COMPUTER-AIDED DESIGN OF BIOCHEMICAL PATHWAYS

by

MICHAEL L. MAVROVOUNIOTIS

Diploma Eng., Chemical Engineering
National Technical University
Athens, Greece
(1984)

Submitted to the Department of Chemical Engineering
in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

at the

Massachusetts Institute of Technology
December 1988

© Massachusetts Institute of Technology 1988. All rights reserved.

Signature of Author ____________________________

Department of Chemical Engineering

Certified by

George Stephanopoulos and Gregory Stephanopoulos
Thesis Supervisors

Accepted by ____________________________

Robert C. Armstrong
Committee for Graduate Students

ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
FEB 23 1989
COMPUTER-AIDED DESIGN OF BIOCHEMICAL PATHWAYS

by

MICHAEL L. MAVROVOUNIOTIS

Submitted to the Department of Chemical Engineering on December 8, 1988 in Partial Fulfillment of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

ABSTRACT

The problem of design of biochemical pathways is addressed through a combination of methods based on Artificial Intelligence, Mathematics, Physical Chemistry, and engineering principles. The methods cover both the development of a new pathway for the production of a bioproduct and the evaluation of an existing pathway.

A database for metabolites and bioreactions contains information needed by various methodologies. Candidate pathways can be generated by a constrained search algorithm, which is based on the iterative satisfaction of constraints through the transformation of a base-set of pathways. Generated pathways can be screened by heuristic rules and detailed analytical examination. Semiquantitative knowledge can be used within the O[M] formalism for Order-of-Magnitude reasoning, which captures engineering commonsense concepts about the relative orders of magnitude of parameters.

Additional methodologies make up for the absence of critical data. They include the estimation of maximum rates for biochemical reactions based on collision limitations, and a group-contribution method for the estimation of equilibrium constants. Using these parameters, thermodynamic and kinetic feasibility analysis of pathways can be accomplished, and the bottlenecks of the pathway can be identified. The bottlenecks can be bypassed by synthesizing appropriate modifications to the pathway.

Thesis Supervisors: Dr. George Stephanopoulos Dr. Gregory Stephanopoulos
Titles: Professor of Chemical Engineering Professor of Chemical Engineering
ACKNOWLEDGEMENTS

I would first like to thank my advisors Profs. George Stephanopoulos and Gregory
Stephanopoulos, whose insights and ideas had a major impact on the directions this
research took and the breadth it covered. Most of all, however, I am indebted to them
for their trust and support, and the freedom they gave me to pursue my ideas. I would
also like thank my thesis committee, Profs. Gene Brown, Aaron Fleisher, Ramesh Patil,
and Daniel Wang, for their guidance and encouragement.

The extensive stimulating discussions with my fellow students were a large part
of my education at MIT and a source of many ideas which influenced my research. I want
to thank in particular Theodore Kritikos, Kevin Joback, and Rama Lakshmanan. I also
thank Joe Vallino and Bob Kiss who provided needed experimental data.

I am grateful to the Biotechnology Process Engineering Center for the financial
support for the project. With the support of the Center, I had the opportunity to work
with Tu-Kien Michael Lam, an undergraduate student whose assistance was very
valuable.

Nothing I can say about my family can express the magnitude of my gratitude to them.
My wife, Gretchen, my father, Λοιζος, my mother, Καλλιοπη, my brothers, Γιωργος
and Δημητρης, and my sister, Νικη, have offered their companionship, love, and
encouragement through easy times and hard; I would not have made it this far without
their support. It is hard to put in words how much I am especially indebted to my
parents for everything they have done for me; I will single them out by dedicating this
thesis to them.

Michael L. Mavrovouniotis
στον πατέρα μου Λοίζο
και τη μητέρα μου Καλλιοπή

to my father Loizos
and my mother Kalliopi
TABLE OF CONTENTS

1 PREFACE 35

CHAPTER 2: OVERVIEW 37

2.1. Summary 38

2.2. The Problem of Pathway Design 39
  2.2.1. MOTIVATION 39
  2.2.2. GENERAL AIMS 39
  2.2.3. SPECIFIC GOALS 39

2.3. Pathway Analysis 41
  2.3.1. NATURE OF THE PROBLEM 41
  2.3.2. CHARACTER OF THE KNOWLEDGE 41
  2.3.3. GROUP-CONTRIBUTION METHOD FOR THE ESTIMATION OF
         EQUILIBRIUM CONSTANTS OF BIOCHEMICAL REACTIONS 42
  2.3.4. COLLISION-LIMIT BASED ESTIMATION OF MAXIMUM RATES FOR
         ENZYMATIC REACTIONS 42
  2.3.5. ORDER-OF-MAGNITUDE REASONING 44
2.4. Pathway Synthesis

