting
covariance using their formulation:

\[
\text{cov}(\hat{x}_c - x_c) = (J_c^T Q J_c)^{-1} \tag{2.32}
\]

\[
\text{cov}(\hat{x}_m - x_m) = \Psi - \Psi J_m^T Q J_m \Psi + \Psi J_m^T Q J_c (J_c^T Q J_c)^{-1} J_c^T Q J_m \Psi \tag{2.33}
\]

We are not aware of similar relationships for general NLP methods beyond what can be obtained through the use of Monte-Carlo simulation.

2.4 Variable Classification

The purpose of this step is to decide which variables can be calculated and which measurements can be checked for errors. Vaclavek (1969) first recognized the need for variable classification. He defined the four basic variable classes (measured, redundant, observable, and unobservable) and developed algorithms for specific linear and nonlinear (Vaclavek & Loucka, 1976) systems. Stanley and Mah (1981a) developed 15 theorems summarizing the concepts and requirements of local and global observability. A summary of research in this field is provided in Table 2.4.

To handle the nonlinear models used in our work, we adapted a matrix projection strategy recently published by van der Heijden (1994a) for linear systems. The algorithm is applicable to the process networks addressed by previous authors as well as the problem of reconciling the large set of measurements and nonlinear relations obtained from a typical bioreactor.
2.4.1 Definitions

The variable classification hierarchy is shown in Figure 2-4. When a measurement is lost or suspected faulty, the variables must be reclassified based on the new configuration. To anticipate these reclassifications beforehand, the algorithm identifies the subsets of the redundant and observable sets that retain their status in these situations.

*Redundant* measurements correspond to variables that remain observable when unmeasured. The remaining measurements are classified as *just measured*. *Tenuously redundant* variables lose their redundancy when another redundant measurement is removed.

The *Tenuously redundant* measurements are identified by sequentially deleting each measurement and determining if any other redundant variables lose their redundancy. If this occurs, this measurement, as well as the reclassified measurements, are labeled as *tenuously redundant*. More importantly, these measurements are identified as a set among which a gross error is indistinguishable by serial elimination. The measurements that retain their redundancy in the face of any particular measurement loss are classified as *robustly redundant*.

*Observable* variables cannot be changed without a change in the values of measured variables. The remaining unmeasured variables are classified as *unobservable*. A subset of the observable variables can be identified that remain observable with the loss of any given measurements and are consequently classified as *robustly observable*. *Robustly observable* and *tenuously observable* variables are identified by sequentially deleting each measurement to see if the observability of the variable is lost.

2.4.2 Classification Method

The observability and redundancy status of a measurement-model system can be seen as falling into one of nine scenarios illustrated by the $3 \times 3$ matrix in Table 2.5. Either none, some, or all measured variables can be redundant and similarly, none, some, or all calculated variables may be observable. Recognizing when a system falls in one of the four corners of the matrix is clear from the equations provided in the table.
Figure 2-4: Variable observability and redundancy classification hierarchy.
Table 2.5: Classification scenarios for variable observability and redundancy for a process system.

<table>
<thead>
<tr>
<th></th>
<th>Redundant Measurements</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>None</td>
<td>Some</td>
<td>All</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>k = rank(R)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e = number of independent equations</td>
<td>Criteria</td>
<td>k=0</td>
<td>0 &lt; k &lt; m</td>
<td>k=m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = number of measurements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c = number of estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observable Estimates None</td>
<td>k=e</td>
<td>e=0</td>
<td>0 &lt; e &lt; m</td>
<td>e=m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observable Estimates Some</td>
<td>e - c &lt; k &lt; e</td>
<td>0 &lt; e &lt; c</td>
<td>m &lt; e &lt; m + c</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observable Estimates All</td>
<td>e = c</td>
<td>c &lt; e &lt; m + c</td>
<td>e = m + c</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The other 5 scenarios require further computations to determine the status of individual variables. The central task is to analyze the structure of the independent equations contained in the process description to determine which calculate unmeasured variables and which establish measurement redundancy. Figure 2-5 provides the algorithm for this determination developed by van der Heijden (1994a). We obtain the tenuous and redundant sub-classifications by iteratively applying this algorithm with particular measurements removed.

2.5 Detecting and Identifying Gross Errors

The validity of the results of the estimation algorithm rests on the assumption that data inconsistency is caused by measurement noise and assumption uncertainty. The presence of undetected gross errors invalidates this assumption and introduces bias into the resulting estimates. Typical gross errors (faults) include miscalibrated sensors, failed sensors, process leaks, unknown cellular byproducts, and neglected dynamic episodes.

There are two steps in gross error handling. The first is to detect the presence of a problem, the second is localizing its source to a particular measurement or model. If we incorrectly reject the null hypothesis ($H_0$) that inconsistency is caused by measurement noise, then a Type I error (false alarm) occurs. Conversely, rejecting a valid $H_0$ results in a Type II error (undetected fault).

The global test (Almasy & Sztano, 1975) compares the measurement adjustments
Figure 2-5: van der Heijden's algorithm adapted for nonlinear observability and redundancy classification.
against a $\chi^2$ distribution threshold to decide if inconsistency results from random error:

$$ h = \delta^T \Psi^{-1} \delta = \epsilon^T \Phi^{-1} \epsilon $$

(2.34)

The null hypothesis $H_0$ is rejected at confidence level $1-\alpha$ if the $h$ statistic exceeds a threshold value for $m$ degrees of freedom:

$$ h(\hat{x}) < \chi^2_{(1-\alpha)}(m) $$

(2.35)

$$ P(h(\hat{x}) \geq \chi^2_{(1-\alpha)}(m)) = 1 - \alpha $$

(2.36)

The number of degrees of freedom is equal to the number of independent constraint equations. The effects of known biases can be added into $\Psi$. The p-value ($\alpha$) represents the probability of incorrectly rejecting a valid null-hypothesis. The converse is a incorrectly accepted null hypothesis which is referred to as a Type II error. The p-value can be plotted graphically over time to monitor the consistency of the process data; ideally, introduction of a fault will result in a sudden increase in the p-value. Figure 2-6A shows the cumulative probability distribution for 1, 3, and 5 degrees of freedom of the $h$ statistic. After experimenting with various visualization techniques, we chose to represent inconsistency by plotting the $\alpha^{-1}$ value that maps to the observed value of $h$ as shown in Figure 2-6B.

Knepper and Gorman (1980) conserve computational effort by testing for gross errors after the first iteration in nonlinear reconciliation.

$$ (f(\hat{x}_m, \hat{x}_c^1))^T \Phi (f(\hat{x}_m, \hat{x}_c^1)) \sim \chi^2_{e-c} $$

(2.37)

$$ f(\hat{x}_m, \hat{x}_c^1) \sim N(0, I) $$

(2.38)

where $\hat{x}_c^{-1}$ is the $\hat{x}$ after the first iteration. One disadvantage of the global test is that it analyzes the entire flow sheet and is therefore less sensitive to particular nodes. It also requires a separate gross error identification step.

Other common inconsistency metrics include the Nodal Test (Mah et al., 1976), the Measurement Test (Mah & Tamhane, 1982) and the Generalized Likelihood Ratio test.
Figure 2-6: The cumulative probability density function for $h$ for 1, 3, and 5 degrees of freedom plotted as $(1-\alpha)$ (A) and $\alpha^{-1}$ (B).
Table 2.6: Progression of gross error detection and diagnosis research.

<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1963</td>
<td>Reilly and Carpani</td>
<td>Global and nodal test</td>
</tr>
<tr>
<td>1965</td>
<td>Ripps</td>
<td>Serial elimination</td>
</tr>
<tr>
<td>1972</td>
<td>Nogita</td>
<td>Serial elimination</td>
</tr>
<tr>
<td>1975</td>
<td>Almasy and Sztano</td>
<td>Global test</td>
</tr>
<tr>
<td>1976</td>
<td>Mah et al.</td>
<td>Nodal test</td>
</tr>
<tr>
<td>1977</td>
<td>Madron</td>
<td>Multidimensional $\chi^2$ test</td>
</tr>
<tr>
<td>1980</td>
<td>Knepper and Gorman</td>
<td>Nonlinear global test</td>
</tr>
<tr>
<td>1981</td>
<td>Romagnoli and Stephanopoulouos</td>
<td>Recursive serial elimination</td>
</tr>
<tr>
<td>1982</td>
<td>Mah and Tamhane</td>
<td>Measurement test</td>
</tr>
<tr>
<td>1985</td>
<td>Madron</td>
<td>Measurement credibility based on bounds violation</td>
</tr>
<tr>
<td>1986</td>
<td>Serth and Heenan</td>
<td>Modified iterative measurement test (MIMT)</td>
</tr>
<tr>
<td>1986</td>
<td>Serth et al.</td>
<td>MIMT applied for nonlinear systems</td>
</tr>
<tr>
<td>1987</td>
<td>Narasimhan and Mah</td>
<td>Generalized likelihood ratio method</td>
</tr>
<tr>
<td>1992</td>
<td>Rollins and Roelfs</td>
<td>Statistical test for bilinear constraints</td>
</tr>
<tr>
<td>1993</td>
<td>Rollins and Davis</td>
<td>Calculation of $\Psi$ from data and related statistical tests</td>
</tr>
<tr>
<td>1994</td>
<td>Crowe</td>
<td>DR and GED using principal component analysis</td>
</tr>
</tbody>
</table>

(Narasimhan & Mah, 1987). The Nodal Test evaluates the constraint residuals to decide which equations might contain a gross error. The measurement test looks at the adjustment of each measurement to identify and rank measurements that may be at fault. The Generalized Likelihood Ratio test is a more general formulation of the measurement test that also considers leaks.

While the Nodal Test avoids the computational cost reconciliation step, it does require further processing to identify the actual measurement error. The drawback of the measurement test is that it assumes that the erroneous measurement will have the largest normalized adjustment. In our experience, the measurement adjustment often smears the impact of gross errors across several measurements or to one particular measurement, referred to as a leveraging point (Rawlings, 1988).

Identification of gross errors can be done using serial elimination (Ripps, 1965; Nogita, 1972). Romagoli and Stephanopoulouos (Romagnoli & Stephanopoulouos, 1981) enhanced this approach by applying it recursively to subsets of the process flow sheet.
2.6 Performance Metrics

The performance of the overall data reconciliation methodology can be evaluated in several ways. The effectiveness of the rectification step can be quantified by the reduction in bias and variance of the reconciled measurements. Gross error detection can be evaluated by how frequently Type I and II errors are made and by the magnitude of an error necessary for detection. This is referred to as the power of the test. Gross error identification can be evaluated by the selectivity in which it finds the correct source of the problem compared with incorrect assignments. Further discussion of performance evaluation can be found in the works of Iordache et al., (1985) and Rosenberg et al., (1987).

2.7 Degrees of Freedom in a Bioprocess

Figure 2-7 breaks down a bioreactor into subunits representing the head space, bubbles, liquid, and cells. The 17 units depicted contain 1 to 19 components and are connected by 43 streams. The components and reactions associated with each unit are listed in Table 2.7 while the resulting sets of equations and unknowns are listed in Table 2.8. For the sake of simplicity we use three lumped reactions to represent growth, respiration, and product formation. Minor nutrients and intracellular accumulations are neglected.

The resulting model has 353 variables constrained by 108 equations, leaving us to find a means to specify the remaining 245 degrees of freedom. We can use the equations in Section 2.1.1 to calculate the degrees of freedom that remain after application of the network balances. This provides a sense of how many additional measurements, assumptions, and constraints are required to describe the process.

Duhem’s Theorem (Prigogine & Defay, 1954) can be used to find the number of equations provided by particular unit operations. Table 2.9 summarizes the results of applying it to several common unit operations (Westerburg et al., 1979). Each unit builds upon $C+2$ equations provided by the fundamental balances using the equations and associated parameters based on the units characteristics. These descriptions can be applied to the unit operations around the reactor.
Figure 2-7: Network representation of a typical bioprocess flow sheet.
Table 2.7: Description of blocks in bioprocess example

<table>
<thead>
<tr>
<th>Unit</th>
<th>Components</th>
<th>$C^u_k$</th>
<th>$R_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substrate tank</td>
<td>H$_2$O, H$_3$O$^+$, OH$^-$, CH$_2$O</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>PumpS</td>
<td>&quot;</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Acid tank</td>
<td>H$_2$O, HCl, Cl$^-$, H$_3$O$^+$, OH$^-$</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>PumpA</td>
<td>&quot;</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Base tank</td>
<td>H$_2$O, NH$_3$, NH$_4^+$, NH$_4$OH, H$_3$O$^+$, OH$^-$</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>PumpB</td>
<td>&quot;</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Inoculum tank</td>
<td>H$<em>2$O, CH$</em>{1.8}$O$<em>{0.5}$N$</em>{0.2}$, H$_3$O$^+$, OH$^-$, CH$_2$O</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Mixer</td>
<td>H$_2$O, O$_2$, CO$_2$, Ar, N$_2$</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Filter1</td>
<td>&quot;</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Filter2</td>
<td>H$_2$O, O$_2$, CO$_2$, Ar, N$_2$, NH$_3$</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Headspace</td>
<td>&quot;</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Bubbles</td>
<td>&quot;</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Condenser</td>
<td>&quot;</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Cooling System</td>
<td>H$_2$O</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Heat exchanger</td>
<td>H$_2$O</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cells</td>
<td>CH$<em>{1.8}$O$</em>{0.5}$N$_{0.2}$</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Liquid</td>
<td>O$_2$, CO$_2$, Ar, N$_2$, CO$_2$</td>
<td>19</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>H$_2$O, NH$_3$, HCl, HPr, CH$_2$O, NH$_4$OH, H$_2$CO$_3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$OH^-$, Cl$^-$, Pr$^-$, HCO$_3^-$, CO$_3^-$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>H$_3$O$^+$, NH$_4^+$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>$\sum_k C^u_k$</td>
<td>91</td>
</tr>
<tr>
<td>System</td>
<td></td>
<td>$\sum_k R_k$</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 2.8: Summary of bioprocess balances and variables used in bioprocess example

<table>
<thead>
<tr>
<th>Balances type</th>
<th>symbol</th>
<th>number</th>
<th>Unknowns type</th>
</tr>
</thead>
<tbody>
<tr>
<td>energy</td>
<td>K</td>
<td>17</td>
<td>enthalpy flows</td>
</tr>
<tr>
<td>energy</td>
<td>$\sum_k C^u_k$</td>
<td>91</td>
<td>enthalpy accumulations</td>
</tr>
<tr>
<td>component</td>
<td>$\sum_k C^u_k$</td>
<td>91</td>
<td>component flows</td>
</tr>
<tr>
<td>component</td>
<td>$\sum_k C^u_k$</td>
<td>91</td>
<td>component accumulations</td>
</tr>
<tr>
<td>elemental</td>
<td>$\sum_k E_k$</td>
<td>0</td>
<td>reaction extents</td>
</tr>
<tr>
<td>elemental</td>
<td>$\sum_k C^u_k$</td>
<td>0</td>
<td>consumption rates</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>108</td>
<td></td>
</tr>
</tbody>
</table>

52
<table>
<thead>
<tr>
<th>Operation</th>
<th>Basic</th>
<th>New</th>
<th>Total</th>
<th>Number</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixer</td>
<td>C+2</td>
<td>-</td>
<td>C+2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flash</td>
<td>C+2</td>
<td>C+2</td>
<td>2C+4</td>
<td>1</td>
<td>Q</td>
</tr>
<tr>
<td>Heat Exchanger</td>
<td>C+2</td>
<td>C+2</td>
<td>2C+4</td>
<td>1</td>
<td>Q</td>
</tr>
<tr>
<td>Pump/Valve</td>
<td>C+2</td>
<td>-</td>
<td>C+2</td>
<td>1</td>
<td>$\Delta p$</td>
</tr>
<tr>
<td>Splitter</td>
<td>C+2</td>
<td>(s-1)(C+2)</td>
<td>s(C+2)</td>
<td>s-1</td>
<td>splits</td>
</tr>
<tr>
<td>Reactor(extent)</td>
<td>C+2</td>
<td>-</td>
<td>C+2</td>
<td>r+2</td>
<td>$\xi_k, Q, \Delta p$</td>
</tr>
</tbody>
</table>

Table 2.9: Balances provided by common unit operations
Figure 2-8: Strategy to manage assumptions used in the reconciliation model to increase the chances that only one gross error is detected at a time.

A similar approach can be used to characterize the reaction and vapor-liquid equilibria present in the reactor. Further constraints can be obtained based on mass transfer, cell physiology (e.g., growth rates, yields), and equipment correlations. At this point, models of the internal structure and behavior of the particular units and components (e.g., reaction and vapor-liquid equilibrium) must be included to reduce the degrees of freedom. Our research has focused on experimentally exploring this hierarchy to identify redundancies reliable enough for use in reconciliation.
Chapter 3

BIOPROCESS GAS MODELING

Gas flow rates, exchange rates, and concentrations contribute to over half of the on-line variables in modern bioprocesses and consequently play an important role in data reconciliation. The detailed description of these variables here provides the basis for the experimental analyses in Chapter 5. This chapter is broken into five sections with the first providing an overview of the off-gas-related variables. The next three sections describe dissolved gas concentrations, gas exchange rates, and derived quantities respectively. The relationships among the variables, which contribute simultaneously and implicitly in reconciliation, are described here individually for use in measurement validation and replacement. The final section describes a strategy to improve the accuracy of off-gas-derived calculations, specifically that of the nitrogen uptake rate and the respiratory quotient.

3.1 Overview of Gas Variables

Figure 3-1 examines the bioreactor from the perspective of the reactor, gas-liquid interface, and cell-liquid interface. From the reactor perspective, compressed gas enters near the base of the reactor, passes through the agitated liquid and head-space above, and leaves via an exit line. The volumetric flow rate ($F_{in}$) is measured for the inlet while the component mole fractions are measured for both the inlet ($y_{i}^{in}$) and outlet ($y_{i}^{out}$) streams. Measurements typically include the reactor temperature ($T$), pressure ($p_{tot}$), pH, agitation rate ($N$), and dissolved oxygen level ($DO$). The broth optical density ($OD$) and volume ($V_{L}$) can either be estimated, measured off-line, or calculated indirectly.

Focusing on the perspective of the gas-liquid interface, there are the partial pressures in the liquid ($p_{i}^{L}$) and gas ($p_{i}^{G}$) phases and the flow of oxygen ($OTR$) and
carbon dioxide (CER) between the phases. The transfer rates are related to the partial pressures by an overall volumetric mass transfer coefficient ($k_La$).

Finally, from the perspective of cell metabolism, we are interested in the rates of gas uptake and production by the cells. Note the calculations in this chapter include water vapor in the summation of gas mole fractions and pressures. Analysis of data reported on a dry-basis (as is the case with some analyzers and spectrometers) requires a corresponding adjustment of the equations. The dimensions used for pressure (atm) and volume (L) are uniform while those for time (h or min) and material quantity (g, mol, or mmol) vary to conform to the established nomenclature of the field. Appendix A.1 provides a complete listing of symbols and their dimensions.
3.2 Dissolved Gases

Concentrations of gases in the liquid phase affect metabolism, mass transfer, and material balances. In aerobic bacterial processes, the selection of the dissolved O\textsubscript{2} set-point is driven by two competing demands. On one hand, a low \( p_{O_2}^L \) level can cause the cells to produce less desired product and more undesired byproduct(s); on the other hand, maintaining a high \( p_{O_2}^L \) level reduces the driving force and therefore the maximum O\textsubscript{2} transfer rate. Failure to balance this trade-off can reduce process performance and product quality. This situation leads many manufacturers to install a second dissolved oxygen probe to detect failures; however, even when this extra measurement is taken, it can not identify which of the two probes is faulty.

Unlike dissolved oxygen, which can be readily measured, the concentration of dissolved CO\textsubscript{2} typically must be calculated indirectly. The calculated value must reflect physically dissolved CO\textsubscript{2} as well as chemically dissolved CO\textsubscript{2} present in species such as bicarbonate. As we will discuss in Section 3.5, the pH dependence of the physical-chemical partitioning of CO\textsubscript{2} impacts the calculation of the RQ.

The dissolved gas partial pressure (\( p_i^L \)) is related to the dissolved gas concentration (\( C_i \)) by its Henry's law coefficient:

\[
p_i^L \approx H_i C_i \tag{3.1}
\]

Assuming a well mixed gas phase, the gas-phase partial pressure is given by

\[
p_i^G = y_i^{\text{out}} p_{\text{tot}} \tag{3.2}
\]

For oxygen, dissolved oxygen probes provide a direct measurement of the \( p_{O_2}^L \). The quantitative relationship between DO and \( p_{O_2}^L \) depends on calibration conditions, specifically the mole fraction of O\textsubscript{2} in the exit gas (\( y_{O_2}^{\text{cal}} \)), absolute reactor pressure (\( p_{\text{tot}}^{\text{cal}} \)), and the dissolved O\textsubscript{2} probe signal (DO\textsuperscript{cal}):

\[
p_{O_2}^L = \left( \frac{\text{DO}}{\text{DO}^{\text{cal}}} \right) y_{O_2}^{\text{cal}} p_{\text{tot}}^{\text{cal}} \tag{3.3}
\]
For other dissolved gases such as N₂ and CO₂, we must estimate this quantity using Henry’s law coefficients and mass transfer correlations relating the gas transfer rate \((TR_i)\) to the pressure driving force:

\[
p_i^L = \frac{p_i^G}{k_La_i} - \frac{TR_i}{k_La_i} \tag{3.4}
\]

### 3.2.1 Dissolved CO₂

Total dissolved CO₂ \(C_{CO₂}^{tot}\) includes the physically dissolved portion \(C_{CO₂}\) as well as the amount chemically partitioned into bicarbonate, carbonate, and carbonic acid:

\[
C_{CO₂}^{tot} \approx C_{CO₂}(1 + K_1 10^{pH}) \tag{3.5}
\]

The concentration of physically dissolved CO₂ is proportional to its liquid phase partial pressure \(p_{CO₂}^L\). The effect of mass transfer resistance, which raises \(p_{CO₂}^L\) above the level that would be in equilibrium with the gas phase, can be represented by combining and rearranging Equations 3.1, 3.2, and 3.4, with the equation (defined later) for calculating the transfer rate (Eq. 3.18) to obtain

\[
C_{CO₂} = \frac{y_{CO₂}^{out}p_{CO₂}^{tot}}{H_{CO₂}} \left(1 + \frac{q_{in}}{p_{CO₂}^{tot}k_La_{CO₂}}\right) \tag{3.6}
\]

\[
C_{CO₂}^{tot} \approx C_{CO₂}(1 + K_1 10^{pH}) = \frac{y_{CO₂}^{out}p_{CO₂}^{tot}}{H_{CO₂}} \left(1 + \frac{q_{in}}{k_La_{CO₂}p_{CO₂}^{tot}}\right) (1 + K_1 10^{pH}) \tag{3.7}
\]

We neglect carbonic acid and carbonate because the concentration of the former is approximately 600 times less than that of CO₂ and the concentration of the latter is insignificant below pH 9. Table 3.1 lists the equilibrium relationships among the various species.

Kinetic limitations in the chemical partitioning of CO₂ are another area of consideration. The rate of carbonic acid dissociation is a potential bottleneck following sudden reductions in CO₂ pressure as might occur with reduced gas residence time,
Table 3.1: Henry’s law constants (atm L mmol⁻¹) from Royce and Thornhill (1991) and equilibrium constants for dissolved gases

\[
\begin{align*}
H_{O_2} &= \frac{1}{1.01 \times 10^5} \exp \left( 12.74 - \frac{133.4}{T(K)-206.7} \right) \\
H_{CO_2} &= \frac{1}{1.01 \times 10^5} \exp \left( 11.25 - \frac{395}{T(K)-175.9} \right) \\
K_1 &= \frac{[HCO_3^-][H^+]}{[CO_2 + H_2CO_3]} = 10^{-6.3} \text{ mol} \frac{mol}{L} \\
K_2 &= \frac{[H^+][CO_3^{2-}]}{[HCO_3^-]} = 10^{-10.3} \text{ mol} \frac{mol}{L}
\end{align*}
\]
pressure, or growth rate:

\[
CO_2(g) \xrightleftharpoons[k_{L'A}^{CO_2}]{}^{k_{L'A}^{CO_2}} CO_2(aq) + H_2O \xrightarrow[H_2O]{}^{0.3s^{-1}} H_2CO_3 \xrightarrow[HCO_3^-]{}^{1s^{-1}} HCO_3^- + H^+
\]

The relative significance of this effect and the effects of mass transfer, chemical dissociation, and reaction kinetics can be assessed by solving the dissolved CO\textsubscript{2} material balance for bicarbonate and substituting it into Equation 3.7 to yield:

\[
C_{CO_2}^{tot} = \frac{y_{CO_2}^{out} p_{tot}}{H_{CO_2}} \left\{ 1 + \frac{q_{in}}{k_{L'A}^{CO_2} p_{tot}} \left[ 1 + \frac{1}{K_{1}10^{pH}} \left( \frac{dC_{CO_2}}{dt} + q_{in} y_{CO_2}^{out} \right) \right] \right\}
\]

This equation summarizes the variable proportionality between \( y_{CO_2}^{out} \) and \( C_{CO_2}^{tot} \). As we will discuss in Section 3.5, this variability manifests itself as noise in the transfer quotient (TQ) and systematic differences between TQ and the respiratory quotient (RQ).