2.4.1. IMPORTANT CHARACTERISTICS OF THE PROBLEM FORMULATION

2.4.2. ALGORITHM

2.4.3. APPLICABILITY

2.5. Integrated Framework

CHAPTER 3: INTRODUCTION

3.1. The Domain: Biotechnology

3.1.1. BIOCHEMICAL PATHWAYS

3.1.2. BIOPROCESSES

3.1.3. MUTATION-SELECTION PROGRAMS

3.1.4. GENETIC ENGINEERING

3.2. The Problem: Design of Biochemical Pathways

3.2.1. SYNTHESIS OF BIOCHEMICAL PATHWAYS

3.2.2. CRITICAL SYNTHESIS ISSUES

3.2.3. ANALYSIS OF BIOCHEMICAL PATHWAYS

3.2.4. CRITICAL ANALYSIS ISSUES

3.2.5. CONVENTIONAL BIOLOGICAL MODELS

3.2.5.1. Fermentation Equations

3.2.5.2. Unstructured Models

3.2.5.3. Cell-Population Models

3.2.5.4. Structured Culture Models
3.2.5.4.1. Universal compartmentalization 67
3.2.5.4.2. Critical compound distinction 68
3.2.5.5. Single-Cell Models 69
3.2.5.6. Discussion 71

3.3. The Approach 73
3.3.1. ARTIFICIAL INTELLIGENCE METHODS 73
3.3.1.1. Rule-Based Expert Systems 74
3.3.1.2. Qualitative Reasoning 78
3.3.1.3. Assumption-Based Truth-Maintenance 78
3.3.1.4. Search 80
3.3.2. COMPUTING ENVIRONMENT AND PROGRAMMING STYLE 81
3.3.3. PREDICTIVE METHODS 84

3.3. The Goal: Study of Pathway Design 85

CHAPTER 4: RESEARCH FRAMEWORK 86

4.1. Introduction 87
4.1.1. CHARACTER OF THE BIG PICTURE 87
4.1.2. CONTEXTS 88

4.2. Graphic Interface 90

4.3. Biochemical Database Context 93
4.3.1. OBJECT TYPES 93
4.3.1.1. Distinction between Bioreactions and Enzymes 94
4.3.1.1.1. Bioreactions 94
4.3.1.1.2. Enzymes 94
4.3.1.1.3. Distinctions and Usage 95
4.3.2. DESIRED DATA 96
4.3.2.1. Small Molecules 98
4.3.2.1.1. Structural Data 98
4.3.2.1.2. Physicochemical Data 98
4.3.2.1.3. Physiological Data 99
4.3.2.1.4. Other Data 99
4.3.2.2. Bioreactions 100
4.3.2.3. Enzymes 100
4.3.2.4. Pathways 101
4.3.2.5. Microorganisms 102
4.3.3. OBJECT CONNECTIONS AND CLASSIFICATION 102
4.3.3.1. Relations among Objects 102
4.3.3.1.1. Relations for Molecules 103
4.3.3.1.2. Relations for Bioreactions 103
4.3.3.1.3. Relations for Enzymes 104
4.3.3.1.4. Relations for Pathways 104
4.3.3.2. Classification of Objects 105
4.3.3.2.1. Metabolites 105
4.3.3.2.2. Bioreactions 105
4.3.3.2.3. Enzymes 106
4.3.3.2.4. Pathways 107
4.3.3.2.5. Microorganisms 107
4.3.4. INTELLIGENT FEATURES OF THE DATABASE 108

4.3.4.1. Consistency Checks 109

4.3.4.1.1. Examples of Inconsistencies that can be Checked for 110

4.3.4.1.2. Apparent Inconsistencies that cannot be Resolved 110

4.3.4.1.3. How to handle Inconsistencies 111

4.3.4.2. Chaining 111

4.3.4.2.1. Search Modes 112

4.3.4.2.2. Applications 112

4.3.4.3. Informative Answers to Queries 113

4.3.4.4. ABstraction 114

4.4. Pathway Synthesis Context 115

4.4.1. INTERACTIVE SYNTHESIS 115

4.4.1.1. Disconnection and Deletion of Objects 115

4.4.1.2. Side Metabolites in Bioreactions 116

4.4.2. AUTOMATIC SYNTHESIS 119

4.4.2.1. Specifications 119

4.4.2.2. Methodology Developed 120

4.4.3. NON-ENZYMATIC REACTIONS 121

4.4.3.1. Spontaneous Reactions 121

4.4.3.2. Conventional Chemical Reactions 123

4.5. Evaluation and Prediction Context 125

4.5.1. DEALING WITH INCOMPLETE KNOWLEDGE 125

4.5.2. DEALING WITH QUALITATIVE KNOWLEDGE 126

4.5.3. DETERMINATION OF RATE-LIMITING STEPS 127
4.5.4. CONJECTURING REGULATORY MECHANISMS 128
4.5.4.1. Heuristics for Wild-Type Regulation 129
4.5.4.2. From Wild-Type to Mutant Regulation 130
4.5.4.3. Testing of Regulatory Mechanisms 131
4.5.4.4. Difficulties 132

4.6. Pathway Recommendation Context 133
4.6.1. INDUSTRIAL ASPECTS 133
4.6.2. GENETICS 134

4.7. Learning Context 136
4.7.1. SEARCH FOR BIOLOGICAL REGULARITIES 136

CHAPTER 5: FORMAL REASONING WITH ORDERS OF MAGNITUDE AND APPROXIMATE RELATIONS 138

5.1. Introduction 139
5.1.1. HIERARCHY OF MODELS 139
5.1.2. QUALITATIVE REASONING 140
5.1.3. ORDER-OF-MAGNITUDE CONCEPTS IN THE ENGINEERING ANALYSIS OF BIOCHEMICAL SYSTEMS 142
5.1.3.1. Knowledge 142
5.1.3.2. Reasoning 143
5.1.4. NEED FOR FORMALIZATION 144

5.2. Previous Work 146
5.2.1. AI CONCEPTS INDIRECTLY RELATED TO ORDERS OF MAGNITUDE 146

5.2.2. THE FOG SYSTEM 147
5.2.2.1. Deficiencies of FOG 147
5.2.2.2. Need for a New Formal System 149

5.3. Advantages of Formal Order-of-Magnitude Reasoning 150

5.4. Quantities and Links 152
5.4.1. VARIABLES AND SIGN SPECS 152
5.4.2. LANDMARKS 152
5.4.3. LINKS 152

5.5. Primitive and Compound Relations 154
5.5.1. PRIMITIVE RELATIONS 154
5.5.2. COMPOUND RELATIONS 156
5.5.3. RESTRICTIONS 156
5.5.4. EXPRESSIVENESS 159

5.6. Semantics and Properties of O[M] relations 161
5.6.1. STRICT INTERPRETATION 161
5.6.1.1. Determination of the Boundaries of the Intervals 163
5.6.1.2. Remaining Degree of Freedom 165
5.6.1.3. Advantages and Disadvantages of the Interpretation 165
5.6.2. HEURISTIC INTERPRETATION 166
5.6.2.1. Construction 168
5.6.2.2. Determination of the Boundaries 170
5.6.2.3. Other Inferences 170
5.6.3. SECONDARY PROPERTIES OF ORDER-OF-MAGNITUDE RELATIONS 172
5.6.3.1. Inverse Relations 172
5.6.3.2. Transitivity of Relations 173

5.7. Other Means of Knowledge Representation 175
5.7.1. ASSIGNMENTS 175
5.7.2. CONSTRAINTS 176
5.7.3. RULES 177

5.8. Reasoning 178
5.8.1. EXAMPLES 178
5.8.2. INFERENCE METHODS 180
5.8.3. TRUTH-MAINTENANCE AND RESOLUTION OF CONTRADICTIONS 183
5.8.4. GOAL DIRECTION 187