3.2.2 Dissolved CO\textsubscript{2} Dynamics

The larger solubility of CO\textsubscript{2} can lead to significant dynamic differences between the CPR and the CER that are sustained over the entire course of a batch. We refer to this difference as the CO\textsubscript{2} accumulation rate (CAR):

\[
CAR = \frac{dC_{CO_2}^{tot}}{dt}
\]

As the cell density increases, the \( p_{CO_2}^{L} \) increases to maintain equilibrium against the increasing \( p_{CO_2}^{G} \) in the exit gas. This results in a significant fraction of the CO\textsubscript{2} remaining in the liquid phase. This fraction decreases with acidity and aeration. RQ calculations that ignore this accumulation in batch culture or the analogous loss in continuous culture product streams can be significantly biased (Aiba & Furuse, 1990). A
dynamic balance on the amount of CO₂ in the liquid phase can be represented as a set of coupled first-order differential equations:

\[
\frac{dC_x}{dt} = \mu C_x
\]

(3.10)

\[
\frac{dC_{CO₂}}{dt} = \frac{\mu C_x}{Y_{XCO₂}} - \frac{q_{in}H_{CO₂}C_{CO₂}}{p_{tot}} \quad \text{(3.11)}
\]

If \( \mu \) is constant, these equations can be solved using Laplace transforms to yield:

\[
C_x = C_{x0}e^{\mu t} \quad \text{(3.12)}
\]

\[
C_c = \frac{C_{x0}e^{\mu t}}{1 + \frac{Y_{XCO₂}q_{in}H_{CO₂}}{p_{tot}(1 + K_110^{pH})}} + Ae^{-\left(\frac{q_{in}H_{CO₂}}{p_{tot}}\right)t} \quad \text{(3.13)}
\]

The magnitude of the second term in Equation 3.13 decreases exponentially over time and represents the difference between the initial dissolved CO₂ concentration and the amount that would be in equilibrium for the operating conditions.

\[
C_c = \left(1 + \frac{Y_{XCO₂}q_{in}H_{CO₂}}{p_{tot}(1 + K_110^{pH})}\right)^{-1} \quad \text{(3.14)}
\]

We draw on these equations later to predict the difference between the TQ and RQ as a function of operating conditions.

### 3.3 Gas Exchange Rates

One important aspect of bioprocess monitoring is calculating the rates of cell-liquid gas exchange. Dynamic changes in concentrations in the liquid, bubble, and head space, can cause the exchange rate to differ from the corresponding transfer rate associated with it.

All gas exchange rate calculations depend on the specific molar flow of gas \( q_{in} \) (mmol L⁻¹h⁻¹). The conversion factor from volumetric flow \( F_{in} \) (L min⁻¹) to molar flow depends on the liquid volume \( V_L \) and the temperature, \( T \) (K), and absolute
pressure \( (p_{\text{tot}}) \) of the gas:

\[
q_{\text{in}} = 6 \times 10^4 \left( \frac{p_{\text{tot}} F_{\text{in}}}{V_L (RT)} \right) \quad (3.15)
\]

The symbol \( R \) represents the ideal gas constant \( (8.206 \times 10^{-5} \text{ L atm mmol}^{-1} \text{ K}^{-1}) \).

Most mass flow meters report data at engineering standard conditions \( (25^\circ \text{C}, 1 \text{ atm}) \) regardless of the state of the gas stream. This leads to a molar concentration of 24.45 L mol\(^{-1}\) and reduces Equation 3.15 to

\[
q_{\text{in}} = 2453 \frac{F_{\text{in}}}{V_L} \quad (3.16)
\]

When volumetric flow meters such as rotameters are used, the molar rate must be calculated using the actual temperature and pressure of the gas. Inadvertently applying a conversion based on scientific standard conditions \( (22.4 \text{ L mol}^{-1}) \) leads to a systematic error of 10 percent.

The exit gas flow, which is difficult to measure directly, is calculated based on constant molar flow of an inert gas. Usually \( N_2 \) is used although Argon can be substituted if necessary. The off-gas composition measurements of the inert and reacting gases allow us to calculate \( \Delta_i \), the normalized mole fraction change for component \( i \) across the reactor:

\[
\Delta_i = y_i^{\text{in}} - y_i^{\text{out}} \left( \frac{y_{\text{inert}}^{\text{in}}}{y_{\text{inert}}^{\text{out}}} \right) \quad (3.17)
\]

Combining this value with the gas flow rate closes the gas balance for component \( i \) around the reactor and calculates the transfer rate of component \( i \) \( (TR_i) \). The transfer rate is taken as negative by convention for evolved gases in this equation:

\[
TR_i = q_{\text{in}} \Delta_i \quad (3.18)
\]

The difference between the transfer rate and \( \text{dissolved} \) uptake rate is the change in dissolved
gas concentration, $C_i$ (mmol L$^{-1}$):

$$UR_i = TR_i + \frac{dC_i}{dt} (10^3)$$

(3.19)

The transfer rates for $O_2$, $CO_2$, and $N_2$ are referred to as oxygen transfer rate (OTR), $CO_2$ evolution rate (CER), and $N_2$ transfer rate (NTR) respectively. The corresponding uptake and production rates are referred to as the $O_2$ uptake rate (OUR), $CO_2$ production rate (CPR), and the $N_2$ uptake rate (NUR).

Because oxygen is only slightly soluble in water, the difference between the rate of oxygen uptake (OUR) and oxygen transfer (OTR) can usually be neglected in rate calculations.

### 3.4 Overall Volumetric Mass Transfer Coefficient

The overall volumetric mass transfer coefficient $k_La_c$ (h$^{-1}$) relates the rate of oxygen transfer to the difference between the concentration of $O_2$ in the liquid phase and the concentration of oxygen required for equilibrium. These concentrations are referred to as $C_{O_2}$ and $C_{O_2}^*$ respectively. The $C_{O_2}^*$ value varies with the reactor pressure and the amount of $O_2$ stripped by the cells. Although the driving force is often represented using molar concentrations, the poorly defined dependence of the Henry’s law constant, $H$ (atm L mmol$^{-1}$), on temperature (Table 3.1) as well as medium composition (Popovic et al., 1979; Quicker et al., 1981) makes it more convenient to work with $k_La$, which is defined in units of mmol L$^{-1}$ h$^{-1}$ atm$^{-1}$:

$$k_La^{O_2} = \frac{k_La^{CO_2}}{H_{O_2}} = \frac{OTR}{\frac{H_{O_2}}{V_L} \int_{V_L} \left(C_{O_2}^* - C_{O_2}\right) dV} = \frac{OTR}{\frac{1}{V_L} \int_{V_L} (p_{O_2}^G - p_{O_2}^L) dV}$$

(3.20)

For small reactors, one can calculate the gas-phase partial-pressure ($p_{O_2}^G$) by assuming a well mixed bubble-phase with an $O_2$ mole fraction equal (in composition) to the exit gas stream ($y_{O_2}^{out}$):

$$k_La^{O_2} = \frac{OTR}{p_{O_2}^G - p_{O_2}} = \frac{OTR}{y_{O_2}^{out} p_{tot} - \frac{DO}{DO_{cal}} y_{O_2}^{cal} p_{cal}^{tot}}$$

(3.21)
Most literature on mass transfer in bioprocesses focuses on the \textit{a priori} prediction of \( k_{La} \) using design parameters such as agitation power and impeller geometry for use in process design and scale-up. Here we are interested in predicting \( k_{La} \) while monitoring data from existing facilities. For an agitated reactor with a Newtonian broth, we combine the correlation for \( k_{La} \) of Richards (1961) with the correlations for agitation power from Michel and Miller (1962) and Rushton (1950) to obtain a relation of the form

\[
k_{La} = (G_1 + G_2 N_i) N^{2.16} q_{in}^{0.56}
\]  

(3.22)

The key variables are the agitation rate, \( N \) (rev min\(^{-1}\)), the number of submerged impellers \( (N_i) \), and the aeration rate \( (q_{in}) \). The lumped parameters \( G_1 \) and \( G_2 \) reflect a number of empirical, process and equipment specific parameters. In situations where the number of submerged impellers is constant, the leading terms lump into the fitted constant \( A \). The exponents on agitation and aeration and the \textit{a priori} prediction of the leading coefficient provide a starting point for fitting the parameters of a log-linear empirical model to historical process data:

\[
\ln (k_{La}) = \ln(A) + B \ln(N) + C \ln(q_{in})
\]  

(3.23)

The effects of agitation and aeration, whose rates are often increased simultaneously to meet oxygen demand, can be difficult to separate without conducting experiments specifically designed for that purpose. Lacking designed experiments, we can ignore the dependence on aeration or use a literature-derived exponent (\textit{e.g.}, \( C = 0.56 \) from Eq. 3.22) leaving the constants \( A \) and \( B \) to be calculated using linear regression.

Experimentally derived values for various reactors studied in this work are summarized in Table 3.2. These were obtained with reactor contents that exhibited Newtonian rheology. The correlation can be used in combination with the maximum \( N \) to determine \( k_{La_{max}} \) and, by using other design constraints in Equation 3.31, \( OUR_{max} \).

The mass transfer coefficients for gases other than \( O_2 \) can be estimated from by assuming their values are proportional to the square root of their respective diffusivities.
(Bird et al., 1960):

\[
\frac{k_{L0c}^{gas_1}}{k_{L0c}^{gas_2}} = \frac{k_{L0}^{gas_1} H^{gas_1}}{k_{L0}^{gas_2} H^{gas_2}} = \sqrt{\frac{D_{AB}^{gas_1}}{D_{AB}^{gas_2}}} \tag{3.24}
\]

The values relevant for our work are \(D_{AB}^{N_2} (2.4 \times 10^{-5})\), \(D_{AB}^{O_2} (2.8 \times 10^{-5})\), and \(D_{AB}^{CO_2} (2.0 \times 10^{-5})\) cm\(^2\)s\(^{-1}\) (Reid et al., 1977).

### 3.5 The Respiratory and Transfer Quotients

The various phenomena that impact RQ and its estimation are shown in Figure 3-2. This section describes the effects of these stoichiometric and operating parameters and their implications for RQ interpretation and utilization. This is followed by a description of the theoretical mapping relating RQ to substrates, products, and yield. We also describe how to incorporate bicarbonate dynamics into the calculation of RQ and evaluate the propagation of measurement errors. Chapter 5 illustrates these concepts using a laboratory scale *Bacillus subtilis* culture.

The prominence of RQ as a bioprocess measurement dates back to its usefulness in early efforts to control *Saccharomyces cerevisiae* processes (Wang et al., 1977). The RQ has been used in on-line monitoring (Nyiri et al., 1975; Cooney et al., 1977; Wang et al., 1979) of substrate depletion (Salmon & Buckland, 1992) and the yield of biomass (Swartz & Cooney, 1979), secondary metabolites (Heinzle et al., 1990; Kiss & Stephanopoulos, 1991), and undesirable byproducts (Wang et al., 1977). Heinzle and Dunn (1993) provide a comprehensive review of calculations, equipment, and error analysis related to off-gas measurements.

Several researchers have pointed to particular factors that must be considered. Royce and Thornhill (1991) investigated the impact of mass transfer on dissolved CO\(_2\) calculations. Royce (1992) notes that the discrepancy between the transfer quotient (TQ) and the true RQ is increasingly significant when the pH rises above 6.5, when the growth rate is high, and when the pH is uncontrolled. Aiba and Furuse (1990) derive the necessary equations to calculate the TQ to RQ ratio as a function of growth rate and aeration. Recent efforts (Bonarius et al., 1995) address the challenges of monitoring...
Figure 3-2: System overview of factors influencing RQ and its measurement.
RQ in bicarbonate-buffered cell cultures. Researchers also have examined the impact on RQ of measurement (Heinzle et al., 1990) and model (Grosz et al., 1984) sensitivity. Our goal is to organize these factors into a common framework to interpret and use RQ more effectively.

While, by definition, RQ is based on the gas exchange between the cells and their liquid environment, computer systems report the ratio of gases transferred. Royce (1992) defines this quantity, which neglects dissolved gas dynamics, as the TQ:

\[
RQ = \frac{CPR}{OUR} \approx TQ + \frac{CAR}{OUR}  \tag{3.25}
\]

\[
TQ = \frac{CER}{OTR} \approx RQ - \frac{CAR}{OTR}  \tag{3.26}
\]

While neglecting dissolved oxygen dynamics is often acceptable given its low solubility (0.25 mmol L\(^{-1}\) at 0.21 atm), the rate of change in the dissolved CO\(_2\) concentration (CAR) can be significant in some regions of operating space. Carbon dioxide is about 20 times more soluble than O\(_2\). Furthermore, this ratio increases significantly with the pH level due to its increasing chemical partitioning into bicarbonate. At a pH level of 7, for example, the effective solubility of CO\(_2\) rises to 100 times that of O\(_2\).

We can estimate the TQ as a function of operating conditions during exponential growth by drawing on the solution of the dynamic balances (Eq. 3.14) (which assume \(H_{O_2} \gg H_{CO_2}\)) and assuming RQ\(\approx\)1.0:

\[
TQ = \frac{CPR - CAR}{CPR} \approx \left(1 + \left(\frac{\mu}{q_{in}}\right)\frac{p_{tot} (1 + K_1 10^{pH})}{H_{CO_2} Y_{XCO_2}}\right)^{-1}  \tag{3.27}
\]

This relationship, presented in Figure 3.3, is described in detail by Aiba and Furuse (1990) who also describe other parameters, such as head space volume, which are included in the model.

### 3.6 Secondary Estimates

In addition to the direct measurement or calculation of variables, we are interested in defining relationships that serve to establish analytical redundancy for use in data
Figure 3-3: The TQ/RQ ratio under steady growth decreases as the pH level increases and VVM/μ ratio decreases.
reconciliation. Here we describe methods to estimate dissolved oxygen, oxygen uptake, nitrogen uptake, and RQ.

### 3.6.1 Dissolved O₂ Based on OTR and k\textsubscript{La}

The dissolved oxygen level also can be calculated indirectly using off-gas measurements and a mass transfer correlation. The equation defining the overall volumetric mass transfer coefficient for oxygen can be rearranged to estimate the dissolved O₂ concentration as a function of gas composition, gas flow, and k\textsubscript{La}O₂:

\[
DO = \frac{DO^{\text{cal}}}{y^{\text{cal}}_{O_2}P_{tot}} \left( y^{\text{out}}_{O_2}P_{tot} - \frac{OTR}{k_{La}O_2} \right)
\]

(3.28)

This relationship provides an independent opportunity to estimate the dissolved oxygen level. This is useful for discriminating between faults in gas composition and dissolved oxygen measurements. Furthermore, using Equation 3.30 to eliminate y\textsubscript{O₂}\textsuperscript{out} allows us to estimate DO using only an estimate of OTR:

\[
DO = \frac{DO^{\text{cal}}}{y^{\text{cal}}_{O_2}} \left( \frac{P_{tot}}{P_{cal}} \right) \left\{ \frac{y^{\text{in}}_{O_2}}{\delta} - OTR \left( \frac{1}{q_{in}\delta} + \frac{1}{k_{La}O_2P_{tot}} \right) \right\}
\]

(3.29)

### 3.6.2 OTR Based on k\textsubscript{La} and Dissolved O₂

We combine and rearrange the equations defining OUR (Eq. 3.18) and k\textsubscript{La} (Eq. 3.20) to obtain a relationship predicting y\textsubscript{O₂}\textsuperscript{out} as a function of agitation and dissolved oxygen:

\[
y^{\text{out}}_{O_2} = \frac{q_{in}y^{\text{in}}_{O_2} + k_{La}p_{O_2}}{q_{in}\delta + k_{La}P_{tot}}
\]

(3.30)

This allows OTR and OTR\textsubscript{max} to be estimated based on the dissolved O₂ measurement and the operating conditions that impact k\textsubscript{La}:

\[
OTR = q_{in} \left[ y^{\text{in}}_{O_2} - \delta \left( \frac{q_{in}y^{\text{in}}_{O_2} + k_{La}p_{O_2}}{q_{in}\delta + k_{La}P_{tot}} \right) \right]
\]

(3.31)

If the reference for the dissolved O₂ probe calibration is the operating pressure and inlet gas composition and \( \delta \approx 1 \), Equation 3.31 can be simplified and rearranged into
Table 3.2: Fitted coefficients used in Equation 3.23 to predict $k_La$ in various reactors. The MIT 20L reactor has a abnormally high coefficient because it was operated at 9 L instead of its working volume of 15 L.

<table>
<thead>
<tr>
<th>Location</th>
<th>Volume</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>$N_{max}$</th>
<th>$k_La_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>80L</td>
<td>-8.167</td>
<td>2.416</td>
<td>-</td>
<td>550</td>
<td>1187</td>
</tr>
<tr>
<td>ARC</td>
<td>20L</td>
<td>-8.62</td>
<td>2.15</td>
<td>-</td>
<td>1200</td>
<td>752</td>
</tr>
<tr>
<td>ARC</td>
<td>150L</td>
<td>-6.86</td>
<td>1.99</td>
<td>-</td>
<td>800</td>
<td>628</td>
</tr>
<tr>
<td>MIT</td>
<td>20L</td>
<td>-13.3</td>
<td>3.29</td>
<td>-</td>
<td>900</td>
<td>6000</td>
</tr>
<tr>
<td>MIT</td>
<td>75L</td>
<td>-15.6</td>
<td>3.47</td>
<td>-</td>
<td>800</td>
<td>2000</td>
</tr>
<tr>
<td>MIT</td>
<td>1500L</td>
<td>-14.7</td>
<td>2.77</td>
<td>0.56</td>
<td>600</td>
<td>1800</td>
</tr>
</tbody>
</table>

three dimensionless groups that show the general relationship among mass transfer, oxygen stripping, and dissolved oxygen pressure:

$$
\frac{OTR}{q_m y_{O_2}^m} = \frac{1 - \frac{P_{O_2}^L}{y_{O_2}^m P_{tot}}}{1 + \frac{q_m}{k_La O_2^2 P_{tot}}} \quad (3.32)
$$

One interesting point to note, show in Figure 3-4A, is that increases in $k_La$ do not result in proportionate increases in OTR because the fraction of incoming $O_2$ stripped asymptotically approaches $(1 - (P_{O_2}^L / y_{O_2}^m P_{tot}))$ as $k_La$ becomes large. The $(P_{O_2}^L / y_{O_2}^m P_{tot})$ term is often equivalent to percent dissolved oxygen but this is contingent on equivalent operating and calibrating conditions. As shown in the numerical example in Figure 3-4B, constraints on the minimum level of dissolved oxygen and maximum agitation rate define the maximum transfer capacity of the reactor.

Table 3.2 lists experimentally derived coefficients, which, when used in Equation 3.23, characterize mass transfer in various reactors used in this work. Alternatively, heat transfer capacity may be the limiting factor that provides the upper bound on OUR$_{max}$ via Equation 3.36. Understanding this dependence allows us to predict the effectiveness of alternate methods of increasing oxygen transfer capacity (e.g., increased aeration, gas-flow, pressure, enriched $O_2$ or reduced $P_{O_2}^L$). It also facilitates the timing of operating decisions to meet biological needs (e.g., induction for recombinant products) based on reactor constraints.
Figure 3-4: General relation (A) showing the fraction of $O_2$ gas removed as a function of $k_La$, dissolved $O_2$, and reactor pressure and the predicted behavior (B) of a 80L laboratory reactor operating at $p_{tot}=1.7$ with sparging of air ($y_{O_2}^{in}=0.21$) at 60 L min$^{-1}$. 
3.6.3 OUR Based on Growth Rate

If we can estimate the total cell growth rate \( r_X \) from the specific cell growth rate \( \mu \) and the dry cell weight \( X \), we can use an assumed value for cell yield on \( O_2 \) \( (Y_{XO_2}) \) to estimate the corresponding OUR:

\[
OUR = \frac{r_X}{Y_{XO_2}} = \frac{\mu X}{cMW_X Y_{XO_2}} \frac{1000}{Y_{XO_2}}
\]  
(3.33)

The molecular weight of cell mass \( (cMW_X) \) is defined as having one mole of carbon per mole \( (\approx 25.6 \text{ g mol}^{-1}) \). The \( Y_{XO_2} \) can be estimated from the degree of reduction balance described in Section 3.6.6:

\[
Y_{XO_2} = \frac{4}{Y_{XS} - \gamma_X} \]  
(3.34)

\[
OUR = \frac{1000 \mu X}{4 cMW_X} \left( \frac{\gamma_S}{Y_{XS} - \gamma_X} \right)
\]  
(3.35)

3.6.4 OUR Based on Heat Evolution

The heat produced by cell metabolism, \( Q_{\text{heat}} \) \( (\text{kcal L}^{-1}) \), is roughly proportional to oxygen uptake \( (\text{Cooney et al., 1968}) \):

\[
OUR = 0.12 Q_{\text{heat}}
\]  
(3.36)

The resulting estimate, while less accurate than that derived from the gas composition measurement, provides additional analytical redundancy for use in data reconciliation.

3.6.5 NUR Based on Growth Rate

To decide if the calculation of NUR from off-gas measurements is feasible, we estimate the NUR required to sustain cell growth:

\[
NUR = \left( \frac{\mu X}{Y_{XN}} \right) \left( \frac{1000}{MW_{N_2}} \right) + \frac{q_p X}{2} + \frac{dC_{N_2}}{dt}
\]  
(3.37)
where $\mu$ is the specific growth rate ($h^{-1}$), $X$ is the cell density (g dry cells L$^{-1}$), $Y_{XN}$ is the yield of cells on nitrogen (mol mol$^{-1}$), MW$_{N_2}$ is the molecular weight of N$_2$, and $q_p$ is the net specific ammonia production rate (mmol g$^{-1}$ h$^{-1}$). Except for mutants lacking feedback regulation (Bali et al., 1992), $q_p$ is negligible since excess ammonia production represses the enzyme. Also, the dissolved N$_2$ concentration can be assumed to remain constant. Figure 3-5 shows the expected NUR trend for an *Azotobacter vinelandii* culture. This profile, which has a maximum NUR of approximately 1.6 mmol L$^{-1}$ h$^{-1}$, assumes the cells have a 10 percent nitrogen content by weight, a specific growth rate of 0.25 h$^{-1}$, and a maximum cell density of 1.5-2.0 g L$^{-1}$. The corresponding estimate of specific nitrogen uptake (0.9 mmol g$^{-1}$ h$^{-1}$) is constant and depends only on $\mu$ and $Y_{XN}''$. 

Figure 3-5: Expected NUR trend for a typical *Azotobacter vinelandii* run based on the amount of nitrogen required for cell growth.
3.6.6 RQ Based on Stoichiometry

A simplified bioprocess representation consists of a lumped chemical reaction (Cooney et al., 1977) where carbon, nitrogen, and energy sources react to form biomass, CO$_2$, and water. Here RQ is the ratio of $r_{CO_2}$ to $r_{O_2}$:

$$r_{SCH_4}O_{b,N}N_{c,N} + r_{NC}C_{d,N}H_{a,N}O_{b,N}N_{c,N} + r_{O_2}O_2 \rightarrow$$

$$r_XCH_xO_{b,x}N_{c,x} + r_{CO_2}CO_2 + r_{H_2O}H_2O + r_{P}C_{d,P}H_{a,P}O_{b,P}N_{c,P}$$

Implicit in this model are balances for carbon, nitrogen and degree of reduction:

$$r_S + r_{Nd,N} = r_X + r_{CO_2} + r_{Pd,P} \quad (3.38)$$

$$r_{Nc,N} + r_{Sc,N} = r_Xc_X + r_{Pc,P} \quad (3.39)$$

$$r_{S}y_S + r_{O_2}y_{O_2} + r_{N}y_N = r_Xy_X + r_{P}y_{P} \quad (3.40)$$

The degree of reduction ($\gamma_i$), while having a chemical significance, is simply the coefficient that results from combining the hydrogen and oxygen balances to eliminate $r_{H_2O}$ as an unknown. Normalizing to one-carbon molecular formulas (e.g., $CH_2O$ for glucose) allows $\gamma_i$ to be calculated as:

$$\gamma_i = 4d_i + a_i - 2b_i + \left(\frac{2b_N - 4d_N - a_N}{c_N}\right)c_i \quad (3.41)$$

which, when ammonia is the nitrogen source, reduces to

$$\gamma_i = 4d_i + a_i - 2b_i - 3c_i \quad (3.42)$$

Table 3.3. uses this equation to calculate $\gamma_i$ values for several common substrates and products. A biomass composition of $CH_{1.8}O_{0.5}N_{0.2}$ is used to calculate $\gamma_X$; most cell lines fall within 5 percent of this value (Roels, 1983).