5.9. Discussion 189

5.10. Application of O[M] In
Process Engineering Systems and Activities 191
5.10.1. ENGINEERING ACTIVITIES IN WHICH ORDER-OF-MAGNITUDE ANALYSIS IS RELEVANT 191
5.10.1.1. Preliminary Design of Process Flowsheets 192
5.10.1.2. Preliminary Design of Control Structures 192
5.10.1.3. Planning of Process Operations 193
5.10.1.4. Process Trend Analysis Fault Diagnosis 194
5.10.2. DETAILED REASONING EXAMPLES 194
5.10.2.1. A Heat Exchanger 194
5.10.2.1.1. Important Parameters 196
5.10.2.1.2. Constraints and Relations 198
5.10.2.1.3. Interences 198
5.10.2.2. Stirred-Tank and Plug-Flow Chemical Reactors 202
5.10.2.2.1. Parameters and Constraints 203
5.10.2.2.2. Order-of-Magnitude Analysis 203

5.11. Summary 205

CHAPTER 6: APPLICATIONS OF ORDER-OF-MAGNITUDE REASONING IN THE ANALYSIS OF BIOCHEMICAL SYSTEMS 206

6.1. Analysis of Isolated Enzymatic Reactions 207
6.1.1. MICHAELIS-MENTEN KINETICS 207
6.1.2. INHIBITION OF AN ENZYMATIC REACTION 210
6.1.2.1. Competitive inhibition 210
6.1.2.1.1. Refinement through Additional Knowledge 212
6.1.2.1.2. Inference of Concentrations from Rates 212
6.1.2.2. Uncompetitive Inhibition 215
6.1.2.3. Noncompetitive Inhibition  215
6.1.3. COMPARISON OF INHIBITION MODES  218
6.1.3.1. Inhibition Mode Diagnosis  219
6.1.4. IDENTIFICATION OF MICHAELIS OR INHIBITION CONSTANTS  219

6.2. Analysis of Biochemical Pathways  221
6.2.1. COMPETITION OF PATHWAYS  221
6.2.2. ANALYSIS OF FLUXES IN BIOCHEMICAL NETWORKS  224
6.2.2. iDEN TIFICATION OF RATE-LIMITING STEPS  225
6.2.2.1. Formulation of the Hypothesis  225
6.2.2.2. Example: Aldolase  228
6.2.2.2. Example: Phosphoglycerate Dehydrogenase  229
6.2.3. TESTING HYPOTHESES ON FEEDBACK REGULATORY MECHANISMS OF PATHWAYS  232

6.3. Concluding Remarks  236

CHAPTER 7: A GROUP CONTRIBUTION METHOD FOR THE ESTIMATION OF EQUILIBRIUM CONSTANTS FOR BIOCHEMICAL REACTIONS  237

7.1. Introduction  238
7.1.1. GROUP-CONTRIBUTION METHODS  238
7.1.2. MOTIVATION FOR THE DEVELOPMENT OF THE METHOD  239

7.2. Estimation of the Contributions of Groups  241
7.2.1. DATA USED IN THE ESTIMATION 241
    7.2.1.1. Data on Molecules 241
    7.2.1.2. Data on reactions 242

7.2.2. DECOMPOSITION OF COMPOUNDS TO GROUPS 243

7.2.3. ASSUMPTIONS 244

7.2.4. RESULTS 245
    7.2.4.1. Accuracy 248

7.3. Examples 249

7.3.1. GIBBS ENERGIES OF FORMATION OF COMPOUNDS 249

7.3.2. EQUILIBRIUM CONSTANTS OF REACTIONS 251

7.4. Summary and Significance 253

CHAPTER 8: COLLISION-LIMIT BASED ESTIMATION OF MAXIMUM RATES FOR ENZYMATIC REACTIONS 276

8.1. Introduction 277

8.1.1. PREVIOUS WORK 277

8.1.2. THE METHOD PRESENTED HERE 277

8.1.2. SIGNIFICANCE 278

8.2. Problem and Assumptions 280

8.3. Treatment of a Two-Reactant Two-Product Ordered Mechanism 282
8.3.1. DERIVATION OF THE RATE EQUATION 282
8.3.2. INTRODUCTION OF CONSTRAINTS 285
8.3.3. TRANSFORMATION OF PARAMETERS 286
8.3.4. DERIVATIVES 286
8.3.5. OPTIMIZATION 288
8.3.6. RESULT 289