The observed RQ reflects a linear combination of metabolic pathways that aggregate to lumped reactions for respiration, growth, and product formation. Rearranging the stoichiometric balances (Eqs. 3.38, 3.39, and 3.40) shows the dependence of RQ on cell
Table 3.3: Values for the degree of reduction ($\gamma_i$) and expected RQ for various substrates when converted to CO$_2$ and water through the respiratory pathway($RQ_r$) or utilized efficiently for cell growth on NH$_3$ ($RQ_g$).

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>$\gamma_i$</th>
<th>$RQ_r$</th>
<th>$RQ_g$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetic Acid</td>
<td>$CH_2O$</td>
<td>4.0</td>
<td>1.0</td>
<td>1.06</td>
</tr>
<tr>
<td>Glucose</td>
<td>$CH_2O$</td>
<td>4.0</td>
<td>1.0</td>
<td>1.06</td>
</tr>
<tr>
<td>Ethanol</td>
<td>$CH_{6\frac{5}{6}}O_{\frac{1}{2}}$</td>
<td>6</td>
<td>0.67</td>
<td>0.46</td>
</tr>
<tr>
<td>Glycerol</td>
<td>$CH_{6\frac{3}{5}}O$</td>
<td>4.67</td>
<td>0.86</td>
<td>0.74</td>
</tr>
<tr>
<td>Mannitol</td>
<td>$CH_{14}O$</td>
<td>4.33</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>Methanol</td>
<td>$CH_4O$</td>
<td>6</td>
<td>0.67</td>
<td>0.46</td>
</tr>
</tbody>
</table>
yield \((Y_{XS})\), and the degrees of reduction of the substrate \((\gamma_S)\) and biomass \((\gamma_X)\):

\[
RQ = \frac{Y_{XO_2}}{Y_{XCO_2}} = \frac{4}{\gamma_X + \frac{Y_{PS}}{Y_{XS}} \gamma_p - \frac{1}{Y_{XS}} \gamma_S - \left(\frac{c_X + \frac{Y_{PS}}{Y_{XS}} c_p - \frac{c_S}{\gamma_N} \gamma_N}{c_N}\right)} \frac{d_p}{d_N}
\]

(3.43)

If the nitrogen source is ammonia and product formation is absent, this relation reduces to

\[
RQ = \left(\frac{4}{\gamma_S}\right) \left(\frac{1 - Y_{XS}}{1 - Y_{XS} \left(\frac{\gamma_X}{\gamma_S}\right)}\right)
\]

(3.44)

This relationship, shown in Figure 3-6, also is found in the work of Erickson (1978a) and Wang et al., (1979). Note that Figure 3-6 assumes a \(\gamma_X\) of 4.06, which is typical of Escherichia coli, and an amino acid composition of \(\text{CH}_{0.5}\text{O}_{0.5}\text{N}_{0.29}\). The shapes of the curves for glucose and protein are sensitive to these two values. Also, all significant byproducts must be represented for the relationship to remain valid. For example, unanticipated byproducts often form when substrates contain both carbon and nitrogen \((e.g.,\) amino acid deamination).

Experimentally observed regularities in thermodynamic efficiency (Roels, 1983) provide an upper limit on \(Y_{XS}\):

\[
Y_{XS}^{\text{max}} \approx \begin{cases} 
0.13\gamma_s & : \gamma_s < 4.67 \\
0.6 & : \gamma_s \geq 4.67
\end{cases}
\]

(3.45)

Conversion from molar to mass yield \((Y'_{XS})\) requires estimates of the biomass ash content (typically 7%) and the one-carbon molar weights \((cMW_i)\) of the cells and substrate:

\[
Y_{XS} = Y_{XS}' \frac{cMW_S(1 - f_{ash})}{cMW_X}
\]

(3.46)
Figure 3-6: Dependence of RQ on $Y_{XS}$ for selected substrates.
Figure 3-7: The dependence of RQ on the product yield on glucose ($Y_{PS}$) assuming ammonia is the nitrogen source and there is no cell growth.

Equation 3.44 can be rearranged to calculate $Y_{XS}$ based on RQ:

$$ Y_{XS} = \frac{\gamma_s RQ - 4}{\gamma_X RQ - 4} \quad (3.47) $$

We also can predict RQ as a function of product yield (Figure 3-7). When RQ data is inconsistent with the assumed model, the RQ range for byproducts the model lacks can be evaluated to identify potential candidates for investigation.

### 3.6.7 Cell Yield Based on RQ

As shown in Figure 3-6, which plots Equation 3.47, RQ is relatively insensitive to yield when $\gamma_S$ and $\gamma_X$ are similar or when $Y_{XS}$ is low (Grosz et al., 1984). The
Figure 3-8: Dependence of $\sigma_{Y_{XS}}$ on $\sigma_{RQ}$, $Y_{XS}$, and $\gamma_X$ for cell growth on carbohydrate. The relationship between the uncertainties in $Y_{XS}$ and RQ is given by

$$\sigma_{Y_{XS}} = \left(\frac{\gamma_X - \gamma_S}{4 \left[ (\frac{\gamma_X}{4} RQ - 1)^2 \right]} \right) \sigma_{RQ}$$  

(3.48)

As shown in Figure 3-8, the $\sigma_{Y_{XS}}$ to $\sigma_{RQ}$ ratio ranges from 2 to 65 for carbohydrates depending on $Y_{XS}$ and $\gamma_X$; consequently, a high degree of RQ accuracy is required to estimate $Y_{XS}$.

3.6.8 Heuristics for RQ Interpretation

Our results can be combined with those of Royce and Heinzle to develop a set of general heuristics, shown in Table 3.4, for interpreting RQ both with and without compensation for gas dynamics.
Table 3.4: Heuristics for interpreting RQ calculations and trends

<table>
<thead>
<tr>
<th>Situation</th>
<th>Potential Causes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TQ &lt; RQ_{expected}$</td>
<td>Steady growth</td>
</tr>
<tr>
<td></td>
<td>Growth increasing</td>
</tr>
<tr>
<td></td>
<td>Airflow decreased</td>
</tr>
<tr>
<td></td>
<td>pH increasing</td>
</tr>
<tr>
<td>$TQ &gt; RQ_{expected}$</td>
<td>Growth slowing</td>
</tr>
<tr>
<td></td>
<td>Airflow increased</td>
</tr>
<tr>
<td></td>
<td>pH falling</td>
</tr>
<tr>
<td>$RQ &lt; RQ_{expected}$</td>
<td>pH &gt; measured pH</td>
</tr>
<tr>
<td></td>
<td>$(\gamma_s/\gamma_X) &gt; \text{ expected}$</td>
</tr>
<tr>
<td>$RQ &gt; RQ_{expected}$</td>
<td>pH &lt; measured pH</td>
</tr>
<tr>
<td></td>
<td>$(\gamma_s/\gamma_X) &lt; \text{ expected}$</td>
</tr>
<tr>
<td>Oscillating TQ</td>
<td>pH control variation</td>
</tr>
<tr>
<td>Flat TQ</td>
<td>Mass spectrometer noise</td>
</tr>
<tr>
<td>Poor $Y_{X,S}$ estimate</td>
<td>$\gamma_X \approx \gamma_S$</td>
</tr>
<tr>
<td></td>
<td>Sensor failure</td>
</tr>
</tbody>
</table>
3.7 Optimization of Off-Gas Calculation Accuracy

We took a particular interest in the experimental monitoring of RQ and NUR because both represented interesting metabolic indicators and were technically challenging to calculate from available data. Our strategy employed error propagation techniques (Box et al., 1978) to evaluate the effects of parameter and measurement uncertainty on the calculation of RQ and NUR and then exploited this information to select operating conditions to improve calculation accuracy.

\[
y = f(x_1, x_2, ...)
\]

\[
\sigma_y = \sqrt{\left( \frac{df}{dx_1} \right)^2 \sigma_{x_1}^2 + \left( \frac{df}{dx_2} \right)^2 \sigma_{x_2}^2 + ...}
\]

3.7.1 Optimization of NUR Accuracy

Noise in mass spectrometer measurements often has a significant impact on the accuracy of derived variables (Heinzle et al., 1990; Royce & Thornhill, 1992). Under normal operating conditions, measurement noise obscures the calculated NUR (Section 5.3.1). Described here is our strategy to optimize the NUR accuracy while maintaining the cell environment necessary for enzyme production.

We estimated the uncertainty of the calculated uptake rates (\(\sigma_{UR_i}\)) by propagating the measurement uncertainties in Equation 3.18 using Equation 3.50:

\[
\sigma_{\Delta_i}^2 = \sigma_{y_{i,n}}^2 + \left( \frac{\sigma_{\Delta_{\text{out}}}}{y_{\text{out}_i}} + \frac{\sigma_{\Delta_{\text{Ar}_{\text{in}}}}}{y_{\text{Ar}_{\text{in}}}} + \frac{\sigma_{\Delta_{\text{Ar}_{\text{out}}}}}{y_{\text{Ar}_{\text{out}}}} \right) y_{\text{out}_i}^2 \frac{y_{\text{Ar}_{\text{in}}}}{y_{\text{Ar}_{\text{out}}}} \nu_{\Delta_i}
\]

\[
\sigma_{UR_{i,n}}^2 = \Delta_i^2 \sigma_{q_{i,n}}^2 + q_{i,n}^2 \sigma_{\Delta_i}^2
\]

In this case, the optimization objective is to minimize the uncertainty in the NUR (\(\sigma_{\text{NUR}}\)) subject to several sets of constraints. The first three constraints represent the operating limits of the reactor for agitation, pressure, and gas flow. The second set of constraints limit \(p_{O_2}^L\) and \(p_{K_2}^L\) so that \(p_{O_2}\) is maintained at a constant level to avoid inhibition and damage of the nitrogenase enzyme and \(p_{K_2}\) is maintained above a level at which it might affect cell growth and enzyme activity.

Culture broth aerated with normal air and operated at 0.5 atm gauge pressure
exposes cells to N\textsubscript{2} partial pressures as high as 1.2 atm. The reduced N\textsubscript{2} partial pressure in new inlet gas is constrained to avoid significantly affecting the behavior of the cells. The pressure is set near the 0.79 atm environmental level. In theory, the pressure can be set lower since the Monod constant, \( K_M \), for N\textsubscript{2} is between 0.04 atm (Dalton & Postgate, 1969) and 0.25 atm (Orme-Johnson, 1985):

Although we can vary gas feed composition, maintaining it constant avoids problems resulting from equilibration time, controller noise, and apparatus complexity \((y_{t=0} = y_{t=t_f})\).

We sought operating conditions that would minimize NUR uncertainty at the end of the batch and at the same time satisfy the constraints for the duration of the run. The high level of mass transfer maintained to satisfy the oxygen transfer requirements leads to a \( \frac{NUR}{k_L a N_2} \) term on the order of \( 10^{-3} \). Therefore \( p_{N_2}^L \) remains essentially at saturation level for the duration of the run.

### 3.7.2 Optimization of RQ Accuracy

The equations defining \( C_{CO_2}^{tot} \), RQ, and TQ include a large number of uncertain measurements and parameters. In order to have a systematic way to decide when calculated RQ is consistent with process measurements and assumptions, we must understand the effects of uncertainty in stoichiometry, biomass composition, and bicarbonate accumulation. In this section we describe methods to address these questions.

Propagating the uncertainties in Equation 3.26, we find the accuracy of TQ proportional to the fraction of oxygen stripped:

\[
\sigma_{TQ} \approx \frac{2Q}{OUR} \sigma_y
\]

Plotting this relation (Figure 3-9) shows that reducing aeration to the level required for minimum dissolved O\textsubscript{2} also minimizes TQ uncertainty.

While minimizing aeration improves TQ, in the context of RQ we find the dynamics of bicarbonate accumulation confound this approach. The calculation of RQ includes both the TQ and the dissolved CO\textsubscript{2} accumulation rate (CAR). Propagating the errors
Figure 3-9: The dependence of $\sigma_{TQ}$ on the stripping rate (OUR/VVM) and gas composition accuracy.
in the equation defining RQ (Eq. 3.25), we see the accuracy of the CAR estimate is implicit in the estimate of the RQ accuracy:

\[ \sigma_{RQ}^2 = \sigma_{TQ}^2 + \frac{\sigma_{CAR}^2}{OUR^2} \]  

(3.54)

\[ CAR = \frac{\Delta C_{CO2}^{tot}}{\Delta t} \]  

(3.55)

\[ \sigma_{CAR}^2 = \frac{2}{\Delta t} \sigma_{C_{CO2}}^{tot} \]  

(3.56)

Applying error propagation to the \( C_{CO2}^{tot} \) estimate from Equation 3.5, while neglecting kinetic and mass transfer effects, we find the \( \sigma_{C_{CO2}^{tot}} \) is highly dependent on the magnitude and accuracy of pH:

\[ \sigma_{C_{CO2}^{tot}}^2 = \left( \frac{y_{CO2}^{out} p_{tot} K_1 10^{pH} e}{H_{CO2}} \right)^2 \sigma_{pH}^2 + \left( \frac{p_{tot}}{H_{CO2} (1 + K_1 10^{pH})} \right)^2 \sigma_y^2 \]  

(3.57)

The uncertainty in the gas accumulation rate originates from uncertainty in the liquid and gas concentrations, the key inputs are pH, \( \sigma_{pH} \), and the total gas sparged during the calculation time step \( \Delta t \) (Figure 3-10).

\[ \sigma_{RQ}^2 = 2 \left( \frac{p_{tot} K_1 10^{pH} e}{H_{CO2} \int_0^{\Delta t} q_{in} dt} \right)^2 \sigma_{pH}^2 + 2 \left( \frac{p_{tot} (1 + K_1 10^{pH})}{H_{CO2} OUR \Delta t} \right)^2 \sigma_y^2 + \left( \frac{q_{in}}{OUR} \right)^2 \sigma_E^2 \]  

(3.58)

While lower aeration reduces off-gas measurement uncertainty, it increases the impact of pH uncertainty. Consequently the aeration rate that balances this trade-off varies with pH, OUR, and the relative accuracy of the instruments (Figure 3-11).
Figure 3-10: The dependence of RQ uncertainty on pH assuming an aeration rate of 1 L L^{-1} min^{-1}, a 2 minute sampling period, and neglecting mass spectrometer inaccuracy.
Figure 3-11: The dependence of the optimum aeration rate for RQ monitoring on pH and OUR assuming $\sigma_y = 10^{-4}$ and $\sigma_{pH} = 0.01$. 
Chapter 4

MATERIALS AND METHODS

Table 4.1 summarizes the various experimental and manufacturing data sets used in this work. 'BIS, Inc. (Kingstree, SC) and Amylin Pharmaceuticals (La Jolla, CA) provided the two large manufacturing and development data sets. Several smaller data sets were collected in the MIT bioseparations Laboratory (MITPP) at the 75L and 1500L scale. The remainder came from experiments in our 20L Biolaftite reactor.

4.1 RQ Interpretation Case Study Example

In Section 5.2.1 we draw on data from the second of the six *B. subtilis* fermentations conducted with Stefan Winkler in collaboration with Roy Magnuson of the MIT department of Biology. We refer to these runs as BSROY1-BSROY6. The medium for these runs is described in Table 4.2. The fed-batch run BSROY6 used a 500 g L\(^{-1}\) glucose solution as the carbon source.

4.1.1 Experimental System

The fermentor (Biolaftite 20L) is equipped with probes for pH, dissolved oxygen (DO) and a Pt 100 temperature sensor. Gas flow is controlled by a Brooks thermal mass flow controller, connected to the control software (Paragon, Intec Controls) via an Opto-22 interface. The Paragon Software resides in an 80286 computer, running under the MS-DOS operating system. System data from the reactor and its peripherals are updated in intervals ranging from 0.25 s to 5 min. The communication between the mass-spectrometer (MGA 1600, Perkin Elmer, Norwalk, CT) and its data server (80286 computer) is done via a RS-232 connection. The communication between Paragon and the G2 control and data logging system is accomplished via the RS-422 nets. The G2 system, residing in a DEC 2100 computer, logs the fermentation data and also makes
Table 4.1: Summary of bioprocesses used in case studies

<table>
<thead>
<tr>
<th>Strain</th>
<th>Product</th>
<th>Scale(L)</th>
<th>Runs</th>
<th>Location</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. licheniformis</td>
<td>α-amylase</td>
<td>150,000</td>
<td>46</td>
<td>Kingstree</td>
<td>K</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>density indicator peptide</td>
<td>20</td>
<td>7</td>
<td>MIT</td>
<td>BSROY</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>biomass</td>
<td>20</td>
<td>2</td>
<td>MIT</td>
<td>RAJAY</td>
</tr>
<tr>
<td>E. coli</td>
<td>angiotensin</td>
<td>20</td>
<td>4</td>
<td>MIT</td>
<td>ECSON</td>
</tr>
<tr>
<td>A. vinelandii</td>
<td>nitrogenase</td>
<td>75</td>
<td>6</td>
<td>MIT PP</td>
<td>NUR</td>
</tr>
<tr>
<td>E. coli</td>
<td>platelet factor 4</td>
<td>1500</td>
<td>4</td>
<td>MIT PP</td>
<td>ECREP</td>
</tr>
<tr>
<td>E. coli</td>
<td>proprietary</td>
<td>75</td>
<td>4</td>
<td>MIT PP</td>
<td>BASF</td>
</tr>
<tr>
<td>E. coli</td>
<td>amylin</td>
<td>80</td>
<td>2</td>
<td>Amylin</td>
<td>API</td>
</tr>
<tr>
<td>E. coli</td>
<td>amylin</td>
<td>20</td>
<td>50</td>
<td>ARC</td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>amylin</td>
<td>150</td>
<td>9</td>
<td>ARC</td>
<td></td>
</tr>
<tr>
<td>E. coli</td>
<td>amylin</td>
<td>1500</td>
<td>1</td>
<td>ARC</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Medium for Bacillus subtilis experiments

<table>
<thead>
<tr>
<th>Component</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>(NH₄)₂SO₄</td>
<td>10 mM</td>
</tr>
<tr>
<td>K₂HPO₄</td>
<td>20 mM</td>
</tr>
<tr>
<td>KH₂PO₄</td>
<td>11 mM</td>
</tr>
<tr>
<td>NaGlutamate</td>
<td>0.1 g L⁻¹</td>
</tr>
<tr>
<td>glucose (batch runs)</td>
<td>20 g L⁻¹</td>
</tr>
<tr>
<td>citrate</td>
<td>0.1 g L⁻¹</td>
</tr>
<tr>
<td>trace metals</td>
<td></td>
</tr>
</tbody>
</table>
the data available over the internet. Data logs are stored locally under both the Paragon and the G2 system. Mass spectrometer data is logged and stored under the G2 system and the mass-spec control computer.

The pH level is controlled by the addition of acid and base solutions to the broth with peristaltic pumps (Masterflex, Cole-Parmer Instrument Co., Chicago, IL). The acid solution was a 2M $H_3PO_4$ and the base a 4.11 M $NH_4OH$. The pH controller works as a proportional on/off controller with a dead zone. The dead zone was set to ±0.1 pH units.

4.1.2 Experimental Operation

Each BSROY run was started with a working volume of 8 L; each ECSON run with a working volume of 9.9 L. To maintain the volume below 10 L, samples were withdrawn when necessary. Before each run, the entire reactor with probes and connected tubing was sterilized. Antifoam (polypropylene glycol MW 6000) was injected when needed. All runs were performed at 37° with air as oxygen source. The essential run data were logged manually every hour.

4.1.3 Gas Analysis

For off-gas measurements, we used a Perkin-Elmer model 1600 MGA mass spectrometer. Inlet and exit gas lines were monitored for $O_2$, $CO_2$, $^{13}CO_2$, $N_2$, $H_2O$, $NH_3$ and Ar concentrations. The data were logged continuously during the run and sent to the Paragon system for use in process control.

4.2 Monitoring of $^{13}CO_2$ Evolution

We monitored $^{13}CO_2$ evolution in three fermentations done in collaboration with Stefan Winkler, Cathryn Shaw, and Kwang Son. We refer to these experiments as ECSON1-ECSON3. These runs used a Escherichia coli strain (BL21DE3) harboring an IPTG inducible plasmid for the production of angiotensin. The fermentation medium (Table 4.3) of Sanchez (1989) was used in experiments. The carbon source for the ECSON1 and ECSON2 runs was glycerol while that for ECSON3 was glucose. A peristaltic pump fed the carbon source to the reactor on an exponential schedule.
Table 4.3: Medium for ECSON experiments

<table>
<thead>
<tr>
<th>Compound</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>(NH$_4$)$_2$SO$_4$</td>
<td>7.0 g L$^{-1}$</td>
</tr>
<tr>
<td>KH$_2$PO$_4$</td>
<td>1.6 g L$^{-1}$</td>
</tr>
<tr>
<td>Na$_2$HPO$_4$</td>
<td>5.3 g L$^{-1}$</td>
</tr>
<tr>
<td>(NH$_4$)$_2$H-citrate</td>
<td>0.5 g L$^{-1}$</td>
</tr>
<tr>
<td>MgSO$_4$</td>
<td>0.3 g L$^{-1}$</td>
</tr>
<tr>
<td>Trace elements soln</td>
<td>1.5 mL L$^{-1}$</td>
</tr>
</tbody>
</table>

Trace Element Solution:
Na$_2$-EDTA 20 g L$^{-1}$

CaCl$_2$ · 2 H$_2$O      | 0.5 g L$^{-1}$
FeCl$_3$ · 6H$_2$O       | 17.0 g L$^{-1}$
ZnSO$_4$ · 7H$_2$O       | 0.2 g L$^{-1}$
CuSO$_4$ · 5H$_2$O       | 0.2 g L$^{-1}$
MnSO$_4$ · H$_2$O        | 0.1 g L$^{-1}$
CoCl$_2$ · H$_2$O        | 0.2 g L$^{-1}$

designed for a specific growth rate of 0.3 h$^{-1}$.

The [1-$^{13}$C] glucose and acetate were obtained from Cambridge Isotope Laboratories (Woburn, MA). A magnetic-sector mass spectrometer (Perkin-Elmer MGA 1600) monitored the gas leaving the reactor at two-minute intervals. Measuring the voltage of the 44 m/e and 45 m/e collectors enable the $^{13}$CO$_2$ concentration to be estimated relative to the calibrated peak measured for $^{12}$CO$_2$. The ambient level of the isotope (1.195%) was subtracted to obtain the CO$_2$ attributable to the labeled substrate.

4.2.1 Sampling and Sample Treatment

Samples from the bioreactor cultures were taken frequently during the runs. Shortly before taking the samples, the sampling valve was sterilized with steam. Approximately 15 ml of broth were collected via the sample valve directly into a 15 ml test tube (Falcon 2095, Becton-Dickinson, Lincoln Park, NJ). The sample was immediately used for the determination of cell concentration, by measuring the optical density at 600 nm. The remaining sample was centrifuged at 800 rpm for 10 minutes to remove the cells. The supernatant was filtered through a 0.2 μL filter (UNIFLOW, Schleicher & Schnell, Keene, NH) and frozen at -40°C for subsequent analysis. The pellet was discarded.
4.2.2 Cell Density

The optical density of the broth was measured at 600 nm in plastic (Vis) 1 cm cuvettes, using a Hewlett-Packard model 8451 diode array spectrophotometer. The samples were diluted with distilled H₂O to remain in the linear area of the spectrophotometer. The conversion to cell mass was done by determining the correlation of OD units and dry cell weight. One optical density unit corresponded to 0.28 g dry cell weight L⁻¹ for *B. subtilis*, and 0.29 g dry cell weight L⁻¹ for *E. coli*.

4.2.3 Analysis of Nutrients and Waste Products

Glycerol, glucose, acetate, lactate, and ethanol levels in the broth were measured by HPLC (R401 Refractometer, Model 510 pump, WISP 710B sample processor, Waters Associates, Milford, MA). The chromatography conditions were: Aminex HPX 87 - H column (Biorad, Richmond, CA) at 40°C, mobile phase 0.0089 N H₂SO₄ at 1 ml min⁻¹. A Hewlett-Packard 3390A Integrator (Hewlett-Packard, Avondale, PA) was used to plot and calculate the heights of the individual peaks. The quantity of the substances was then determined via standard curves made by serial dilution from standard solutions of the above mentioned substances. Glucose, glycerol, acetate and ammonia concentrations also were measured according to the manufacturer's recommendations by using standardized enzymatic assay kits (Boehringer Mannheim, Indianapolis, IN). These were used when the HPLC analysis was not clear or when we wanted to obtain the concentration of a substance in real time.

4.3 NUR Monitoring Experiments

In order to test the strategy for monitoring of NUR by mass spectrometry, we conducted five *Azotobacter vinelandii* fermentations in the MIT pilot plant 75L reactor. The wild-type strain (OP) was used in four of the cultures while a non-nitrogen-fixing mutant (UW45) was used in the control. Table 4.4 summarizes the strains and experimental conditions used.

Cells were cultured on Burk’s nitrogen-free agar plates for 5 days. Single colonies were transferred to 5-mL tubes of Burk’s nitrogen-free media and grown for 24 h. Each
Table 4.4: Run conditions for NUR experiments

<table>
<thead>
<tr>
<th>Run number</th>
<th>Strain used</th>
<th>Aeration rate (L L^{-1} min^{-1})</th>
<th>Gas feed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OP</td>
<td>0.1-0.2</td>
<td>Special</td>
</tr>
<tr>
<td>2</td>
<td>OP</td>
<td>0.1-0.2</td>
<td>Normal</td>
</tr>
<tr>
<td>3</td>
<td>OP</td>
<td>0.1-0.2</td>
<td>Special</td>
</tr>
<tr>
<td>4</td>
<td>OP</td>
<td>0.1</td>
<td>Special</td>
</tr>
<tr>
<td>5</td>
<td>UW45</td>
<td>0.1</td>
<td>Special</td>
</tr>
</tbody>
</table>
Table 4.5: Fermentation medium used in NUR experiments

<table>
<thead>
<tr>
<th>Substance</th>
<th>Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sucrose</td>
<td>20 g L⁻¹</td>
</tr>
<tr>
<td>K₂HPO₄</td>
<td>1.602 g L⁻¹</td>
</tr>
<tr>
<td>KH₂PO₄</td>
<td>0.408 g L⁻¹</td>
</tr>
<tr>
<td>NaCl</td>
<td>0.197 g L⁻¹</td>
</tr>
<tr>
<td>FeCl₃</td>
<td>1.62 mg L⁻¹</td>
</tr>
<tr>
<td>NaCitrate</td>
<td>2.94 mg L⁻¹</td>
</tr>
<tr>
<td>MgSO₄.7H₂O</td>
<td>200 mg L⁻¹</td>
</tr>
<tr>
<td>CaCl₂.2H₂O</td>
<td>88 mg L⁻¹</td>
</tr>
<tr>
<td>NaMo₄.2H₂O</td>
<td>25 mg L⁻¹</td>
</tr>
<tr>
<td>H₃BO₃</td>
<td>0.88 mg L⁻¹</td>
</tr>
<tr>
<td>ZnSO₄</td>
<td>0.6 mg L⁻¹</td>
</tr>
<tr>
<td>CoCl.6H₂O</td>
<td>0.6 mg L⁻¹</td>
</tr>
<tr>
<td>MnSO₄.H₂O</td>
<td>0.017 mg L⁻¹</td>
</tr>
<tr>
<td>CuSO₄.5H₂O</td>
<td>0.01 mg L⁻¹</td>
</tr>
<tr>
<td>Urea (Run. 5)</td>
<td>0.609 g L⁻¹</td>
</tr>
</tbody>
</table>

5-mL tube was transferred to 500 mL of medium and cultured in baffled, 2.8-L Fernbach shake flasks at 30°C in an orbital shaker (Queue) at 225 rev min⁻¹ for 24 h. Table 4.5 lists the composition of the fermentation medium used in the runs. The inoculum medium was identical to the fermentation medium except that the concentrations of K₂HPO₄ and KH₂PO₄ were doubled to buffer against pH changes. Two to five flasks (1.0-2.5 L total) of cells at a density of 1.3-1.8 g L⁻¹ were aseptically added to a 50-L working volume reactor (Chemap AG, Switzerland).

The 50 L cultures were grown for 10-15 h at 30°C. Aeration was initiated at 5 L min⁻¹ and increased to 10 L min⁻¹ in situations where the upper limit of agitation might otherwise be reached before the run was completed (Runs 1-3). The pH was controlled at 7.0 through additions of 1 N KOH. The back pressure was maintained at 1.5 atm absolute. Agitation was varied automatically to control dissolved O₂ at 0.04-0.06 atm using a polarographic dissolved oxygen probe (Ingold). The gas mixture, fed to the reactor from size 300 cylinders, was pre-mixed gravimetrically to obtain a molar composition of 40% O₂, 45% N₂, and 15% Ar. The gas feeding system is shown in Figure 4-1.
Figure 4-1: Gas handling system for NUR measurement experiments.
4.3.1 Acetylene Reduction Assays

Three assays based on the acetylene reduction method (Postgate, 1972) were used to monitor the enzymatic activity of nitrogenase. They were performed by Patricia Christi, a doctoral student in the department of Chemistry. These are referred to as the in vivo, unsaturated in vitro, and saturated in vitro assays.

In the in vivo assay, empty glass vials were degassed and back filled with an assay mixture of 10% acetylene in argon. Various volumes of air were added to the assay vials to generate a titration curve of activity as a function of the amount of air added. The overpressurized vials were allowed to equilibrate through a bubbler to atmospheric pressure. An aliquot of 500 or 1000 μL of culture broth (depending on cell density) was next added to the assay vials and incubated at 30°C for 30 min at 120 rev min⁻¹. The reaction was stopped by the addition of 0.1 mL of a 30% trichloroacetic acid solution. A 100 μL sample of the head space gas was injected into a gas chromatograph (Perkin Elmer, Sigma 3B) equipped with a 10 foot × 1/8 inch Porapak-N column operated at a back pressure of 300 kPa. The amount of ethylene produced was compared to a calibrated ethylene standard.

In the unsaturated in vitro acetylene reduction assays, a 250 mL sample of cell broth was centrifuged at 4080g min at 4°C for 5 min. The supernatant was poured off and the pellet that remained was frozen in liquid nitrogen. We degassed the frozen pellet by repeatedly placing under vacuum and back filling the vial with argon. Aerobically degassed 0.1 M Tris buffer (pH 8) was added to obtain a 1:8 weight to volume dilution of the pellet. The cell suspensions was disrupted anaerobically in a French pressure cell at 4°C. An aliquot of 50 μL of the cracked cell mixture was assayed by an adenosine triphosphate (ATP) regenerating system using creatine phosphokinase (Davis & Orme-Johnson, 1976). After the assay was run for 30 min at 30°C, a 100 μL sample of the head space gas was injected into the gas chromatograph (GC) system described above. The amount of ethylene produced was compared with a calibrated ethylene standard.

Finally, the saturated in vitro assay seeks to approximate the in vivo activity by adding purified iron protein to the cell homogenate until the GC measurement
described above indicates saturation. In order for nitrogen fixation to proceed, the two enzyme subunits (iron protein and molybdenum-iron protein) must come into contact with each other for electron transfer to take place. The dilution of the two sub-units in the cell homogenate leads the reaction rate to be much lower than that seen in vivo (Thorneley & Lowe, 1984).

4.3.2 Reactor Monitoring

Optical density (OD) at 660 nm was measured off line using a spectrophotometer (Perkin Elmer). Optical density at 510 nm was measured in situ and on-line using an OD probe (Fundalux, Chemap AG, Switzerland). Dry cell weight was measured by vacuum filtering cell broth through a 0.22 μm filter (Millipore) and drying it overnight in a 60°C oven. We found that 1 g L⁻¹ of dry cell weight was equivalent to 4.5±0.06 OD₆₆₀ units and 28.7±0.3 OD₅₁₀ units.

Gas flow to the reactor was measured by a rotameter and, in runs 4 and 5 by a thermal mass flow meter (Brooks). Samples of the inlet and outlet gas streams were delivered continuously by metering pumps (FMI) to the mass spectrometer. After completing the experiments, we found evidence in the data that one gas-metering pump continuously entrained a small amount of ambient air. We corrected the data using an average leak rate of 0.12%. This value was determined by feeding custom gas in parallel directly to the mass spectrometer and through the gas metering pumps.

A magnetic-sector mass spectrometer (model MGA 1600, Perkin-Elmer, Pomona, CA) was used in all the experiments. The instrument ionizes incoming gases using electron bombardment in a vacuum envelope of 10⁻⁶ torr. The ionized gases are directed into four Faraday cup collectors allowing quantification of ions with a mass to charge ratio (m/z) between from 2 to 120. In this case, the inlet and outlet gases were monitored for O₂, N₂, Ar, CO₂, and H₂O; these compounds having m/z ratios of 32, 28, 40, 44, and 18 respectively. The instrument calibration contained a slight bias in the N₂ signal which varied linearly with CO₂ concentration. This occurred because a portion of the CO₂ ionizes to a CO fragment with a m/z ratio equal to that of nitrogen. After the bias in the instrument calibration matrix was quantified by feeding several mixtures of
CO₂ and custom gas to the mass spectrometer, more accurate N₂ concentrations were recalculated.

Data from two control systems (Satt, Alfa-Laval, Sweden; Paragon, Intec Controls, Walpole, MA) were transmitted over RS-422 serial connections to a software package for data logging and analysis (G2, Gensym, Cambridge, MA). The accuracy optimization problem was solved using a spreadsheet (Lotus 1-2-3 ver. 3.1+, Lotus Development Corporation, Cambridge, MA) and verified with custom optimization software (GAMS, Gams Development Corporation, Washington, DC).
Chapter 5

RESULTS AND DISCUSSION

This chapter begins with an example of using elemental balance data reconciliation to analyze a set of industrial fermentation data. This is followed by sections that apply the theoretical developments of Chapter 3 to experimental analyses of four off-gas variables – RQ, NUR, $k_La$, and $^{13}$CER. Finally, we present a case study that further illustrates the data reconciliation concepts described in Chapter 2 to an off-gas diagnosis problem.

The studies of RQ and NUR grew from a need to understand and evaluate elemental balance inconsistencies. The RQ and balance closure observed in a series of *Bacillus subtilis* cultures are discussed here. The fixation of nitrogen by *Azotobacter vinelandii*, which also exhibits unusual RQ, was monitored using a novel strategy developed for this purpose. We sought to use $k_La$ correlations and $^{13}$CER monitoring to increase the number of variables that could be reconciled and estimated.

5.1 Elemental Balance Data Reconciliation

An industrial *Bacillus licheniformis* fermentation producing $\alpha$-amylase served as the first model system studied in this work. The process had problems with variability and declining process performance in the preceding two years. To understand these problems, the company’s technical resources, located overseas, needed complete and accurate process data. This motivated the development of methods to assess the consistency of the data.

The process, shown schematically in Figure 5-1, typically operates for 110 to 120 h. The reactor has a working volume of 100 m$^3$ and is operated with an initial batch phase followed by a fed-batch phase with periodic harvesting of material. The initial reactor charge contained glucose, yeast extract, citrate, various salts and trace elements. Gaseous ammonia and phosphoric acid were used for pH control and corn syrup (40%
Figure 5-1: Schematic representation of process flows and measurements for the industrial enzyme fermentation.
Figure 5-2: On-line measurement for the total rates of O$_2$ uptake, CO$_2$ evolution, NH$_3$ addition, and sugar addition for Run K890.

dextrose equivalents) was added during the fed-batch phase. Figure 5-2 shows the trends for the four available on-line rate measurements: O$_2$ uptake, CO$_2$ evolution, NH$_3$ addition, and sugar addition. These values are calculated for the entire 100 m$^3$ of reactor broth with units of mol s$^{-1}$. The enzymatic activity of $\alpha$-amylase was measured in reactor samples intermitently and in each product withdrawal. No biomass or byproduct concentration measurements were available.

The lumped bioconversion can be modeled as:

$$CH_2O + O_2 + NH_3 + \text{other substrates} \rightarrow$$

$$\text{cells} + \text{amylase} + CO_2 + H_2O + \text{byproducts}$$

Because $\alpha$-amylase ($CH_{1.47}O_{0.30}N_{0.27}$) and B. licheniformis cell mass ($CH_{1.83}O_{0.57}N_{0.23}$) have similar degrees of reduction (4.07 and 4.02 respectively), they could not be
simultaneously estimated by elemental balancing and were lumped together as one
generic biomass term. Because the identity and quantity of byproduct(s) being
produced were unknown, we set the initial model to assume they were absent. We
assumed the concentrations of ammonia and sugar remained constant after exhaustion
of the initial glucose charge. This results in the simplified model:

\[ r_S \text{CH}_2\text{O} + r_O \text{O}_2 + r_N \text{NH}_3 \rightarrow r_X \text{CH}_{1.83}\text{O}_{0.57}\text{N}_{0.23} + r_C \text{CO}_2 + r_W \text{H}_2\text{O} \]

Using the elemental balancing approach described in Chapter 3, this results in five
unknowns and three balances for carbon, nitrogen, and degree of reduction. With three
of the five unknowns measured, there are two more measurements than necessary to
specify the system. The extra redundancy provides an opportunity to detect
inconsistencies in the data and to locate the most likely cause of disagreements when
they occur.

The O\textsubscript{2}, CO\textsubscript{2}, and sugar measurements were assigned an uncertainty of 0.3 mol s\textsuperscript{-1}
and the NH\textsubscript{3} measurement was assigned an uncertainty of 0.15 mol s\textsuperscript{-1}. Because all
these uncertainties were about 10 percent of the measurement’s values, each molar flow
was given equal weight in the reconciliation. Because the uncertainties may be
incorrect, the choice of the \(\alpha\) tolerance in the \(\chi^2\) test is somewhat arbitrary.

Hourly measurements from the batch were analyzed using data reconciliation.
Figure 5-3A shows the hourly p-value for rejection of the null hypothesis. The least
consistent measurement for each sample where \(\alpha\) was above 0.5 are shown in Figure
5-3B. Serial deletion of measurements for the batches with p-values higher than 50
percent indicated that the ammonia addition measurement was faulty with the
reconciled flow falling below the measured value. Plant operators investigated and
found the sensor had a calibration error that resulted in the reported measurement
being 50 percent above the actual flow. This correction was made retroactively to the
data set and the analysis repeated. Figure 5-4 shows the resulting improvement in data
set consistency.

Further review of this data showed that the calibration constant used in the
Figure 5-3: The upper plot shows p-values corresponding to the $\chi^2$ statistic for the hourly rate data. The lower plot shows the least consistent measurement at each time step.
Figure 5-4: The p-values for Run K890 after the $NH_3$ flow meter bias was removed.
reconciliation to calculate the sugar feed was eight percent too low. When this was corrected, the inconsistency of the data increased. After broth concentration measurements were made, this inconsistency was traced to the fact that about ten percent of the substrate was being used toward the accumulation of polysaccharides and techoic acid in the fermentation broth. Including this behavior in the model essentially offsets the previous correction for the sugar feed calibration and brings the consistency back to that seen in Figure 5-4. Ironically, if the calibration constant was correct in the first iteration, the ammonia error would have masked the polysaccharide accumulation and neither would have been detected.

We found that to incorporate all the available measurements in the reconciliation model and to identify the primary causes of inconsistency, nonlinear models and nonlinear data reconciliation would be required. In particular, there were numerous bilinear terms (e.g., flow and concentration) that required this. The sections that follow focus expanding the reconciliation model beyond the elemental balance.

5.2 Measurement Consistency Analysis

5.2.1 Use of RQ in Consistency Analysis

Our interest in RQ interpretation arose from inconsistencies found in a series of Bacillus subtilis cultures. The data from one of these runs (BSROY2) is presented in Figure 5-5. These experiments were done to scale-up the production of a density-indicator peptide (Magnuson et al., 1994). Our experimental goal was to implement a reconciliation-based strategy for controlling the rate of glucose addition in order to minimize nutrient imbalances and their consequent physiological effects.

The process exhibited unusual behavior including an unexpected demand for acid and material balance non-closure that suggested that unidentified substrates or products might be present. The inconsistent degree of reduction, evidenced by RQ levels far below unity, was particularly puzzling. Although RQ-derived estimates of yield can be inaccurate (see Figure 3-6), they do provide an upper bound on $Y_{XS}$ and establish redundancy between OUR and CER. The potential causes we investigated to explain the inconsistencies found in these runs included sensor bias and inaccuracy as
Figure 5-5: Measured and calculated values obtained from the BSROY2 experiment.
well as neglected reactants, products, and process dynamics. The calculated RQ trends for one of the runs (Run 2) are shown in Figures 5-6A-D to illustrate the consistency improvement after each fault was identified and corrected.

Mass spectrometer accuracy, evaluated through experiment and error propagation, was ruled out as a cause (Figure 3-9). First we found calibration errors in the mass spectrometer configuration. The CO₂ signal had to be scaled by 0.77, the O₂ signal by 0.95, and the N₂ ratio by 1.003. These eliminated most of the drift in RQ, however its value was below the expected level of approximately 1.05 (Figure 5-6B). Was the RQ indicating the cells were producing an unknown byproduct? The mappings between stoichiometry and RQ we developed in Section 3.6.6, showed the RQ level to be inconsistent with any common *Bacillus subtilis* byproducts.

It turned out that the cause of the inconsistent RQ was accumulation of dissolved CO₂ which we had neglected. The value we were calculating was actually the transfer quotient (TQ) not the RQ. Bioprocess control research is often conducted using model systems operated at acidic pH levels where chemically dissolved CO₂ and the CAR are negligible. Thus, the RQ and TQ are often used interchangably in the literature. Here the CAR was significant due to both the high pH level and the low rate of aeration. Ironically, we expected the low aeration to improve measurement accuracy. Looking again at Figure 5-6B, we see that the TQ shifts at 6 h and 9 h correspond to a decrease in the pH and an increase in the aeration rate respectively; exactly the behavior predicted by Equation 3.27.

Accounting for the CAR made the RQ of the other experiments consistent (RQ ≈ 1) but the RQ data of Run 2 remained in the 0.7 to 0.8 range (Figure 5-6C). Referring back to Figures 3-3 and 3-10, we see that both the (TQ-RQ) differential and the σ_{RQ} increase with pH. We found the computer-recorded pH level was 0.7 units below the true value, which we had recorded manually during the runs. We had left this bias in the control system intentionally at the start of the experiment in order to avoid a controller disturbance after calibration. The result was an under-estimation of the CAR and therefore the RQ. After we corrected the pH, the calculated RQ and σ_{RQ} became more consistent with the expected RQ of 1.05 (Figure 5-6D). The pH bias also was
Figure 5-6: Plots A-D depict the chronological progression of the RQ trends we calculated for experiment BSROY2. Figure B included a correction for off-gas measurement bias while Figure C incorporates bicarbonate accumulation estimated with biased pH reading. Finally, Figure D includes a correction for pH measurement bias.
Figure 5-7: Calculated RQ derived from TQ and pH measurements for BSROY2. The solid lines represent the expected σ limits for RQ based on Equation 3.58.

detectable by examining the noise in the calculated RQ. In Chapter 3, we derived a relationship predicting RQ uncertainty as a function of the pH level (Figure 3-10). The RQ noise observed here was far above that expected for pH 6.5 (Figure 5-7).

5.2.2 Use of $k_{La}$ in Consistency Analysis

The purpose of the following example is to demonstrate the utility of using a mass transfer correlation to add analytical redundancy to a reconciliation model. Figure 5-8 compares the dissolved oxygen measured by two independent probes with the dissolved oxygen estimated based on a correlated value of $k_{La}$ using Equation 3.29. The parameters of the $k_{La}$ correlation were obtained independently using data from the
experiment in the upper left (PPT191). These experiments were conducted in the MIT bioseparations laboratory’s 1500L Chemap fermentor. As can be seen in the figure, the estimated value tracks well with the values reported by the two probes installed in the reactor. In experiment ECREP3, the estimate is about 15 percent lower that the measured values. The consistency of the probes indicates a systematic mistake in the probe calibrations or in the measurement of gas flow or oxygen consumption, although we could not be verify this. In experiment ECREP4, the two probe measurements diverge by approximately 20%. In this case, the estimate was uncharacteristically noisy early in the run but over time the estimate converged with the value of the first probe. Encouragingly, a review of the operator logs shows the second probe was determined to be faulty before the run.

5.2.3 Use of Nutrient Analysis in Consistency Analysis

Figure 5-9 illustrates the results of using media nutrient analysis to troubleshoot a *Saccharomyces cerevisiae* fermentation process (Faruggio, 1995). The experimenter had expected the process to achieve a cell density of around 40 g L\(^{-1}\) but only obtained 7.5 g L\(^{-1}\). This analysis is described in more detail by Truong (1993). Comparing the range of cell densities predicted from the yields and medium composition with the target and actual cell densities, we found that the medium had insufficient amounts of histidine, tryptophan, and uracil; nutrients required by the auxotrophic strain. When the experimenter removed these limitations, the process performance improved significantly.

In several of the model systems we studied, we found that the algorithm was able to detect poorly-designed media. In several cases, the exercise found the investigators had not scaled the amount of an auxotrophic nutrient appropriately or it was being catabolized due to the carbon limitation imposed by fed-batch strategies.

5.3 Measurement of NUR

The results of our study of NUR monitoring in *Azotobacter vinelandii* are divided into six sections which follow. First we describe the baseline noise in the calculated
Figure 5-8: Comparison of dissolved oxygen measured by redundant probes with an estimate based on the $k_La$ correlated in the top left experiment.
Figure 5-9: Comparison of target and observed cell density with the expected range for each nutrient for a *Saccharomyces cerevisiae* process. This medium required greater amounts of histidine, tryptophan, and uracil to achieve the target.
NUR under normal conditions. This is followed by the selection of new operating conditions using the accuracy optimization strategy described in Section 3.7.1. The next two sections report the observed trends in cell growth and nitrogen uptake. Finally, we evaluate how the NUR compared with nitrogenase assays and discuss the systematic errors we found in the experimental apparatus.

5.3.1 Baseline Measurement Noise

At the start of this work, we were unsure if noise and bias in the measurement of Ar and N₂ gas concentrations would prevent accurate calculation of the NUR. To assess the stability and accuracy of the NUR calculation, we monitored three non-N₂ fixing *Escherichia coli* batches. We found the uncertainty in the calculated NUR (±6 mmol L⁻¹ h⁻¹) unacceptable for accurate NUR measurement given that the anticipated maximum NUR was less than 2 mmol L⁻¹ h⁻¹ (Figure 5-10). The high rate of aeration needed to meet the organism's demand for O₂ leads the cells to capture less than 0.2 percent of the N₂ fed. The expected noise level in the measurement of HER was lower (± 0.2 mmol L⁻¹ h⁻¹), but as stated earlier, *in vivo* the cells reuse the H₂ instead of releasing it stoichiometrically.

5.3.2 Gas Feed Optimization

To maximize the production of nitrogenase, the dissolved O₂ must be held at a partial pressure of 0.04 atm. Although reducing the N₂ fraction would marginally improve the accuracy, we chose to keep N₂ pressure well above the *Kₘ* for the enzyme to ensure sustained growth. To achieve this, the minimum N₂ partial pressure was set at 0.65 atm (75% of the atmospheric level). The optimization algorithm improved NUR accuracy by minimizing *qₘ* and *y₅₂* and maximizing *yₒ₂* and *pₜₐ₉*. Table 5.1 lists the assumptions used in calculating the optimum values. The result was *Fₘ* set at 5 L min⁻¹, *pₜₐ₉* set at 1.5 atm and *y₅₂* at 0.45. This left the *y₅₉* / *yₒ₂* ratio to be chosen based on the OUR and the minimum *pₒ₂*.

Figure 5-11 graphically summarizes the optimization results. Curve A illustrates the accuracy improvement when the inlet gas is modified by adding Ar to ambient air. Curve B represents the accuracy when the N₂ and O₂ fractions are varied.
Figure 5-10: The original variability in NUR measurement was much greater than the expected value of the NUR. This noise level was significantly reduced by optimizing the feed gas composition and operating conditions.

Table 5.1: Assumptions used in NUR gas mix optimization

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Min</th>
<th>Max</th>
<th>Initial</th>
<th>Final</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{vi}$</td>
<td></td>
<td></td>
<td>$5 \times 10^{-5}$</td>
<td>$5 \times 10^{-5}$</td>
<td>abs. vol %</td>
</tr>
<tr>
<td>$\sigma_{q_{in}/q_{in}}$</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>%</td>
</tr>
<tr>
<td>ln(A)</td>
<td></td>
<td></td>
<td>-15.45</td>
<td>-15.45</td>
<td>mmol L$^{-1}$ h$^{-1}$</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>3.44</td>
<td>3.44</td>
<td></td>
</tr>
<tr>
<td>OUR</td>
<td></td>
<td></td>
<td>2</td>
<td>150</td>
<td>mmol L$^{-1}$ h$^{-1}$</td>
</tr>
<tr>
<td>CER</td>
<td></td>
<td></td>
<td>2</td>
<td>150</td>
<td>mmol L$^{-1}$ h$^{-1}$</td>
</tr>
<tr>
<td>NUR</td>
<td></td>
<td></td>
<td>0.04</td>
<td>2</td>
<td>mmol L$^{-1}$ h$^{-1}$</td>
</tr>
<tr>
<td>$p_{O_2}^L$</td>
<td></td>
<td></td>
<td>0.04</td>
<td>0.04</td>
<td>atm</td>
</tr>
<tr>
<td>$p_{N_2}^L$</td>
<td>0.7</td>
<td></td>
<td></td>
<td></td>
<td>atm</td>
</tr>
<tr>
<td>N</td>
<td>100</td>
<td>800</td>
<td></td>
<td></td>
<td>rev min$^{-1}$</td>
</tr>
<tr>
<td>$F_{in}$</td>
<td>5</td>
<td>50</td>
<td></td>
<td></td>
<td>L min$^{-1}$</td>
</tr>
<tr>
<td>$p_{tot}$</td>
<td>1.0</td>
<td>1.5</td>
<td></td>
<td></td>
<td>atm</td>
</tr>
</tbody>
</table>
Figure 5-11: Estimated NUR uncertainty as a function of $y_{Ar}^{in}$ for air mixed with Ar assuming a OUR$_{max}$ value of 150 mmol L$^{-1}$ h$^{-1}$ (A) and for custom mixtures of Ar, O$$_2$$, and N$$_2$$ assuming OUR$_{max}$ values of 150 (B), and 50 (C) mmol L$^{-1}$ h$^{-1}$.

independently. The resulting higher O$$_2$$ fractions allow lower gas flows and larger changes in N$$_2$$ concentration. These results assumed an OUR$_{max}$ of 150 mmol L$^{-1}$ h$^{-1}$ while curves C and D show the results for values of 100 and 50 mmol L$^{-1}$ h$^{-1}$ respectively. As the Ar fraction in the inlet gas increases, the declining O$$_2$$ fraction must be offset by increased gas flow to meet O$$_2$$ demands. The resulting loss in accuracy leads to the inflection in the curves. The upper limit on Ar concentration is constrained by the minimum $p_{N_{2}}^{L}$ and the maximum $q_{in}$.

Note that raising the Ar concentration above 15% results in only a slight improvement in accuracy. Using an adjustable gas mixing system to tailor the feed to the current O$$_2$$ demand results in only a small improvement in NUR accuracy. On this basis, we chose the custom gas mixture (15 % Ar, 40% O$$_2$$, 45% N$$_2$$) for these
Figure 5-12: Cell density (A), specific growth rate (B), and expected NUR (C) profiles for experiments 1 (■), 2 (♦), 3 (▲), 4 (□), and 5 (◊).

5.3.3 Observed Cell Growth

The cell densities and specific growth rates observed at the end of the experiments ranged from 1.0 to 1.3 g L\(^{-1}\) (Figure 5-12A) and 0.1 to 0.4 h\(^{-1}\) (Figure 5-12B) respectively. The cell densities and growth rates at the end of the batches led to maximum NUR estimates of 0.4-0.6 mmol L\(^{-1}\) h\(^{-1}\), as shown in Figure 5-12C. These estimates, which assume cell nitrogen content of 10 percent by weight, were slightly lower than those we had predicted before before the experiments. After 7 h of cultivation, the four nitrogen-fixing cultures appear to converge to a common total growth rate (\(\mu X\)) of 0.11-0.17 g L\(^{-1}\) h\(^{-1}\). This behavior, indicative of an unidentified rate limitation in the reactor, might merit follow-up work by other researchers seeking to improve the process.
5.3.4 NUR Measurement

Figure 5-13 shows the calculated NUR trends for four *Azotobacter vinelandii* cultures (Runs 2-5). These NUR values include corrections for the systematic errors described in Section 4.3.2. Compared to the control run shown in Figure 5-10, Run 2 demonstrated the decrease in noise possible when the flow of ambient air was minimized. The noise level was further reduced in Runs 4 and 5, which were grown using the custom gas-mix. In contrast to Runs 2 and 4, which were harvested when the cells were still growing, we operated Run 3 until the cells depleted of the carbon source to observe the cessation of respiration and N\textsubscript{2} fixation. As expected, we did not measure any hydrogen evolution in the experiments. The data gap apparent at 5 h in Run 3 was due to a temporary communication failure while the disturbance at 5 h in Run 4 was due to the replacement of a gas cylinder. The drift seen in Run 5, which used a non-N\textsubscript{2}-fixing mutant, reflects unresolved systematic errors in the monitoring system.

During the first 6-8 h of the experiments using custom gas, the cells exhibited a high \( \mu \), a low NUR, and a low acetylene reduction rate (ARR). This pattern – which was seen in Runs 1, 3, and 4 – suggests that an alternate nitrogen source was consumed (*e.g.*, lysed cells present in the inoculum flasks). We did not find evidence to link the pattern to the use of the custom gas.

Figure 5-14 shows the specific uptake rates calculated by dividing the NUR and OUR by the on-line estimate of cell density. The negative values early in the run reflect the systematic uncertainty (±0.5 mmol L\textsuperscript{-1} h\textsuperscript{-1}) remaining in the NUR measurement. The specific NUR trends of three experiments converge to approximately 1.2 mmol g\textsuperscript{-1} h\textsuperscript{-1}. The specific OUR ranged from 50 to 100 mmol g\textsuperscript{-1} h\textsuperscript{-1}; a value 50 times higher than the specific NUR and five times higher than the specific OUR typically seen in *Escherichia coli*.

5.3.5 Assay Correlations with NUR

Figure 5-15 compares the calculated NUR with the assays for *in vivo*, unsaturated *in vitro*, and saturated *in vitro* acetylene reduction. We also compare the NUR with an estimate based on total cell growth rate. The observed relationships can be summarized
Figure 5-13: Measured NUR trends for runs with the N2 fixing strain aerated using regular air at a volume-to-volume per minute (VVM) rate of 0.1 L L\(^{-1}\) min\(^{-1}\) (Run 2), special gas at 0.1-0.2 L L\(^{-1}\) min\(^{-1}\) (Run 3), special gas at 0.1 L L\(^{-1}\) min\(^{-1}\) (Run 4), and for the non-N2-fixing run aerated using special gas at 0.1 L L\(^{-1}\) min\(^{-1}\) (Run 5).

Figure 5-14: Trends of measured OUR/X and NUR/X (mmol g\(^{-1}\) L\(^{-1}\)) for runs 1 (■), 2 (♦), 3 (▲), 4 (□) and 5 (◊).
Figure 5-15: Measured NUR correlated against in vivo (A), unsaturated in vitro (B), and saturated in vitro (C) acetylene reduction assays along with an estimate of N₂ demand based on cell growth (D).

\[ NUR = 0.8 \pm 0.1 \text{ in vivo } ARR \]  
\[ = 6.0 \pm 0.8 \text{ unsaturated in vitro } ARR \]  
\[ = 0.4 \pm 0.1 \text{ saturated in vitro } ARR \]  
\[ = 1.7 \pm 0.7 \frac{\mu X 1000}{Y_{XN} MW_{N_2}} \]

The expected ratio of the NUR to the ARR is 0.33 mmol N₂ (mmol C₂H₂)⁻¹ because the electron transfer required to reduce N₂ are three times that of acetylene. The ratio seen in the in vivo assay (≈ 1 mmol mmol⁻¹) suggests the assay has a rate-limitation other than electron transfer. The ratio seen in the unsaturated in vitro
assay (6 mmol mmol⁻¹) shows an apparent 20-fold reduction in activity; this reflects the dilution of the two enzyme subunits. The ratio of the saturated in vitro assay (0.4 mmol mmol⁻¹) approaches the theoretical value of 0.33 because excess iron protein is added to the cell homogenate compensate for the enzyme dilution. The correlation between the various assays and cell growth is weak. This might have been caused by noise in the cell density estimate, variations in nitrogen content, or transient accumulation and depletion of $NH_4^+$ during the run.

5.3.6 Systematic Errors

Although the use of the modified inlet gas reduced the random measurement error, we found it also created and magnified systematic errors. In Run 1 we found the aseptic sampling bottles released approximately 400 cm³ of ambient air into the reactor with each sample, perturbing the measurement for approximately 20 minutes. We purged the bottles with the custom gas in subsequent runs to eliminate this problem. Similarly, we found that one of the gas metering pumps allowed a small amount of normal air (0.1-0.2% of total flow) to leak into the sampling stream. We quantified this leak experimentally and corrected the data as described in the methods section. The O₂ enrichment and low gas flow led to high concentrations of CO₂ in the exit stream. This accentuated small errors present in the mass spectrometer calibration matrices. We also experimentally quantified and corrected these errors. While these corrections improved the quality of the data, the possibility of corruption by remaining systematic errors must be considered when interpreting our quantitative results.

5.4 Monitoring of $^{13}$CO₂ Evolution

In these experiments we wanted to test the hypothesis that $^{13}$CO₂ evolution could be measured on-line by mass spectrometry and used to monitor cell metabolism. Our first experiment measured $^{13}$CER after we added labeled glucose to an Escherichia coli culture growing on glycerol. We were curious how quickly the cells would use the alternate substrate and if the signal to noise ratio in the measurement would be reasonable. In the second experiment, we added labeled acetate to a culture growing on
glucose. Our aim was to see if we could use the $^{13}$CER measurement to help control the rate of glucose addition. Specifically, we wanted to determine if the measurement revealed whether the cells were consuming or producing acetate, an undesirable byproduct of the glucose metabolism.

5.4.1 $^{13}$CER Profile and Label Balance

After 20 h of cultivation, the cells had grown to a optical density of 49 (600nm). At this point we injected 0.5 g L$^{-1}$ of glucose, 20\% $^{13}$C-labeled in the 1st position, into the reactor. Figure 5-16A shows the baseline measurement noise ($\sigma_n = 0.0004$) of the exit CO$_2$ concentration. Figure 5-16B shows the observed $^{13}$CER profile shortly before and after the pulse addition. Forty percent (2.23 of 3.33 mmol) of the labeled carbon exited the reactor as CO$_2$. This is consistent with a cell yield of approximately 0.25 g L$^{-1}$ (Bruinenberg et al., 1986). The actual yield is likely higher due to the concomitant utilization of glycerol. The data show a 5 min lag from addition to the first sign of label at the mass spectrometer and a 5-7 min transient from the initial rise to the peak concentration. These reflect the effects of the transport, mixing, and reaction steps in the path between label injection and off-gas measurement as well as exit-line transport lag, reactor headspace dilution, and timing uncertainty.

5.4.2 Monitoring Acetate Concentration and Consumption

In this experiment, 0.1 g L$^{-1}$ $^{13}$C-labeled acetate was added to a Bacillus subtilis culture at 24.61 h; just before the cells exhausted the glucose. Label consumption began immediately due to the low concentration of glucose at that point. When the glucose was exhausted (25 h), a momentary drop in the CER was followed by a symmetric rise in both the labeled and unlabeled CER (Figure 5-18). The symmetry indicated growth on acetate and the relative rates suggested an acetate concentration (3 g L$^{-1}$), which was consistent with off-line assays. The total $^{13}$CO$_2$ attributed to labeled substrate was 29\% above our expectations. This suggests either measurement error or the need for more detailed accounting of the fate of carbon based on its position in the substrates.
Figure 5-16: Measured concentration of $^{13}$CO$_2$ above ambient level (A) translated to estimates of $^{13}$CER and CER from tracer solution (B).
Figure 5-17: Trends for glucose, cell mass, and estimated acetate during experiment ECSN4
Figure 5-18: Labeled and unlabeled response to labeled acetate added just prior to glucose exhaustion. The top figure shows label evolution relative to normal CO₂. The bottom figure shows the relative levels of CER.
5.4.3 $^{13}$CER Dynamics

We recommend follow-up studies to develop a method to separate physiological parameters from the various physical, chemical, and transport phenomena involved. Figure 5-19 illustrates some of the phenomena that must be modeled to predict the $^{13}$CER profile after labeled substrate is added. The phenomena include transport of labeled material into and out of the cell, its residence time in the bicarbonate pool, and the transport lag and dispersion in the exit gas stream.

Several of these aspects can be modeled as CSTRs in series:

$$\frac{dx}{dt} = D(x - x_{in}) \quad (5.5)$$
$$x = x_{in} - (x_{in} - x_0)e^{-Dt} \quad (5.6)$$

Here $x$ is the molar fraction of the label in the head space and exit stream, $x_0$ is the initial concentration, $x_{in}$ is the mole fraction of the label in the gas leaving the broth, and $D$ is the head space dilution rate. This model predicts that 99% of the transient seen in the second experiment would be achieved in 2.3 min. The 45 minute tail from the peak back to baseline likely due to the wash-out of CO$_2$ in the bicarbonate pool.

The potential for use of this measurement in data reconciliation, process modeling, and process control is promising. We discuss several areas of potential research in Chapter 6.

5.5 Nonlinear Data Reconciliation

As summarized in Table 4.1, this project examined data reconciliation applied to data collected from a variety of processes, equipment configurations, and organisms. Here we illustrate the use of data reconciliation by examining a fermentation from the Amylin data set. The Amylin data set contains approximately 50 runs conducted in six sets of experiments at the Alberta Research Council facility in Edmonton, Canada. In addition, it contains data collected at Amylin's research facility in San Diego, CA and from the University of Georgia. Overall, the data sets contained well over $10^6$ data points. The experimental objective was to scale up and optimize the production of a
Figure 5-19: Flow diagram of the dynamics associated with $^{13}$CER monitoring
recombinant form of amylin, a peptide for treating Type I diabetes. In contrast to manufacturing processes where conditions are relatively routine and stable, the operating conditions in these experiments were highly variable due to intentional variation of parameters as well as variability associated with early-stage control strategies and operational learning curves.

5.5.1 Experimental Background

Figure 5-20 plots the trends of a subset of the measurements and calculations from experiment 2033. The data appear noisy, particularly the pH and dissolved oxygen measurements and off-gas derived calculations of OUR, CER, RQ, and $k_La$. The incident log for these experiments did not mention abnormalities in the run. The DO and pH transients seen in the data are artifacts of the operators adding glucose when sudden increases in DO signaled glucose depletion. The OUR and CER trends were more puzzling: data in the 18-20 h range appeared to have definite faults while the validity of the remaining data was unclear.

We should note we had previously identified and corrected three errors in the calculations provided by the facilities computer systems. First, the computer system pre-dated the use of oxygen enrichment and consequently the calculations of OUR and CER were incorrect. Secondly, the computer data log also implied independent measurements of oxygen, air, and total-gas flow whereas the latter value was simply a (noisy) calculation by the flow controller. Finally, the $k_La$ values in the data log were calculated using the incorrect OUR and a $k_La$ equation (Buckland et al., 1985) valid only for small O$_2$ stripping rates. These problems were typical of what we found in the other systems examined (including our own) and illustrate the need for data reconciliation even in cases where instruments function properly. These problems also motivate the use of model-based data reconciliation before black-box statistical methods for modeling and fault diagnosis that assume that base line data are valid.

5.5.2 Observability and Redundancy Classification

Table 5.2 summarizes the equations used in the full reconciliation model for this system. In this example, we draw on the subset of these equations that link the
Figure 5-20: Measured and calculated data from Run A2033.
Table 5.2: Relationships used in the bioprocess data reconciliation example

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon, nitrogen, and degree of reduction balances</td>
<td>3</td>
</tr>
<tr>
<td>Liquid ammonia and sugar balances</td>
<td>2</td>
</tr>
<tr>
<td>OTR, CER, and RQ definitions</td>
<td>3</td>
</tr>
<tr>
<td>$k_{La}$ calculation and correlation</td>
<td>2</td>
</tr>
<tr>
<td>$p_{O_2}^L$ and $p_{O_2}^L$ definitions</td>
<td>2</td>
</tr>
<tr>
<td>Cell mass, cell moles, OD correlations</td>
<td>2</td>
</tr>
<tr>
<td>SFR calculation</td>
<td>1</td>
</tr>
<tr>
<td>$F_{gasin}=f(q_{in}, V_L)$</td>
<td>1</td>
</tr>
<tr>
<td>Cell yield calculation</td>
<td>1</td>
</tr>
<tr>
<td>Biomass = growth rate * cell mass</td>
<td>1</td>
</tr>
<tr>
<td>Equal DO Probes</td>
<td>1</td>
</tr>
<tr>
<td>Know inlet Ar/N2, O2/Ar ratios</td>
<td>2</td>
</tr>
<tr>
<td>Constant Ar/N2 inert ratio</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.3: Variable classifications for models using only the O2 balance, the O2 balance and the second O2 measurement, and these plus the $k_{La}$ correlation. The symbols M, O, R, RO, and RR refer to classification as measured, observable, redundant, robustly observable, and robustly redundant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>O2 balance</th>
<th>O2 analyzer</th>
<th>$k_{La}$ correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{air}$</td>
<td>M</td>
<td>M</td>
<td>R</td>
</tr>
<tr>
<td>$F_{O_2}$</td>
<td>M</td>
<td>M</td>
<td>R</td>
</tr>
<tr>
<td>OTR</td>
<td>O</td>
<td>O</td>
<td>RO</td>
</tr>
<tr>
<td>$y_{O_2}^{MS}$</td>
<td>M</td>
<td>R</td>
<td>RR</td>
</tr>
<tr>
<td>$y_{O_2}^{analyzer}$</td>
<td>M</td>
<td>R</td>
<td>RR</td>
</tr>
<tr>
<td>$p_{tot}$</td>
<td>M</td>
<td>M</td>
<td>R</td>
</tr>
<tr>
<td>N</td>
<td>M</td>
<td>M</td>
<td>R</td>
</tr>
<tr>
<td>DO</td>
<td>M</td>
<td>M</td>
<td>R</td>
</tr>
</tbody>
</table>

gas-related variables shown in Figure 5-21. After these models were assembled, the next step was to classify the redundancy and observability of the variables. Table 5.3 summarizes the classification results of using the methods described in Section 2.3. The first column assumes the use of the oxygen balance and ignores the safety oxygen analyzer measurement and mass transfer equation. Here all the variables were classified as measured except for the calculated OUR, which was observable.

In addition to the mass spectrometer that monitored gas composition, this reactor also had an oxygen analyzer installed at the exit stream due to the explosion risks of using pure oxygen. Fortunately, the data from this instrument also was directed to the
Figure 5-21: Schematic representation of the system used in the variable classification example.
data log, providing us with more redundancy to assess the inconsistent gas data. When we include the relationship representing the direct redundancy between the analyzer and the mass spectrometer in the reconciliation model, we see a consequent reclassification (column 2) with these variables now redundant. Finally, adding a third relationship equating the calculated and correlated values of the overall volumetric mass transfer coefficient shifts several of the classifications. First, the exit measurements become robustly observable, that is, removal of any other particular measurement leaves them redundant. Second, the OTR becomes robustly observable since its calculation can withstand the loss of any particular measurement. Finally, new tenuously redundant groups emerge relating pressure, agitation rate, and dissolved oxygen as well as the flows of air and oxygen.

This information is useful in selecting the models for reconciliation and for determining what gross errors can be detected and identified. The results indicated the DR algorithm could detect significant problems with any of these measurements and positively identify them if they are present in the composition measurements.

5.5.3 Detection of Gross Errors

Figure 5-22A shows that the original data are inconsistent, particularly in the neighborhood of 16-19 h. The larger the value of $\alpha^{-1}$, the less consistent the data. The y-axis is plotted as the inverse of the p-value ($\alpha$) at which the inconsistency could be considered random. The horizontal line at $\alpha^{-1} eq 10$ represents the tolerance above which we chose to investigate inconsistencies. The lower line shows the p-values after detection and removal of faulty data caused by a leak in the mass spectrometry sampling system. This leak led to bias in the reported values for the exit gas compositions. Evidence for this bias can be seen in Figure 5-22B which plots the MS value along with that reported by the safety analyzer. The mass transfer correlation provided the tie-breaking redundancy to discern which of these two measurements is correct. Off-line data on cell mass production and sugar consumption, available after the run, provide an estimate of OUR which also supports this conclusion.
Figure 5-22: P-value trend for Run 2033. The y-axis is plotted as the inverse of the p-value; the larger the value, the less likely the inconsistency is due to random error. The large values in the upper plot reflect the major fault in the measurement of exit gas concentration shown in the lower plot.
5.5.4 Significance

The impact of the undetected fault on the carbon balance can be seen in Figure 5-23 which plots the molar inventory of carbon over the course of Run 2034, one of the other affected experiments. This method of presentation shows the total amount of glucose present in the medium and addition vessels at the start of the run. As this material disappears, the carbon is accounted for in CO₂, biomass, and acetate. Ideally the height of the stacked bars, which are normalized based on the initial inventory, should remain at 100%. Contrasting the two plots we see the erroneous inventory of evolved CO₂ leads to a large error in the carbon balance relative to the corrected data. The acetate portion is absent from the first plot from 22h onward since its value was not measured. The values are present in the second plot because we were able to estimate them based on the measured values of ammonium accumulation.

Lacking the diagnosis of the mass spectrometer problem, we would have had to question the validity of the entire data set and the data collected in the five other experiments of the campaign which were similarly corrupted. We found the final product concentration achieved was highly correlated with the maximum OUR achieved during the run. Discovering this relationship would have been difficult if the erroneous data were not corrected. Because the large-scale facilities planned for this process had much lower oxygen and heat transfer capabilities, the economic potential of the process was far below that implied by the small-scale experiments. Although the impact of the oxygen transfer constraint could likely be reduced by adjusting the operating strategy, this problem was likely a factor in the company’s shift in development effort towards chemical synthesis rather than fermentation.
Figure 5-23: Carbon balance closure before (A) and after (B) MS fault removal.
Chapter 6

CONCLUSIONS AND
RECOMMENDATIONS

This research evaluates and improves the methods for using data reconciliation to analyze bioprocess data. It defines and organizes the necessary assumptions, models, and correlations into a data reconciliation (DR) framework to assist in process diagnosis. The challenges encountered when analyzing data from a variety of industrial and lab-scale data sets helped clarify modeling and methodology priorities. Our experiments using DR for process control illustrated that a the large portion of real-time data is contributed by off-gas measurements. In order to interpret inconsistent off-gas data, we examined how to calculate and predict respiratory quotient (RQ) and mass transfer coefficient ($k_La$) values. We developed and evaluated two novel strategies using mass spectrometry to extract information from off-gas data. We measured nitrogen uptake rate (NUR) in $A.\ vinelandii$ cultures in order to monitor nitrogenase enzyme activity and used measurements of $^{13}$CO$_2$ evolution as a means to monitor the metabolism of labeled substrates.

6.1 Conclusions

6.1.1 Nonlinear Data Reconciliation

Our case study of a large scale enzyme production fermentation process showed that reconciling the complete set of measurements and assumptions necessitates the use of nonlinear data reconciliation. Using only linear models limits the effectiveness of DR because the elemental balances because the models include only a subset of the measurements, and these often impact more than one calculated rate. For example, faults in the calculation of OUR and CER are usually caused by problems in the measurement of gas flow or in the mass spectrometer system. This means that the
calculated OUR and CER can be faulty yet consistent with each other. We found OUR estimates based on historical correlations of the overall volumetric mass transfer coefficient \((k_La)\) useful in providing additional information to detect OUR faults as well as dissolved oxygen probe errors. We also found the need to establish tie-breaking analytical redundancy through the use of additional models such as mass transfer correlations and nutrient yields.

6.1.2 Nonlinear Variable Classification

A data reconciliation methodology must be able to classify dynamically the redundancy and observability of the variables. Techniques offered in the literature for this task apply only to linear models or particular nonlinear models. We adapted a matrix projection technique proposed by van der Heijden (1994a) to address this problem. We also classified the redundant and observable variables into subsets based on whether they remained redundant (or observable) when other measurements were lost. This approach identifies subsets of indistinguishable faults and thereby reduces the computation time and clarifies the interpretation of the diagnostic results.

6.1.3 Respiratory Quotient

Bicarbonate dynamics cause the RQ to deviate from the transfer quotient (TQ) in processes that are operated near or above neutral pH levels. This deviation is apparent during sudden process upsets but also is significant during the quasi-steady-state exponential growth. Overlooking the difference between RQ and TQ can lead data reconciliation to incorrectly suggest problems with sensors, equipment, or cell metabolism. Conversely, we showed that measuring this difference provides a means to detect incorrect values of pH, gas flow, and growth rate. We derived models and heuristics to summarize the effects of accuracy, stoichiometry, substrate composition, biomass composition, and bicarbonate dynamics on RQ and TQ values. As the pH level rises, the determining factor in the accuracy of RQ calculations shifts from off-gas accuracy to pH accuracy. Consequently, the optimum aeration rate for RQ accuracy increases with pH and OUR.
6.1.4 Monitoring of $^{13}$CO$_2$ Evolution

We evaluated the feasibility and utility of using mass spectrometry to monitor $^{13}$CO$_2$ evolution during fermentations. The results validated the capability of the MGA 1600 mass spectrometer to measure small amounts of $^{13}$CO$_2$ with a fairly low noise level ($\sigma_v = 0.004$). This approach can help identify substrates in use and, when used with NMR analysis, monitor the metabolic fluxes through various pathways. This approach allows us to continuously monitor the process and to avoid the use of radioactive substances.

6.1.5 NUR monitoring

Our NUR study estimated *A. vinelandii*’s rate of gaseous nitrogen uptake, confirmed the feasibility of monitoring it using mass spectrometry, and compared the measurement with traditional assays of nitrogenase activity. We determined the operating adjustments necessary to optimize the accuracy of the uptake rate, which was extremely noisy under normal conditions. The NUR has a maximum value of approximately 2 mmol L$^{-1}$ h$^{-1}$ and a specific rate of approximately 1.2 mmol g$^{-1}$ h$^{-1}$. As expected, the cells did not produce $H_2$ gas. The NUR measurement correlated fairly well with the three acetylene reduction assays used in the experiment. The saturated *in vitro* assay approached the expected 1:3 ratio of the NUR to ARR. It appeared that factors other than electron transfer were rate-limiting for the *in vivo* and unsaturated *in vitro* assays. This study demonstrated a novel method to improve measurement accuracy through rational adjustment of the operating conditions. The measurement technique provides a unique method to monitor enzyme activity on-line for use in process development and harvest timing.

6.2 Recommendations for Future Research

6.2.1 Data Analysis Automation

Reviewing our research results, we find three areas that merit further study. The first is the ongoing development of methodologies and tools necessary to automate analysis of large data sets. These tasks include data filtering and pre-processing, interpretation of reconciliation results, incorporation of dynamic models, improvement
of nonlinear optimization computational speed, convergence reliability, and management of models and assumptions.

6.2.2 Nutrient Medium Design and Analysis

A second research area is a more in-depth study of bioprocess medium design. Ideally, an automated analysis would consider salt precipitation, nutrient degradation by sterilization, nutrient toxicity, catabolism of auxotrophic supplements, substitution of nutrients, and influences on product formation. This could help yield more cost-effective processes and reduce variability associated with the use of semi-defined media.

6.2.3 Monitoring of $^{13}$CO$_2$ Evolution

A third research area is the use of $^{13}$CER measurements to monitor substrate utilization, substrate concentration, byproduct formation, and byproduct degradation. In particular, it would be interesting to monitor and control acetate metabolism during high cell density bacterial fermentations. Obtaining accurate metabolic data requires development of a detailed model of the dynamics of CO$_2$ accumulation. This model can be developed from the models for RQ interpretation and bicarbonate dynamics described here. The high cost of labeled substrates makes it important to minimize their use. We found future experiments can obtain accurate results using 25 percent of the substrate used in our preliminary experiments.
References


142


Chapter A

Appendix

A.1 Nomenclature

Variables

\( cMW_i \)  molecular weight of component i normalized to 1 c-mol mol\(^{-1} \)

\( C_i \)  concentration of component i in liquid phase (mol L\(^{-1} \))

\( C_{O_2}^* \)  concentration of O\(_2\) in equilibrium with gas phase (mol L\(^{-1} \))

\( DO \)  partial pressure of O\(_2\) relative to reference condition (%)

\( \gamma_i \)  degree of reduction of component i

\( G_1 \)  empirical constant for \( k_{La} \) design correlation

\( G_2 \)  empirical constant for \( k_{La} \) design correlation

\( A \)  empirical constant for \( k_{La} \) correlation (Eq. 3.23)

\( B \)  empirical constant for \( k_{La} \) correlation (Eq. 3.23)

\( C \)  empirical constant for \( k_{La} \) correlation (Eq. 3.23)

\( H_i \)  Henry’s law coefficient for component i (atm L mol\(^{-1} \))

\( K_1 \)  equilibrium constant for CO\(_2\) and HCO\(_3^-\)  

\( K_2 \)  equilibrium constant for HCO\(_3^-\) and CO\(_3^{2-}\)

\( k_{CO_2} \)  rate constant for CO\(_2\) reaction with H\(_2\)CO\(_3\)

\( k_{La}^i \)  overall volumetric mass transfer coefficient (mmol L\(^{-1} \) h\(^{-1} \) atm\(^{-1} \))

\( k_{La}^i_c \)  overall volumetric mass transfer coefficient (h\(^{-1} \))

\( p_{tot} \)  total pressure (atm)

\( p_{cal} \)  total pressure at time of probe calibration (atm)

\( p_i^G \)  partial pressure of component i in the gas phase (atm)

\( p_i^L \)  partial pressure of component i in the liquid phase (atm)

\( Q_{heat} \)  rate of heat production (kcal L\(^{-1} \))

\( r_i \)  rate of consumption of component i (mmol h\(^{-1} \))

\( R \)  ideal gas constant

\( T \)  temperature (K)

\( X \)  dry cell weight concentration (g L\(^{-1} \))

\( V_L \)  liquid volume (L)

\( Y_{X_i} \)  yield of cell mass on component i (mol mol\(^{-1} \))

\( y_i^{in} \)  mole fraction of component i in the inlet gas stream

\( y_i^{out} \)  mole fraction of component i in the outlet gas stream

\( \delta \)  ratio of outlet molar gas flow to inlet molar flow

\( \Delta_i \)  normalized change in gas mole fraction across reactor
\( \mu \) specific cell growth rate (h\(^{-1}\))

**Gas Rates** (mmol L\(^{-1}\) h\(^{-1}\))

\( F_{in} \) volumetric flow of gas addition (L min\(^{-1}\))

\( q_{in} \) specific molal rate of gas addition (mmol L\(^{-1}\) h\(^{-1}\))

CER CO\(_2\) evolution rate

CPR CO\(_2\) production rate

OUR O\(_2\) uptake rate

OTR O\(_2\) transfer rate

VVM volume gas sparged per volume liquid per minute (L L\(^{-1}\) min\(^{-1}\))

UR\(_i\) Uptake rate of component i

RQ Respiratory quotient

TQ Transfer quotient

TR\(_i\) Transfer rate of gas i to the liquid

**Scalars**

m number of measured variables

c number of unmeasured variables

e number of constraints

n number of variables

**Matrices**

\( A \) Linear constraint matrix

\( \Sigma \) Covariance matrix

\( \Psi \) Measurement covariance matrix

\( \Phi \) Residual covariance matrix

\( R \) Redundancy matrix

\( P \) Projection matrix

\( I \) Identity matrix

**Vectors**

\( x \) state vector

\( p \) parameter vector

\( z \) general parameter vector

\( e \) constraint residual error vector

\( \delta \) parameter error vector

\( \lambda \) Lagrange multiplier vector

**Functions**

\( f \) equality constraint function

\( J \) Jacobian

**Superscripts**

\( \# \) pseudo inverse

**Subscripts**

c calculated subset

m measured subset

**Diacriticals**

\( \hat{\cdot} \) measured

\( \check{\cdot} \) true

150
<table>
<thead>
<tr>
<th><strong>Statistics</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$E[\cdot]$</td>
<td>Expected value operator</td>
</tr>
<tr>
<td>$m$</td>
<td>degrees of freedom</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>chi-square</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>p-value; probability of $\chi^2$ statistic greater than tolerance</td>
</tr>
<tr>
<td>$h$</td>
<td>chi-square statistic</td>
</tr>
<tr>
<td>$\text{cov}(x)$</td>
<td>Variance-covariance operator $= E[xx^T]$</td>
</tr>
<tr>
<td>$\chi_{1-\alpha}^2(m)$</td>
<td>chi-square tolerance</td>
</tr>
<tr>
<td>$\sigma_i$</td>
<td>standard deviation of variable i</td>
</tr>
<tr>
<td>$H_0$</td>
<td>Null hypothesis</td>
</tr>
</tbody>
</table>


A.2 Media Yields

The following four tables list information useful for fermentation media design and analysis that we collected from the literature. The use of these data is described in Chapter 2.
Table A.1: Composition of biologically derived nutrients in terms of elements (g g\(^{-1}\)), amino acids (g g\(^{-1}\)) and vitamins (\(\mu g g^{-1}\))

<table>
<thead>
<tr>
<th>Elements and Pseudo-elements</th>
<th>Yeast Extract</th>
<th>Trypticase Peptone</th>
<th>Yeast Extract</th>
<th>Tryptone</th>
<th>N-Z-Case</th>
<th>Whey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nitrogen</td>
<td>0.103 g</td>
<td>0.117 g</td>
<td>0.092 g</td>
<td>0.131 g</td>
<td>0.133 g</td>
<td>0.025 g</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.002 g</td>
<td>0.0032 g</td>
<td>0.027 g</td>
<td>0.026</td>
<td>0.0081</td>
<td>–</td>
</tr>
<tr>
<td>Chlorine</td>
<td>0.003 g</td>
<td>0.003 g</td>
<td>–</td>
<td>0.0029</td>
<td>0.004</td>
<td>0.0088</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.0006 g</td>
<td>0.0035 g</td>
<td>0.0004 g</td>
<td>0.00096</td>
<td>0.0003</td>
<td>0.0039</td>
</tr>
<tr>
<td>Iron</td>
<td>0.002 g</td>
<td>0.0003 g</td>
<td>0.00028 g</td>
<td>0.00007</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.034 g</td>
<td>0.0024 g</td>
<td>0.00042 g</td>
<td>0.003</td>
<td>0.0007</td>
<td>0.023</td>
</tr>
<tr>
<td>Magnesium</td>
<td>0.0007 g</td>
<td>0.003 g</td>
<td>0.0003 g</td>
<td>0.00045</td>
<td>0.0002</td>
<td>0.0019</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>0.012 g</td>
<td>0.0065 g</td>
<td>0.0089 g</td>
<td>0.0075</td>
<td>0.0081</td>
<td>0.0041</td>
</tr>
<tr>
<td>Sulfur</td>
<td>–</td>
<td>0.0073 g</td>
<td>–</td>
<td>0.0004</td>
<td>0.0023</td>
<td>0.0003</td>
</tr>
<tr>
<td>Alanine</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.027</td>
<td>0.012</td>
</tr>
<tr>
<td>Arginine</td>
<td>0.035 g</td>
<td>0.026 g</td>
<td>0.0078 g</td>
<td>0.033</td>
<td>0.032</td>
<td>0.005</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>–</td>
<td>0.051 g</td>
<td>0.051 g</td>
<td>0.064</td>
<td>0.063</td>
<td>0.019</td>
</tr>
<tr>
<td>Cysteine</td>
<td>0.016 g</td>
<td>0.003 g</td>
<td>–</td>
<td>0.0019</td>
<td>0.0032</td>
<td>–</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>–</td>
<td>0.17 g</td>
<td>0.065 g</td>
<td>0.19</td>
<td>0.18</td>
<td>0.027</td>
</tr>
<tr>
<td>Glycine</td>
<td>–</td>
<td>0.018 g</td>
<td>0.024 g</td>
<td>0.024</td>
<td>0.017</td>
<td>0.013</td>
</tr>
<tr>
<td>Histidine</td>
<td>0.015 g</td>
<td>0.024 g</td>
<td>0.0094 g</td>
<td>0.02</td>
<td>0.025</td>
<td>0.0059</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>0.047 g</td>
<td>0.050 g</td>
<td>0.029 g</td>
<td>0.048</td>
<td>0.046</td>
<td>0.015</td>
</tr>
<tr>
<td>Leucine</td>
<td>0.064 g</td>
<td>0.071 g</td>
<td>0.036 g</td>
<td>0.035</td>
<td>0.079</td>
<td>0.016</td>
</tr>
<tr>
<td>Lysine</td>
<td>0.065 g</td>
<td>0.053 g</td>
<td>0.040 g</td>
<td>0.068</td>
<td>0.073</td>
<td>0.016</td>
</tr>
<tr>
<td>Methionine</td>
<td>0.020 g</td>
<td>0.024 g</td>
<td>0.0079 g</td>
<td>0.024</td>
<td>0.028</td>
<td>0.0053</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>0.035 g</td>
<td>0.038 g</td>
<td>0.022 g</td>
<td>0.041</td>
<td>0.044</td>
<td>0.0062</td>
</tr>
<tr>
<td>Proline</td>
<td>–</td>
<td>0.115 g</td>
<td>–</td>
<td>–</td>
<td>0.074</td>
<td>0.011</td>
</tr>
<tr>
<td>Serine</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.050</td>
<td>0.029</td>
</tr>
<tr>
<td>Threonine</td>
<td>0.033 g</td>
<td>0.035 g</td>
<td>0.034 g</td>
<td>0.031</td>
<td>0.038</td>
<td>0.012</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>0.010 g</td>
<td>0.009 g</td>
<td>0.0088 g</td>
<td>0.015</td>
<td>0.011</td>
<td>0.0012</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>0.040 g</td>
<td>0.023 g</td>
<td>0.0060 g</td>
<td>0.071</td>
<td>0.033</td>
<td>0.0055</td>
</tr>
<tr>
<td>Valine</td>
<td>0.048 g</td>
<td>0.056 g</td>
<td>0.034 g</td>
<td>0.063</td>
<td>0.053</td>
<td>0.012</td>
</tr>
<tr>
<td>Riboflavin</td>
<td>50 (\mu g)</td>
<td>5.8 (\mu g)</td>
<td>–</td>
<td>–</td>
<td>19 (\mu g)</td>
<td>0.18 (\mu g)</td>
</tr>
<tr>
<td>Thiamine</td>
<td>100 –</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>Pantothenic Acid</td>
<td>100 2.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Pyridoxin</td>
<td>30 0.06</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>20</td>
<td>2.6</td>
</tr>
<tr>
<td>Biotin</td>
<td>4 0.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.4</td>
<td>0.36</td>
</tr>
<tr>
<td>Folic Acid</td>
<td>20 0.67</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Choline</td>
<td>2000 0.45</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Niacin</td>
<td>400 8</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>279</td>
<td>11</td>
</tr>
<tr>
<td>PABA</td>
<td>24 0.21</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Table A.2: Estimates of cell yield (g g⁻¹) on various elements

<table>
<thead>
<tr>
<th>Elements</th>
<th>Suzuki 85 C brassicae</th>
<th>Blanch 93 E. coli</th>
<th>Haggstrom 75 Bacteria</th>
<th>Mateles 74 Pseudomonas c</th>
<th>Kuhn 79 Bacillus caldopenax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon</td>
<td>2.2–2.2</td>
<td>–</td>
<td>1.9</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>12–13</td>
<td>10</td>
<td>8*</td>
<td>9.1</td>
<td>6</td>
</tr>
<tr>
<td>Sulfur</td>
<td>280–310</td>
<td>160</td>
<td>94</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>53–63</td>
<td>34</td>
<td>36</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>Potassium</td>
<td>59–67</td>
<td>53</td>
<td>60</td>
<td>160</td>
<td>51</td>
</tr>
<tr>
<td>Magnesium</td>
<td>380–430</td>
<td>280</td>
<td>170</td>
<td>130</td>
<td>571</td>
</tr>
<tr>
<td>Sodium</td>
<td>3600–3700</td>
<td>–</td>
<td>260</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Chlorine</td>
<td>–</td>
<td>–</td>
<td>333</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Iron</td>
<td>2400–4500</td>
<td>–</td>
<td>2800</td>
<td>1700</td>
<td>–</td>
</tr>
<tr>
<td>Zinc</td>
<td>5000–6700</td>
<td>–</td>
<td>6000</td>
<td>280000</td>
<td>–</td>
</tr>
<tr>
<td>Manganese</td>
<td>71000–125000</td>
<td>–</td>
<td>17000</td>
<td>770000</td>
<td>–</td>
</tr>
<tr>
<td>Calcium</td>
<td>53000–100000</td>
<td>–</td>
<td>280</td>
<td>830</td>
<td>10000</td>
</tr>
<tr>
<td>Cobalt</td>
<td>25000–45000</td>
<td>–</td>
<td>280000</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Copper</td>
<td>16000–27000</td>
<td>–</td>
<td>28000</td>
<td>240000</td>
<td>–</td>
</tr>
<tr>
<td>Molybdenum</td>
<td>–</td>
<td>–</td>
<td>420000</td>
<td>560000</td>
<td>–</td>
</tr>
<tr>
<td>Boron</td>
<td>–</td>
<td>–</td>
<td>140000</td>
<td>1700000</td>
<td>–</td>
</tr>
</tbody>
</table>

Table A.3: Estimates of cell yield (g dry cells g⁻¹) on various amino acids based on bacterial cell composition

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Yield</th>
<th>Amino acid</th>
<th>Yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>19</td>
<td>Leucine</td>
<td>22</td>
</tr>
<tr>
<td>Arginine</td>
<td>33</td>
<td>Lysine</td>
<td>28</td>
</tr>
<tr>
<td>Asparagine</td>
<td>40</td>
<td>Methionine</td>
<td>63</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>40</td>
<td>Phenylalanine</td>
<td>53</td>
</tr>
<tr>
<td>Cysteine</td>
<td>106</td>
<td>Proline</td>
<td>44</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>37</td>
<td>Serine</td>
<td>45</td>
</tr>
<tr>
<td>Glutamine</td>
<td>37</td>
<td>Threonine</td>
<td>38</td>
</tr>
<tr>
<td>Glycine</td>
<td>16</td>
<td>Tryptophan</td>
<td>172</td>
</tr>
<tr>
<td>Histidine</td>
<td>102</td>
<td>Tyrosine</td>
<td>71</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>33</td>
<td>Valine</td>
<td>23</td>
</tr>
</tbody>
</table>
Table A.4: Estimates of cell yield (kg g⁻¹) for various vitamins based on cell composition (Thompson, 1972)

<table>
<thead>
<tr>
<th>Vitamin</th>
<th><em>Aerobacter aerogenes</em></th>
<th><em>Pseudomonas fluorescens</em></th>
<th><em>Clostridium butylicum</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Thiamine</td>
<td>91</td>
<td>39</td>
<td>11</td>
</tr>
<tr>
<td>Riboflavin</td>
<td>23</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Nicotinic Acid</td>
<td>4.2</td>
<td>4.8</td>
<td>4.0</td>
</tr>
<tr>
<td>Pantothenic Acid</td>
<td>7.1</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Pyridoxin</td>
<td>147</td>
<td>175</td>
<td>161</td>
</tr>
<tr>
<td>Biotin</td>
<td>256</td>
<td>141</td>
<td>–</td>
</tr>
<tr>
<td>Folic Acid</td>
<td>71</td>
<td>114</td>
<td>357</td>
</tr>
<tr>
<td>Inositol</td>
<td>0.71</td>
<td>0.59</td>
<td>1.2</td>
</tr>
</tbody>
</table>
Chapter B

SOURCE CODE

This appendix lists some of the Matlab and GAMS scripts and functions written in the course of this research for the reader’s reference. Where possible, I have cross-referenced each program to the figure in the document it was used to create. All software is copyright (c) Massachusetts Institute of Technology (1997).

B.1 RQ, $k_{L_a}$, and NUR Calculations

B.1.1 aibabias.m

This script created Figure 3-3 showing the TQ/RQ ratio as a function of pH and VVM/$\mu$.

% aibabias. Script to show tq/rq ratio as function of ph,u,and q
%
u = 0.5;
vvmrange = [0.10/u 0.25/u 0.50/u 1.00/u];
phrange = [5.0:0.1:8.0];
phindex = [1:length(phrange)];
vvmindex = [1:length(vvmrange)];
%  
rr = zeros(length(phrange),length(vvmrange));
%  
for phj=phindex;
   for vvmk=vvmindex;
      rr(phj,vvmk)=aiba(phrange(phj),u,vvmrange(vvmk));
   end;
end;
plot(phrange,rr);
grid
xlabel('16\times pH')
sylabel('16\times TQ/RQ')
setext(6.95,0.62,'16\times VVM/\mu =0.2');
setext(7.5,0.76, '16\times 0.5');
setext(7.6,0.84, '16\times 1.0');
setext(7.7,0.89, '16\times 2.0');
printsto --deps2 d:\working\thesis\plots\aibabias\eps
B.1.2 aiba.m

This function is called by aibabias.m.

function rr=aiba(ph,u,vvm);
%
aiba.m Script to generate plots derived from Aiba & Fuerese (1990) equations:
%
k1 = 10^-(-6.3);
k2 = 10^-(-10.25);
P = 1;           % pressure (atm)
alpha = 1;       % out/qin
phi = 0.5;       % vh/vl
ho2 = 0.029;

hco2 = 0.738;

hco2eff = hco2*(1+k1*10^-((ph)+k1*k2*10^-((ph)));
rho = (vvm*60)/u;

sigma = alpha*rho/(phi+ho2*P);
sigmaprime = alpha*rho/(phi+hco2eff*P);

rr = (1+sigma)*sigmaprime/(1+sigmaprime)*sigma);

B.1.3 rqvyield.m

This script created Figure 3-6.

% script to generate a plot of rq vs. yield for various substrates
nsubs = 7;
subcomp = zeros(nsubs,4);

subname = zeros(nsubs,21);
dors = zeros(nsubs);
ash = 0.05;
x = [1 1.8 0.5 0.2];

dorx = 4.06;

subname(1,:)='\texttimes 16\text{glucose}';
subcomp(1,:)=[1 2 1 0];

subname(2,:)='\texttimes 16\text{glycerol}';
subcomp(2,:)=[1 8/3 1 0];

subname(3,:)='\texttimes 16\text{ethanol}';
subcomp(3,:)=[1 3 0.5 0];

subname(4,:)='\texttimes 16\text{methane}';
subcomp(4,:)=[1 4 0 0];

subname(5,:)='\texttimes 16\text{citric acid}';
subcomp(5,:)=[6/6 8/6 7/6 0];

subname(6,:)='\texttimes 16\text{protein}';
subcomp(6,:)=[1 2.03 0.5 0.29];

subname(7,:)='\texttimes 16\text{methanol}';
subcomp(7,:)=[1 4 1 0];

index=0.00:0.01:1.00;
plot([0 1.2],[0 2],'.')
set(gca,'FontSize',16)
set(gca,'FontName','Times')

hold on
for i=1:nsubs
    dors(i)=dor(subcomp(i,:));
    if dors(i)>4.67
yxsm(i)=0.6;
else
  yxsm(i)=0.13*dors(i);
end;

for j=1:length(index)
  srq(i,j) = rcalc(dors(i),dorx,index(j));
  yindex(i,j) = yxsgcalc(index(j),mwc(x),mwc(subcomp(i,:)),ash);
end;

plot(yindex(i,index<yxsm(i)),srq(i,index<yxsm(i)));%  
setext(yxsgcalc(x,yxsm(i),mwc(x),mwc(subcomp(i,:)),ash),...
      rcalc(dors(i),dorx,yxsm(i)),subname(i,:));
end;

xlabel('\16times Y_{-}XS~-' (g g^{-1}));
sxlabel('\16times RQ')
grid
hold off
printsto -deps2 d:\working\thesis\plots\rqvyield.eps

B.1.4 rcalc.m
  This function calculates the expected RQ for a given substrate and biomass degree of reduction and yield on substrate.
function rq=rqc(dors,dorx,yxs)
  rq=(4/dors)*(1-yxs)/(1-yxs*(dorx/dors));

B.1.5 yxsgcalc.m
  This function converts molar cell yield to mass cell yield.
function yxsg=yxsgcalc(yxs,mwx,mws,ash);
yxsg=yxs*(mwx/(1-ash))/mws;

B.1.6 syxssrq.m
  This function calculates the error in the cell yield estimate as a function of the degrees of reduction of the substrate and biomass, the yield, and the error in the RQ.
function syxs= syxssrq(gamma, gammas, yxs, srq);
  %syxssrq.m estimate yield uncertainty based on RQ
  
  rq=rcalc(gamma, gammas, yxs);
  gain=((gamma-gamma)/(4*(gamma/4*rq-1)^2));
  syxs=gain*srq;

B.1.7 rqvyps.m
  This script created Figure 3-7 showing the relationship between RQ and product type and yield.

nprods = 8;
prodcomp = zeros(nprods,4);
prodname = zeros(nprods,21);
dors = zeros(nprods);
ash = 0.05;
glucose = [1 1.8 0.5 0.2];
dorglucose = 4;
%
prodname(1,:)='\16\times Glucose'; prodcomp(1,:)=[1 2 1 0];
prodname(2,:)='\16\times EtOH'; prodcomp(2,:)=[1 3 0.5 0];
prodname(3,:)='\16\times Methane'; prodcomp(3,:)=[1 4 0 0];
prodname(4,:)='\16\times Citric Acid'; prodcomp(4,:)=[1 8/6 7/6 0];
prodname(5,:)='\16\times Protein'; prodcomp(5,:)=[1 2.03 0.5 0.29];
prodname(6,:)='\16\times MeOH'; prodcomp(6,:)=[1 4 1 0];
prodname(7,:)='\16\times Lysine'; prodcomp(7,:)=[1 14/6 2/6 2/6];
%
index=0.00:0.01:1.00;
plot([0 1.2], [0 2],('.'))
hold on
for i=2:nprods
  dors(i)=dor(prodcomp(i,:));
yxsmax(i)=0.4; %arbitrary
for j=1:length(index)
  sqg(i,j)=rncalc(dorglucose,dors(i),index(j));
  yindex(i,j)=yxsgcalc(index(j), mwc醛(prodcomp(i,:)), mwc醛(glucose),0);
end;
plot(yindex(i,index<yxsmax(i)), sqg(i,index<yxsmax(i)));
axis([0 0.71 0.8 1.6]);
if rncalc(dorglucose,dors(i),yxsmax(i))>1.7,
  stext(max(yindex(i,index<yxsmax(i))),-0.05,...
    1.55,...
    prodname(i,:));
else
  sttext(max(yindex(i,index<yxsmax(i))),...
    rncalc(dorglucose,dors(i),yxsmax(i)),...
    prodname(i,:));
end
end;
sxlabel('\16\times Y_{PS}'' (g g^{-1})');
sylabel('\16\times RQ')
grid
hold off
pause
printsto -deps2 d:\work\thesis\plot\r\sympy.png
B.1.8  yxserr.m

This script created Figure 3-8.

% yxserr.m script to look at yield estimation sensitivity
%
gammas = 4.0;
gammaxrange = [4.06 4.2 4.3 4.8];
yxsrange = [0:0.1:0.6];
%
for i=1:length(gammaxrange)
  for j=1:length(yxsrange)
    syxs(i,j)=syxssrq(gammaxrange(i),gammas,yxsrange(j),1);
  end;
end;
%
plot(yxsrange,syxs(1,:),yxsrange,syxs(2,:),yxsrange,syxs(3,:),yxsrange,syxs(4,:))
xlabel('
16\times Yxs (mol mol^{-1})')
sxlabel('
16\times \sigma_{Yxs} / \sigma_{RQ}')

stext(0.1,60,'\text{16\times \gamma_{x}=4.0\over}}')
stext(0.1,20,'\text{16\times 4.2}')
stext(0.1,12,'\text{16\times 4.3}')
stext(0.1,5,'\text{16\times 4.8}')

pause

printsto -deps2 d:\working\thesis\plots\yxserr.eps

B.1.9  tqerr.m

This script created Figure 3-9.

% tqerr.m script to generate s_tq=f(our/vvm) plot
fgin = 1;
otr = 100;
tq = 1;
sg = [0.1 0.01 0.001]/100;
ourovervvm = [0.01:0.1:10 [10:3:50]];
%
for i=1:length(ourovervvm)
  for j=1:length(sg)
    s(i,j)=tquncert(otr/ourovervvm(i),otr,tq,sg(j));
  end;
end;
%
index =s(:,1)<1;
index2=s(:,2)<1;
index3=s(:,3)<1;
%
semilogy(ourovervvm(index), s(index,1),...
  ourovervvm(index2),s(index2,2),...
  ourovervvm(index3),s(index3,3));
%
161
B.1.10 rqerrbic.m

This script created Figure 3-10.

%rqerrbic.m script to generate plot of rqerr=f(ph,sph)
%
klaco2 = 100000*30;
phs = [5.5:0.1:7.5];
syoutco2 = 0; %0.001/100;
sphs = [0.01 0.02 0.04];
sampleperhr = 30;
%
ptot = 1.5;
hco2 = 0.039;
pK1 = 6.3;
%
sptot = 0.0; %0.1;
shco2 = 0.0; %0.01;
sq = 0.0; %100;
spK1 = 0.0; %0.05;
slaco2 = 0.0; %50;
%
q = 2453;
otr = 100;
youtco2 = otr/q;
%
for i=1:length(phs)
    for j=1:length(sphs);
        [s(i,j) ctot(i,j)]=ctotacc(youtco2,ptot,...
            hco2,q,...
            pK1,klaco2,...
            phs(i),syoutco2,...
            sptot,shco2,sq,...
            spK1,slaco2,sphs(j));
        se(i,j)=(youtco2*phtot/hco2)*10^(-pK1)*10^phs(i)*log(10)*sphs(j);
    end;
end;
%
% RQ=(CPR+dctot/dt)/otr=cpr/otr +dctot/dt/otr
sdiff = s*sqrt(2)*sampleperhr;
sRQ = sdif/otr;
seRQ = se*sqrt(2)*sampleperhr/otr;
%
semilogy(phs,sRQ(:,1),phs,sRQ(:,2),phs,sRQ(:,3))
sxlabel(’\(\times\) pH’)
sylabel(’\(\sigma_{RQ}\)’)  
sext(7.2,0.2,'\(\times\) 0.01');
sext(7.2,0.38,'\(\times\) 0.02');
sext(7.05,0.7,'\(\times\) \(\sigma_{pH}=0.04\)’);
printsto -deps2 d:\working\thesis\plots\rqerrbic.eps

B.1.11 optaer.m
This script created Figure 3-11.

% Script to look at effect of operating conditions and measurement accuracy on uncertainty in RQ.

% Ranges of variables to look at
phs = [5.00 : 0.01 : 8.00];
vvms = [0.010: 0.005 : 1.000];
orss = [1 10 100];
%
% Measurement Accuracies
sph = 0.01;
sy  = 0.0001;
%
aopt = zeros(length(orss),length(phs));
vopt = aopt;
%
for k=1:length(orss)
z = zeros(length(vvms),length(phs));
   for i=1:length(phs)
      for j=1:length(vvms)
         z(j,i)=uncert(phs(i),vvms(j),sph,sy,orss(k));
      end;
   end;
   for i=1:columns(z)
      aopt(k,i)=min(z(:,i));
      vopt(k,i)=vvms(z(:,i)=aopt(k,i));
   end;
end;
%
figure(1)
plot(phs,vopt(1,:),phs,vopt(2,:),phs,vopt(3,:))
sxlabel(’\(\times\) pH’)
sylabel(’\(\times\) VVM for \(\sigma_{RQ}\)={\it min}’)
sext(7,0.05,'\(\times\) OUR=1')
sext(6.5,0.3,'\(\times\) OUR=10')

163
B.1.12 bsr2sig.m

This script created Figure 5-7.

```matlab
% bsr2sig.m
%
sy = 0.00005;
sph = 0.005;
%
[data mids]=loadbsr('/mit/jjprior/matlab/data/bsr2.dat');
gasindex;
%
onevec= ones(rows(data),1);
unc = zeros(rows(data),1);
%
% fixing ph and ms calibrations
data(:,PTOT) = onevec*1.2;
data(:,PH) = data(:,PH)+onevec*0.7;
data(:,YINO2) = data(:,YINO2).*onevec*0.949;
data(:,YOUTO2) = data(:,YOUTO2).*onevec*0.949*1.003;
data(:,YINCO2) = data(:,YINCO2).*onevec*0.769;
data(:,YOUTCO2) = data(:,YOUTCO2).*onevec*0.769*1.003;
%
for i=4:rows(data)-3
    [unc(i) fp(i)]=uncert(median(data(i-3:i+3,PH)),...
data(i,FGIN)/10,sph,sy,m[median(data(i-3:i+3,OTR))];
end;
%
xguesses=gasevec(data,mids);
whitebg('w')
figure(1)
plot(xguesses(:,TIME),xguesses(:,RQ),'b',...%0.05+1*unc,'r-',...
xguesses(:,TIME),onevec*1.05-1*unc,'r-')
axis([0 12 0 1.5]);
grid;
sxlabel('\text{Time(h)}')
sylabel('\text{Calculated RQ}')
set(3.5,1.3,'\times\text{Sigma}')
set(4, 0.3,'\times\text{RQ-Sigma}')
subw=zeros(rows(xguesses),1);
meanw=subw;
subset=[20:rows(xguesses)-20];
for i=subset
```

164
$stw(i)=std(x\text{guesses}(i-19:i+19,RQ));$
$meanw(i)=mean(x\text{guesses}(i-19:i+19,RQ));$
end;

printsto -depst2 d:\working\thesis\plots\bbr2sig.eps
%
figure(2);
plot(x\text{guesses}(:,TIME),x\text{guesses}(:,RQ),'b',... 
     xguesses(:,TIME),onevec*1.05+unc,'r-',... 
     xguesses(:,TIME),onevec*1.05-unc,'r-',... 
     x\text{guesses(subset,TIME)},meanw(subset)+stw(subset),'b--',... 
     x\text{guesses(subset,TIME)},meanw(subset)-stw(subset),'b--');
axis([0 12 -1 3]);
sxlabel('/20\times Time(h)')
sylabel('/20\times RQ')
printsto -depst2 d:\working\thesis\plots\bbr2sig.es
%
figure(3)
plot(x\text{guesses}(:,TIME),x\text{guesses}(:,RQ),'r',... 
     x\text{guesses}(:,TIME),x\text{guesses}(:,TQ),'b'.)
ggrid
xlabel('Time (h)')
ylabel('TQ and RQ')

B.1.13 nur3.gms
GAMS program to find optimum gas feed.

* NUR3.GMS
* GAMS program to calculate the optimum gas mixture for
* maximizing NUR calculation accuracy. It was written
* to generate a graphical representation of the accuracy
* vs argon concentration.
* The original version of the program had an incorrect
* equation for the error propagation, using standard
* deviations where it should have used variances.
* J. Prior 10/96
*
$TITLE NUR ACCURACY MODEL
SETS
   I gases /N2, AR, O2, CO2/
   J streams /IN, OUT/
   K time steps /INIT,FINAL/

TABLE UPTAKES(I,K)
* Sets the initial and final gas uptake rates

    INIT      FINAL
    N2        0.04     2
    AR        0        0
    O2        2        350

165
PARAMETER MSACCEST(I) ms accuracy (fraction x100)
* Gas composition is represented as volume percent
/W2 0.005
AR  0.005
O2  0.005
CO2 0.005;
SCALAR QACC airflow and vl accuracy est /0.04/;
SCALAR VL liquid volume (L) /1/;
SCALAR GASCONV mmol min per (L h) by 100 /24.53/;
SCALAR KLALNA ln(kla pre exponent) /-15.45/;
SCALAR KLAB kla exponent /3.44/;
VARIABLES
Y(I,J,K) gas composition
QGAS(K) specific gas flow (mmol per l per h)
FGAS(K) inlet gas flow (l per min)
UR(I,K) uptake rate specified
URACC(I,K) uptake rate accuracy
MSACC(I) mass spec accuracy
DELTA(I,K) normalized changed in conc of i
KLA(K) mass transfer coefficient
P(I,K) partial pressure of gases
PRES(K) reactor pressure
Z obj function;
POSITIVE VARIABLES Y,QGAS,F GAS,PRES,P,N,KLA;

MSACC.LO(I) = MSACCEST(I);
MSACC.UP(I) = MSACCEST(I);
UR.LO(I,K) = UPTAKES(I,K)-ABS(UPTAKES(I,K)*0.01);
UR.UP(I,K) = UPTAKES(I,K)+ABS(UPTAKES(I,K)*0.01);
URACC.LO(I,K) = 0.01;
URACC.L(I,K) = 10;
PRES.LO(K) = 1.0;
PRES.UP(K) = 1.5;
P.LO(I,K) = 0.0;
P.UP(I,K) = 1.5;
FGAS.LO(K) = 0.05;
FGAS.UP(K) = 1.0;
P.LO("O2",K) = 0.04;
P.UP("O2",K) = 0.04;
Y.LO(I,"OUT",K) = 0.01;
Y.UP(I,J,K) = 100;
P.LO("N2",K) = 0.66;
Y.UP("CO2","IN",K) = 0.04;
KLA.LO(K) = EXP(KLALNA) * 100**KLAB;
KLA.UP(K) = EXP(KLALNA) * 800**KLAB;

UR.L(I,K) = UPTAKES(I,K);
KLA.L(K) = KLA.LO(K);
FGAS.L(K) = FGAS.LO(K);
QGAS.L(K) = FGAS.L(K)*GASCONV/VL;
Y.L("N2", "IN", "INIT") = 78.0;
Y.L("N2", "OUT", "INIT") = 77.9;
Y.L("O2", "IN", "INIT") = 21.0;
Y.L("O2", "OUT", "INIT") = 20.9;
Y.L("CO2", "IN", "INIT") = 0.04;
Y.L("CO2", "OUT", "INIT") = 0.1;
Y.L("AR", "IN", "INIT") = 1.0;
Y.L("AR", "OUT", "INIT") = 1.0;
Y.L("N2", "IN", "FINAL") = 78.0;
Y.L("N2", "OUT", "FINAL") = 77.6;
Y.L("O2", "IN", "FINAL") = 21.0;
Y.L("O2", "OUT", "FINAL") = 20.0;
Y.L("CO2", "IN", "FINAL") = 0.04;
Y.L("CO2", "OUT", "FINAL") = 1;
Y.L("AR", "IN", "FINAL") = 1.0;
Y.L("AR", "OUT", "FINAL") = 1.0;

EQUATIONS
OBJ Objective function: minimum uncertainty
GASTOT(J,K) sum of gas mole fractions = 1
URCALC(I,K) gas exchange rate calculation
GFLOWCALC(K) gas flow rate conversion
USEAIR(K) just supplement with O2
PCALC(I,K) liquid partial pressure estimate
ONEGAS(I) one gas used throughout
URACCCALC(I,K) uptake rate accuracy estimate;

OBJ.. Z =E= + uracc("N2", "FINAL")/UR("N2", "FINAL")
      + 0.01*uracc("N2", "INIT")/UR("N2", "FINAL");
GASTOT(J,K).. SUM(I,Y(I,J,K)) - 100 =E= 0;
URCALC(I,K).. QGAS(K)*
      (Y(I, "IN", K) - Y(I, "OUT", K)*
      (Y("AR", "IN", K)/Y("AR", "OUT", K)))
      =E= UR(I, K);
URACCCALC(I,K).. URACC(I,K) -
      SQRT(
      (QACC*QGAS(K))**2
      *
      POWER((
      Y(I, "IN", K) - Y(I, "OUT", K)*
      })

167
(Y("AR", "IN", K)/Y("AR", "OUT", K))
), 2)
+
(MSACC(I)**2 +
(Y("AR", "IN", K)/Y("AR", "OUT", K))**2 * MSACC(I)**2 +
Y(I, "OUT", K)**2 * Y("AR", "OUT", K)**(-4) *
(Y("AR", "IN", K)**2 * MSACC("AR")**2 +
Y("AR", "OUT", K)**2 * MSACC("AR")**2)
)

* QGAS(K)**2

=E=0;
P(CLC(I, K)) = P(I, "OUT", K)*PRES(K)/100.0 + UR(I, K)/KLA(K) =E=0;
GFCALC(K) = FGAS(K)*gass conv/VEL =E=0;
*
USEAIR(K) = 3.71*Y("O2", "IN", K) - Y("N2", "IN", K) =E=0;
ONEGAS(I) = Y(I, "IN", "INIT") - Y(I, "IN", "FINAL") =E=0;

MODEL NUROPT /ALL/;

OPTION OPTCR = 0.002;
OPTION ITERLIM = 10000;
OPTION RESLIM = 10000;

SOLVE NUROPT USING NLP MINIMIZING Z;

PARAMETER REP SUMMARY REPORT;
REP("AR", "ACC") = Z.L;
REP("AR", "F") = FGAS.L("FINAL");
REP("AR", "YO2") = Y.L("O2", "IN", "FINAL");
REP("AR", "OUR") = UR.L("O2", "FINAL");
REP("AR", "YAR") = Y.L("AR", "IN", "FINAL");

UR.L("O2", "FINAL") = 300;
UR.L("CO2", "FINAL") = -300;
UR.UP("O2", "FINAL") = 300;
UR.UP("CO2", "FINAL") = -300;
UR.L("O2", "FINAL") = 300;
UR.L("CO2", "FINAL") = -300;
SOLVE NUROPT USING NLP MINIMIZING Z;
REP("AR1", "ACC") = Z.L;
REP("AR1", "F") = FGAS.L("FINAL");
REP("AR1", "YO2") = Y.L("O2", "IN", "FINAL");
REP("AR1", "OUR") = UR.L("O2", "FINAL");
REP("AR1", "YAR") = Y.L("AR", "IN", "FINAL");
UR.LO("O2", "FINAL")=250;
UR.LO("CO2", "FINAL")=-250;
UR.UP("O2", "FINAL")=250;
UR.UP("CO2", "FINAL")=-250;
UR.L("O2", "FINAL")=250;
UR.L("CO2", "FINAL")=-250;

SOLVE NROOPT USING NLP MINIMIZING Z;
REP("AR2.5", "ACC")= Z.L;
REP("AR2.5", "F") = FGAS.L("FINAL");
REP("AR2.5", "YQ2") = Y.L("O2", "IN", "FINAL");
REP("AR2.5", "OUR") = UR.L("O2", "FINAL");
REP("AR2.5", "YAR") = Y.L("AR", "IN", "FINAL");

UR.LO("O2", "FINAL")=200;
UR.LO("CO2", "FINAL")=-200;
UR.UP("O2", "FINAL")=200;
UR.UP("CO2", "FINAL")=-200;
UR.L("O2", "FINAL")=200;
UR.L("CO2", "FINAL")=-120;

SOLVE NROOPT USING NLP MINIMIZING Z;
REP("AR5", "ACC")= Z.L;
REP("AR5", "F") = FGAS.L("FINAL");
REP("AR5", "YQ2") = Y.L("O2", "IN", "FINAL");
REP("AR5", "OUR") = UR.L("O2", "FINAL");
REP("AR5", "YAR") = Y.L("AR", "IN", "FINAL");

UR.LO("O2", "FINAL")=100;
UR.LO("CO2", "FINAL")=-100;
UR.UP("O2", "FINAL")=100;
UR.UP("CO2", "FINAL")=-100;
UR.L("O2", "FINAL")=100;
UR.L("CO2", "FINAL")=-100;

SOLVE NROOPT USING NLP MINIMIZING Z;
REP("AR10", "ACC")= Z.L;
REP("AR10", "F") = FGAS.L("FINAL");
REP("AR10", "YQ2") = Y.L("O2", "IN", "FINAL");
REP("AR10", "OUR") = UR.L("O2", "FINAL");
REP("AR10", "YAR") = Y.L("AR", "IN", "FINAL");

UR.LO("O2", "FINAL")=50;
UR.LO("CO2", "FINAL")=-50;
UR.UP("O2", "FINAL")=50;
UR.UP("CO2", "FINAL")=-50;
UR.L("O2", "FINAL")=50;
UR.L("CO2", "FINAL")=-50;

DISPLAY REP;
B.2 Data Reconciliation Programs

B.2.1 analyze.m

Matlab script demonstrating the data reconciliation algorithm.

```matlab
%analyze.m
disp('Script to Demonstrate Data Reconciliation methodology');
global SCALE STEST MODEL_LEAK ARUN
MODEL_LEAK=1;
STEST=1;
ARUN=2033;
%
%(1) Process Description
%(2) Variable Classification
%(3) Data Rectification
%(4) Data Consistency Analysis
%(5) Gross Error Detection
%(6) Results Interpretation
%%%%%%%%%%%%%%%%

//The 'stage' variable allows the program to be stopped and restarted at
//intermediate points in the algorithm without having to repeat earlier
//calculations.
if ~exist('stage')
    stage=0;
end;
%
%
if stage==0
    [data mids]=loadstef('/mit/jjprior/Private/stefan/hepdata/data7.dat');
    [data mids]=loaddbsr('/mit/jjprior/matlab/data/bsr2.dat');
    [data mids]=loadamy('/mit/jjprior/matlab/data/a2033.dat');
    [fermtime data mids]=loadjust(ARUN);
    stage=stage+1;
end;
%
global xhist chihist conhist
global jp_eps
jp_eps=1e-9; % important setting for variable classification
global JP_DEBUG
if ~exist('JP_DEBUG')
    JP_DEBUG=0; % 0 or 1=0; 0 or 1
end;
JP_DEBUG=1;
global POS_VARS
gas_model=1;
stoic_model=2;
nh3_model=3;
stefan=4;
amylin=5;
```

170
if stage==1
    disp('Process Description');

eval = eval neoliberal(15060); % Use eval neoliberal for stage 1

if use_model==gas_model
    disp('gastrain model');
    POS_VARS=0;
    gasindex;
    load rsge2 %get some consistent data
    [x, midx, fun, xvlb, xvub] = gasmeas(10);
    uncert = gasuncer(x);
    x = C(100, :);
    lfun = 'gaslabel';
    estfun = 'gasest';
    one_drop = 1;
    midx(DISO2) = 1;
    midx(OD) = 0;
end;

if use_model==stefan
    disp('gastrain model (stefan)');
    if ~exist('POS_VARS') POS_VARS=0; end;
    gasindex;
    rhofeed = 1.25;
    caddad = sum(data(:, GLYADDED));
else
    global mws
    mws = 29.66; %glycerol C1 H5/3 O1 =12+5/3+16=29.66
    stage = 1;
    [x, midx, fun, xvlb, xvub] = gasmeas(10);
    x = C(100, :);
    estfun = 'gasest';
    x = data(250, :);
midx=mids(250,:);
% x=data(430,:);
% midx=mids(430,:);
% midx(DISO2)=1;
% x=data(64,:);
% midx=mids(64,:);
lfun='gaslabel';
one_drop=1;
if estfun=[]
xguess=feval(estfun,x,midx);
else
  disp('warning: no initial estimate function');
xguess=x;
end;
if JP_DEBUG
gastest(xguess,midx,uncert);
end;
end;

%---------------------

if use_model==amylin
  disp('justgas model (amylin)');
  if ~exist('POS_VARS')  POS_VARS=0;  end;
  justgidx;
  [x,midx,fun,xv1b,xvub]=gasmeas(10);
  fun  = 'justgas';
  estfun='justest';
  lfun  = 'justlabl';
  one_drop=1;
  x=data(801,:);
  midx=mids(801,:);
  x=mean(data(790:810,:));
  midx=mids(801,:);
  x=mean(data(340:360,:));
  midx=mids(350,:);
  x=data(250,:);
  midx=mids(250,:);
  if estfun=[]
    xguess=feval(estfun,x,midx);
  else
    disp('warning: no initial estimate function');
    xguess=x;
  end;
  uncert=justunce(xguess);  % need to make generic
  if JP_DEBUG justtest(xguess,midx,uncert);end;
end;
%---------------------
if use_model==stoic_model
    disp('Using stoic model');
    estfun=[];
    [x,midx,uncert,fun]=stoimeas;
    lfun='stoilabs';
    end;
if use_model==nh3_model
    [x,midx,uncert,fun]=nh3synm;
    lfun='nh3label';
    end;
stage=stage+1;
end;

%-----------------------------------------------------
%
%-----------------------------------------------------

%(2) Variable Classification
%-----------------------------------------------------
%
% Variables are classified as measured, redundant,
% observable, or unobservable
% The redundant variables are then sub-classified into just_redundant
% or diagnosable and the observables are classified as tenously
% or robustly observable
%
if stage==2
    TEST=1;
    SCALE=max(abs(dfdx(fun,xguess)));
    TEST=0;
    disp('Variable Classification')
    Time0=cputime;
    disp(' Observability classification')
    disp(' put xguess in here?')
    [xclass k]=classify(fun,xguess,midx);
    otime=cputime-Time0;
    disp(sprintf(']%i degrees of redundancy present'))
    disp(sprintf(']%.1f sec',otime));
    disp(' Diagnosability classification')
    disp(' Variables are further classified as uniquely "diagnosable".
% Should identify groups of "causes". Need to look at RQ
% issue
    [allclass same_as]=diagclas(fun,xguess,midx,xclass); %xguess use important
dtime=cputime-Time0-otime;
    disp(sprintf(']%.1f sec',dtime));
stage=stage+1;
end;
%-----------------------------------------------------
if stage==3
%disp('generating fault scenarios to test');
%[sidx]=senario(1,fun,x,uncert,midx,xclass)
%[sidx,saclass]=senario(1,fun,x,uncert,midx,xclass,same_as)
stage=stage+1;
end;
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%(3) Data Rectification
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% The measurements are adjusted to fit the process constraints
% while minimizing the weighted adjustments. Interpretation
% of the adjustments will depend on the degrees of freedom of
% the process.
%
if stage==4
disp('Data Rectification')

[xfit,optionsout]=recon(x,xguess,uncert,midx,fun,xvlb,xvub,10);
% [xfit,optionsout]=recon(x,xguess,uncert,midx,fun,xvlb,xvub,1);
rtime=cputime-Time0-dtime-otime;
disp(sprintf('%4.1f sec',rtime));
    disp(sprintf('%i evaluations',optionsout(10)));
stage=stage+1;
end;
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%(4) Data Consistency Analysis
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% The magnitude of the adjustments are tested against the
% chi-square distribution to test the hypothesis that the
% adjustments are the result of random error. 'fun' is passed
% to the routine only to allow the number of independent
% relations in the model to be determined.
%
% [what is effect of unobs on dof?]
if stage==5;
disp('Data consistency analysis')
[p,chi_sq]= consist(x,xfit,uncert,midx,fun);
ctime=cputime-Time0-dtime-otime;
disp(sprintf('pvalue=%5.2f, chisqr=%5.2f',p,chi_sq))
disp(sprintf('%4.1f sec',ctime));
stage=stage+1;
end;
%(5) Gross Error Diagnosis

% If the p-value is above a set threshold, the data is
% suspected of containing a gross error an effort is made
% to determine if a particular sensor can be identified
% as the cause.

% In order for a measurement to be "diagnosable" it must be
% redundant and upon deletion, not reduce any other redundant
% measurement to "measurable" (this is a theory, not proven yet).

% If the process description has
% redundant(xclass=2) measurements and to determine if deleting
% 1 or more of them results in the remaining data-set
% becoming significantly more consistent.

if stage==6;
disp('Gross error diagnosis')
ptresh=-0.1;
if p>=ptresh % comment out to run get anyway for now
    one_drop=1;
    if p<0.999 % fit is probably better than guess
        [pvalues,xfits,chi_sqrs] = gedetect(fun,x,xfit,uncert,midx,xclass,same_as,xvlb,xvub,10);
    else
        [pvalues,xfits,chi_sqrs] = gedetect(fun,x,xguess,uncert,midx,xclass,same_as,xvlb,xvub);
    end
    % [pvalues,xfits,chi_sqrs] = ged2(fun,x,xfit,uncert, sidx);
else
    one_drop=0;
end
gtime=cputime-Time0-ctime-dtime-rtimexotime;
disp(sprintf(' %4.1f sec',gtime));
stage=stage+1;
end;
%
%--------------------------------------------------------------------------
%(6) Results Interpretation and Presentation
%--------------------------------------------------------------------------
save runhome
if stage==7
disp('Results interpretation');
areport;
end;
end;
B.2.2 classify.m

function to classify variables.

function [xclass, k]=classify(fun,x,midx)

%classify variable observability and redundancy using SVD
%function xclass=classify(fun,x,midx)
%
% Arguments: fun function returns vector of constraint violations
% x system variable measurements and estimates
% midx measurement index (1=measured, 0=unmeasured)
%
% Results: xclass vector summary (see clascode)
%

%-------------------------------------------------------------------------------

global jp_eps
if isempty(jp_eps), error('jp_eps not set'); end;
if jp_eps=1e-9, disp('Warning: jp_eps <> 1e-9'); end;

%-------------------------------------------------------------------------------

n = length(midx); % number of variables
m = sum(midx); % number of measurements
R = redund(fun,x,midx); % redundancy matrix
ro = R; %Save copy of orig R
Fx = dfdx(fun,x); % Estimate Jacobianof F(x)=0

c = n-m; %number of unmeasured

% using rref was found more stable at one point in research; but slow
R = rref(R,jp_eps); % row echelon reduction of R
k = rank(R,jp_eps); % degree of redundancy
e = rank(Fx,jp_eps); % independent equations
ea = rows(Fx);

if ea>e, error(sprintf('i redundant equations in model',ea-e)); end;
%
%ko=rank(ro,jp_eps);
%modsum;
%---test effect of row echelon-----------------------------------------------
% Need to develop a way to eval inconsistent constraints

if k=ko
% disp(sprintf(' WARNING:rank of rref R (%4i) and R (%4i) differ',k,ko));
% gaslabel; %gas train specific function
% xlabel(midx==1,:)
%end
%-------------------------------------------------------------------------------

if m>0
meas = sum(abs(R))<jp_eps; % Vars with 0 columns not redund.
else
meas = [];
end;

if e-k==c, % all vars obs if no more redund.

if c>0, % in R than expected (e-c=k)
obs(1:c)=ones(c,1);
else
    obs=[];
end
else % If not, need to determine which
    nullspace=mynull(Fx(:,~midx),jp_eps); %find nullspace using svd
    if nullspace==[]
        disp('Classification Error')
        disp('nullspace=[])'
        cond(Fx(:,~midx))
        disp(sprintf('rank of Fx is %i at eps', rank(Fx(:,:),eps)))
        disp(sprintf('rank of Fx is %i at eps', rank(Fx(:,~midx),eps)))
        disp(sprintf('rank of Fx is %i at jp_eps',rank(Fx(:,~midx),jp_eps)))
        disp(sprintf('rank of Fx is %i at 1e-5', rank(Fx(:,~midx),1e-5)))
    end;
    if columns(nullspace)>1, % which constitute unobservable
        obs=sum(abs(nullspace'))<jp_eps; % lin. combinations of vars.
    else % If zero in all cols then obs.
        obs=abs(nullspace)<jp_eps;
    end
end
xclass = zeros(n,1); %summarize classification results
classcode; % load classification codes
if m>0
    xclass(midx) = meas'*MEASURED +(1-meas')*REDUNDANT; % in a vector
end;
xclass(~midx)=obs*OBSERVABLE +(1-obs)*NOT_OBSERVABLE;
if rows(xclass)<n,
    disp('classification algorithm error')
    disp(sprintf('n=%i variables', n))
    disp(sprintf('m= %i measurements', m ))
    disp(sprintf('c= %i calculated ',c))
    disp(sprintf('e= %i equations', e))
    disp(sprintf('k= %i deg of red', k ))
    disp(sprintf('e-k-c= %i ', e-k-c ))
    disp('Observability')
    if k==e disp(' none'); end;
    if k==(e-c) disp(' all');end;
    if (e>k)&(k>(e-c)) disp('some');end;
    disp('Redundancy');
    if e==0 disp(' none');end;
    if k==m disp(' all'); end;
    if (0<k)&(k<m) disp('some');end;
    error('Notice:classification algorithm failure')
end;
if 1==0 % commented out

177
ddisp(sprintf('rank of Fx is %i at eps',rank(Fx(:,,:),eps)))
disp(sprintf('rank of Fxc is %i at eps',rank(Fx(:,~midx),eps)))
disp(sprintf('rank of Fxc is %i at jp_eps',rank(Fx(:,~midx),jp_eps)))
disp(sprintf('rank of Fxc is %i at 1e-5',rank(Fx(:,~midx),1e-5)))
disp(sprintf('n=%i variables', n))
disp(sprintf('m= %i measurements', m ))
disp(sprintf('c= %i calculated ',c))
disp(sprintf('e= %i equations', e))
disp(sprintf('k= %i deg of red', k ))
disp(sprintf('e-k-c= %i ', e-k-c ))
disp('Observability')
if k==e    disp(' none');end;
if k==(e-c) disp(' all');end;
if (e>k)&&(k>(e-c)) disp(' some');end;
disp('Redundancy');
if k==0    disp(' none');end;
if k==m    disp(' all'); end;
if (0<k)&&(k<m) disp(' some');end;
pause;
end;

B.2.3  redund.m
function r=reund(fun,x,midx)
  %reund    redundancy matrix calculated based on model and measurements
  %
  % r=reund(fun,x,midx)
  %
global jp_eps  
Fx=dfdx(fun,x);
Ec=Fx(:,~midx);
Em=Fx(:,midx);
if length(Ec)>0,
  r=Em - Ec*pinv(Ec,jp_eps)*Em;
else
  r=Em;
end

B.2.4  recon.m
  function to reconcile data.
function  [xcalc, options] = recon(x,xguess,uncert,midx,fun,xvlb,xvub,factor)
  %recon    function to do reconciliation using optimization
  %
global C_COUNT POS_VARS JP_DEBUG SCALE TEST
C_COUNT=0;
if POS_VARS==1
  vlb=xvlb;
vub=xvub;
else

178
vlb=[]; 
vub=[]; 
end; 
%
GRADFUN='gfun';
%----------------------------------------------------------
% constr options:
% OPTIONS(1)-Display parameter (Default: 0). 1 displays some results
% OPTIONS(2)-Termination tolerance for X. (Default: 1e-4).
% OPTIONS(3)-Termination tolerance on F. (Default: 1e-4).
% OPTIONS(4)-Termination criterion on constraint violation. (Default: 1e-6)
% OPTIONS(5)-Algorithm: Strategy: Not always used.
% OPTIONS(6)-Algorithm: Optimizer: Not always used.
% OPTIONS(7)-Algorithm: Line Search Algorithm. (Default 0)
% OPTIONS(8)-Function value. (Lambda in goal attainment.)
% OPTIONS(9)-Set to 1 if you want to check user-supplied gradients
% OPTIONS(10)-Number of Function and Constraint Evaluations.
% OPTIONS(11)-Number of Function Gradient Evaluations.
% OPTIONS(12)-Number of Constraint Evaluations
% OPTIONS(13)-Number of equality constraints.
% OPTIONS(14)-Maximum number of iterations. (Default 100*no. of variables)
% OPTIONS(15)-Used in goal attainment for special objectives.
% OPTIONS(16)-Minimum change in variables for finite difference gradients.
% OPTIONS(17)-Maximum change in variables for finite difference gradients.
% OPTIONS(18)-Step length. (Default 1 or less).
%
options=zeros(18,1); 
options(1)=JP_DEBUG; % 0 or 1 
options(2)=1e-4; %Termination tolerance for X. (Default: 1e-4) 
options(3)=1e-4; %Termination tolerance on F. (Default: 1e-4) 
options(4)=1e-6; %Termination criterion on constraint violation. (Default: 1e-6) 
options(9)=0; % i = check derivs 
options(13)=length(feval(fun,xguess)); % number of constraints 
options(14)=500; 
R=diag(uncert(midx).^2); % need to add cov terms, check missing 
pinvR=inv(R); % pinv caused problems, 
STEST=1; 
SCALE=max(abs(dfdx(fun,xguess)' )); 
STEST=0; 
while factor>1 
  disp(sprintf('Uncertainties relaxed by factor of %i\n',factor)) 
  [xcalc,options]=myconstr('objfcn',xguess,options,vlb,vub,... 
  GRADFUN,fun,x,midx,pinvR./factor); 
  xguess=xcalc; 
  factor=factor/10; 
end; 
[xcalc,options]=myconstr('objfcn',xguess,options,vlb,vub,GRADFUN,fun,x,midx,pinvR);
B.2.5 consist.m

function to assess data consistency.

function [p,chi_sqr]=consist(x,xfit,uncert,midx,fun)

%consist calculate p-value for rejection null hypothesis of no gross errors
%
invR = diag(uncert(midx))^-2; % Covariance matrix
% need to deal with nondiag
chi_sqr = (xfit(midx)-x(midx))'*invR*(xfit(midx)-x(midx));
p = chi2cdf(chi_sqr,t dof(fun,xfit,midx));

B.2.6 gedetect.m

function to detect gross errors.

function [pvalues,xfit,chi_sqr] = gedetect(fun,x,xguess,uncert,midx,xclass,same_as,xv lb,xv

% Gross Error Detection
%
global jp_eps
classcode;
%
xfit = zeros(length(midx),length(midx));
pvalues = zeros(length(midx),1);
chi_sqr = zeros(length(midx),1);
%
for i=1:length(midx),
tidx = midx;
if (xclass(i)==REDUNDANT)||(same_as(i)==i), % only test 1st red.
tidx(i)=0;
[xfit(:,i),optionsout]=recon(x,xguess,uncert,tidx,fun,xv lb,xvub,factor);
%
% calculate p-value for rejection null hypothesis of no gross errors
R = diag(uncert(tidx))^-2; % Covariance matrix
% need to deal with nondiag
% and sig inversion
chi_sqr(i) = (xfit(tidx,i)-x(tidx))'*pinv(R)*(xfit(tidx,i)-x(tidx));
disp(sprintf('%%i evaluations checking var %i: X2=%10.3f',optionsout(i),i,chi_sqr(i)));
% Need to deal with unobservable entries effect on t dof faster
% than currently done
% should test for independent equations
eqs = length(feval(fun,x));
vars = length(x);
meas = length(x(tidx));
classcode;
unobs = sum(xclass==NOT_OBSERVABLE); % assumes we wouldn’t delete
% a variable that would cause more variables to be unobservable
dof(i) = eqs+meas-vars+unobs;
pvalues(i) = chi2cdf(chi_sqr(i),dof(i));
% if pvalues(i)<p best
% xguess=xfit(:,i);
% pbest=pvalues(i);
% end
end

B.2.7 gasest.m

Function to generate initial guess vector for reconciliation.

function xguess = gasest(x,midx)
%gasest Constraint Equations for gas train reconciliation
global JP_DEBUG
%load index to variable names
gasindex

%--DEFINE CONSTANTS FOR MODEL-----------------------------
mwsub=29.6;
%A = exp(-7);
A = -8.116688; %Amylin 80 was 9.1188e-4;
%KLAB = 2.416; % was 2.4 kla_slope := 2.416
x(KLAB)=2.416;
DOCAL = 100;
YOUT0CAL = 20.9;
PTOTCAL = 1.1;
ODDCW = 3;
CXDCW = 25/1000;
%stoindex
RX=1;
RS=2;
RN=3;
RC=4;
RO=5;
RW=6;
% confirm not measured here
if ~midx(D02)
x(D02)=x(DIS02);
end;
x(QIN) = 2453*x(FGIN)/x(VL); % VOL TO SP. MOLAR CONVERSION
x(OTR) = x(QIN)*(x(YIN02)-x(YOUT02)*(x(YINN2)/x(YOUTN2)));
x(CER) = x(QIN)*(x(YUTCO2)*(x(YINN2)/x(YOUTN2))-x(YINCO2));
x(PLO2) = (x(DIS02)/DOCAL)*YOUT0CAL/100*PTOTCAL; %plo2 definition
x(PG02) = x(YOUT02)*x(PTOT); %pg02 definition(WELL MIXED)
x(KLA) = x(OTR)/(x(PG02)-x(PLO2)); % kla calculation
if x(KLA)>0
x(KLAB)= (log(x(KLA))-A)/log(x(N)); % kla correlation
else
x(KLAB)=0;
end;
% rq definition
x(TQ) = x(CER)/x(OTR);

x(RQ)=x(TQ);
%if x(RQ)>0.85
% x(RQ)=0.85;
end; % unstable if RQ is above 0.86 since RQ vs. Y is discontinuous
%x(DCCO2DT)=x(OTR)*1.05-x(CER);
x(CX)=x(DCW)/CXDCW;
if ~midx(MU)
x(MU)=(x(SUR)-x(CER))/x(CX);
end;
x(BPR)=x(MU)*x(CX);

% try to estimate sur based on cbalance and dorb balance...
% this assumes glycerol substrate...
x(SUR)=((x(BPR)+x(CER))+(4*x(OTR)+4.2*x(BPR))/4.67)/2;
if ~midx(YXS)
x(YXS)=0.4; % should improve based on RQ and substrate
end;
x(YXS)=x(BPR)/x(SUR);
if (~midx(SF) | ~midx(FS))
  if x(SF)==0
    x(SF)=0.5;
  end;
  % x(FS)=x(SFR)/1000/x(SF)*mwsub*x(VL);
else
  x(SFR) = x(FS)*1000*x(SF)/mwsub/x(VL);
end;
if ~midx(DCSDT)
x(DCSDT)=x(SFR)-x(SUR);
else
  % x(DCSDT)=((x(SFR)-x(SUR))*x(DCSDT))/2;
%end;
if ~midx(DCW) & midx(OD)
x(DCW) = x(OD)/ODDCW; % dry cell to od correlation
end;
if ~midx(OD) & midx(DCW)
x(OD)=x(DCW)*ODDCW;
end;

x(CCTOT)=bicarb(x(YOUTCO2),x(PTOT),x(CER),x(KLA),x(PH),x(TEMP));
if ~midx(FN)
x(FN)=x(CX)*x(MU)*0.17*x(VL);
else
  %x(FN)=(x(FN)*x(CX)*x(MU)*0.17*x(VL))/2;
end;
x(NUR)=x(MU)*x(CX)*0.17;
x(DCNDT)=x(FN)/x(VL)-x(NUR);
%x(HION)= 10^(-x(PH));
%x(WPR)=x(CER);
gaslabel
if JP_DEBUG
    disp('these values are 0 (bad for classification); consider replacing');
xvlabel(x==0,:)
end;
% x(x==0)=1e-6*ones(sum(x==0),1);
xguess=x;

B.2.8 gastrain.m

Function to represent process model. The contents of the model varied depending on the process being analyzed. Two scripts (gasindex.m and gaslabel.m) were used to associate variable names with entries in the variable vector.

function fx = gastrain(x)
%gastrain Constraint Equations for gas train reconciliation
%
% global variable names
global N PTOT DISO2 FGIN WS WN DCW OD S CNH4 TEMP VL PH
global YINO2 YINCO2 YINN2 YOUTO2 YOUTCO2 YOUTN2
global SFR FN KLA PLO2 PGO2 QIN CER OTR CO2 CCO2 CHC03 CS
global CX HO2 HCO2 HION DC02DT DCC02DT DCN03DT DCSDT MU RQ
global FS SF TIME D02 WA YINH2O YOUTH2O MSTIME YINNH3 YINAR YOUTNH3 YOUTAR
global YINC13C YOUTC13C GLYADDED BPR NUR DCSDT SUR DCN03T YXS
global KLAB CCTOT TQ
%
% global SCALE STEST
mwsup=29.6;
if columns(x)==1
    x=x';
end;
%--DEFINE CONSTANTS FOR MODEL-----------------------------------------------
A = -8.116688; %Amylin 80was 9.1188e-4;
%KLAB = 2.416; %was 2.4 kla_slope := 2.416
DCAL = 100;
YOUTO2CAL = 20.9;
PTOTCAL = 1.1;
ODDCW = 3;
CXDCW = 25/1000;
%stoindex
RX=1;RS=2;RN=3;RC=4;RO=5;
rcomp=zeros(3,RO); %c dor n

183
%disp('Glycerol substrate');
rcmp(:,RX)=[1.4 2.0 0.17]; %1 1.83 0.56 0.17
rcmp(:,RS)=[1 4.67 0];
rcmp(:,RN)=[0 0 1];
rcmp(:,RC)=[1 0 0];
rcmp(:,RO)=[0 -4 0];
%
pvars=zeros(RO,rows(x));
pvars(RS,:)=-x(:,SUR); %
pvars(RN,:)=-x(:,NUR);
pvars(RO,:)=-x(:,OTR);
pvars(RC,:)=-x(:,CER);
pvars(RX,:)=-x(:,BPR);
elebal=rcmp*pvars;
%
--Calculate constraint errors--------------------------
%numeqs=23;
numeqs=22;
fx=zeros(rows(x),numeqs);
fx(:,1)=x(:,OTR) - x(:,KLA).*((x(:,PGO2)-x(:,PLO2)); %kla calculation
fx(:,2)=x(:,PLO2) - (x(:,DISO2)/DOCAL)*YOUTO2CAL/100*PTOTCAL; %plo2 definition
fx(:,3)=x(:,PGO2) - x(:,YOUT2).*x(:,PTOT); %pgo2 definition(WELL MIXED)
fx(:,4)=x(:,OTR) - x(:,QIN).*((x(:,YINO2) - x(:,YOUT2).*((x(:,YIN2)/x(:,YOUT2))/)); %QTRdef
fx(:,5)=x(:,CER) - x(:,QIN).*((x(:,YIN2).*((x(:,YIN2)/x(:,YOUT2))- x(:,YINCO2)); %CERdef
fx(:,6)=x(:,QIN) - 2453.0*x(:,FGIN)/x(:,VL); % VOL TO SP. MOLAR CONVERSION
fx(:,7)=x(:,TQ) - x(:,SUR)/x(:,VL); %tq definition
fx(:,8)=x(:,DCW) - x(:,BP)*DXDCW; %dry cell to molar conc.
fx(:,9)=x(:,DCW) - x(:,OD)/ODDCW; %dry cell to od correlation
fx(:,10)=x(:,SFR) - x(:,SF)/1000.*x(:,SF)/mwsu./x(:,VL); %
%fx(:,11)=x(:,BPR) - x(:,SUR) + x(:,CER)
%fx(:,11)= elebal(1,:); %carbon balance
%fx(:,12)= elebal(2,:); %dor balance
%fx(:,13)= elebal(3,:); %nitrogen balance
fx(:,14)=x(:,BPR) - x(:,SUR).*x(:,YXS);
fx(:,15)=x(:,DISO2) - x(:,DO2);
fx(:,16)=x(:,YINO2)/x(:,YIN2) - (20.95/78.1); % assume atmospheric air(const ratio
fx(:,17)=x(:,YINO2)/x(:,YINAR) - (20.95/0.934);
fx(:,18)=x(:,YIN2)/x(:,YINART) - x(:,YOUT2)/x(:,YOUTAR);
fx(:,19)=x(:,SFR) - 0 - x(:,DCSDT) - x(:,SUR);
fx(:,20)=x(:,BPR) - x(:,MU).*x(:,CX);
fx(:,21)=x(:,FN)/x(:,VL)-0 - x(:,DCNDT) - x(:,NUR); %input-output-accumulation-reaction
fx(:,22)=x(:,KLA) - exp(A)*x(:,N)).*(-x(:,KLAB)); %kla correlation
%fx(:,23)=x(:,YINCO2)/0.0004*ones(rows(x),1);
% x(QR) - (x(CER)-x(DCCO2DT))/x(OTR); %rq definition
% x(QR)-1.05;
%Remember to update gasconst.m when revising this model.
% (10^-(-x(PH)) - x(HION))/10^-(-7); %added to test tenuous
fx=fx';
if ~STEST
    fx=fx./[SCALE(ones(columns(fx),1),1)];
end;

B.2.9 justgas.m
Subset of process model used in amylin gas train reconciliation example
function fx = justgas(x)
  %justgas  Constraint Equations for gas train reconciliation
  global ARUN MODEL_LEAK
  global FO FA FLEAK
  global YEO2 YMO YMC YMN
  global KLAB N D01 PTOT
  global SCALE STEST
  if columns(x)==1
    x=x';
  end;
  %--DEFINE CONSTANTS FOR MODEL---------------------------------------
  YOUTO2CAL = 20.9;
  if     ARUN==2033, DOCAL = 1.39; PTOTCAL = 1.67;
    elseif ARUN==2034, DOCAL = 1.4; PTOTCAL = 1.85;
    elseif ARUN==2042, DOCAL = 1.39; PTOTCAL = 1.67;
    elseif ARUN==2043, DOCAL = 1.43; PTOTCAL = 1.68;
    else error('run not in sensor calibration database');
  end;
  %----------------------------------
  onavec=ones(rows(x),1);
  cVL =onavec*16;
  cY0D=onavec;
  cY0N=onavec*0;
  cYOC=onavec*0;
  cYAO=onavec*0.2095;
  cYAN=onavec*0.7900;
  cYAC=onavec*0.0006;
  %calculations----------------------
  %-mixer FG=FO+FA
  cFG=x(:,FO)+x(:,FA);
  cYG0= (x(:,FO).*cY0D + x(:,FA).*cYAO)./cFG;  % mixer
  cYGc= (x(:,FO).*cY0C + x(:,FA).*cYAC)./cFG;  % mixer
  cYGN= (x(:,FO).*cY0N + x(:,FA).*cYAN)./cFG;  % mixer
  cQIN=2453.0*cFG./cVL/1000; %scaled
  %-measurements and leak model----
  cYEN=cYGN; % assume constant dry molar flow i/o
  if MODEL_LEAK
    cYEO = x(:,YMO)-x(:,FLEAK);
    cYEC = x(:,YMC)+x(:,FLEAK);
else
cYEO = x(:,YMO);
cYEC = x(:,YMC);
end;
cFE = cFG.*cYGN./cYEN; % nitrogen balance
cOTR = cQIN.*(cYGO - cYEO.*cYGN./cYEN); % oxygen
cCER = -cQIN.*(cYG - cYEC.*cYGN./cYEN); % carbon dioxide
% mass transfer-------------
cPLO2 = (x(:,D01)/DICAL)*YOUTO2CAL/100*PTOTCAL; % plo2 do probe estimate
cPGO2 = cYEO.*x(:,PTOT); % pgo2 well mixed estimate
cLNA = (x(:,KLAB)-0.6745)/(-0.1639);
cKLAB = x(:,KLAB);
cKLACORR = exp(cLNA).*(x(:,N)*60).^cKLAB/1000; % scale
cOTRkla = cKLACORR.*(cPGO2-cPLO2);
%------------------------
numeqs=2;
fx=zeros(rows(x),numeqs);
fx(:,1) = cOTR - cOTRkla; % kla correlation = calculation
fx(:,2) = cYEO - x(:,YEO2); % redundant oxygen sensor
if MODEL_LEAK
  fx(:,3) = cYEN-x(:,YMN); % est ymn
  fx(:,4) = cOTR - cCER; % RQ = 1
end;
fx=fx';
if ~STEST % Scale output for better convergence properties
  fx=fx./[SCALE(ones(columns(fx),1),:,:)]';
end;

186