8.4. Nondimensionalization 291
8.4.1. GENERAL ORDERED ENZYMATIC REACTION 291
8.4.2. DIMENSIONLESS PARAMETERS 292
8.4.2.1. Physical Significance of the Parameters 292
8.4.3. DIMENSIONLESS TREATMENT OF THE EXAMPLE 293
8.4.4. RESULTS FOR OTHER ORDERED MECHANISMS 294

8.5. Example of Asymptotic Behavior 305
8.5.1. EFFECT OF f IN A SIMPLIFIED SYSTEM 305
8.5.2. REACTION VERY FAR FROM EQUILIBRIUM 306
8.5.2.1. First-Order Approximation 306
8.5.2.2. Second-Order Approximation 306
8.5.2.3. Region of Influence of the Reversibility of the Reaction 307
8.5.3. REACTION NEAR EQUILIBRIUM 307

8.6. Estimation of Collision Parameters 310
8.6.1. STERIC FACTORS 310
8.6.1.1. Substrate 310
8.6.1.2. Enzyme 311
8.6.2. APPROXIMATE RATE OF ENCOUNTER 311
8.6.2.1. Simplifications Due to the Relative Sizes of the Species 312
8.6.2.2. Substrate Diffusion Coefficient 312
8.6.2.3. Enzyme Radius 312
8.6.2.4. Resulting Rate of Encounter 313
8.6.3. FINAL COLLISION PARAMETER 313
8.6.4. RANGES FOR OTHER PARAMETERS 314

8.7. Numerical Examples 316
8.7.1. ISOLATED REACTIONS 316
8.7.2. A SHORT PATHWAY 318
8.7.2.1. Parameters for Small Molecules 318
8.7.2.2. Parameters for Enzymes 318
8.7.2.3. Results 321

8.8. Extensions 324
8.8.1. EXTREMA FOR OTHER PARAMETERS 324
8.8.2. EXPLOITATION OF ADDITIONAL CONSTRAINTS 325
8.8.2.1. Constraints on the Maximum Enzyme Turnover 325
8.8.2.2. Constraints on the Michaelis Constant 326
8.8.2.3. Other Constraints 326

8.9. Discussion 328
8.9.1. SIMPLIFICATION OF THE RESULTS 328
8.9.2. HANDLING RANDOM-ORDER MECHANISMS 328
8.9.3. RANGES FOR OTHER PARAMETERS 329
8.9.4. APPLICATIONS
8.9.5. MULTI-ENZYME COMPLEXES
8.9.6. ACTIVE CONCENTRATIONS OF CURRENCY METABOLITES
8.9.7. METABOLIC EFFICIENCY AND EVOLUTION OF ENZYMES
8.9.8. CONCLUSIONS

8.10. Summary

CHAPTER 9: SYNTHESIS OF BIOCHEMICAL PATHWAYS SATISFYING STOICHIOMETRIC SPECIFICATIONS

9.1. Introduction
9.1.1. VARIANTS OF THE SYNTHESIS PROBLEM
9.1.1.1. Base Problem: Synthesis of Whole Pathways
9.1.1.2. Modifications to Existing Pathways
9.1.1.2.1. Bottlenecks in Pathways
9.1.1.2.2. Removal of Bottlenecks
9.1.1.2.3. Equivalence to Base Problem
9.1.1.3. Completion of Partial Pathways
9.1.2. NATURE OF THE SPECIFICATIONS AND CONSTRAINTS
9.1.2.1. Stoichiometric Specifications
9.1.2.1.1. Reactants
9.1.2.1.2. Products
9.1.2.1.3. Intermediates
9.1.2.1.4. Bioreactions
9.1.2.2. Thermodynamic Specifications
9.1.2.3. Handling Specifications
9.1.3. PREVIOUS WORK

9.2. Problem Formalization
9.2.1. STOICHIOMETRIES OF A PATHWAY
9.2.2. FORMALIZATION OF STOICHIOMETRIC SPECIFICATIONS
9.2.2.1. Constraints on the Molecular Stoichiometry
9.2.2.2. Constraints on the Intermediate Stoichiometry
9.2.2.3. Constraints on the Reaction Stoichiometry

9.3. Algorithm for the Synthesis of Biochemical Reaction Pathways
9.3.1. INTRODUCTION OF ADDITIONAL CONSTRAINTS
9.3.1.1. Pathway Properties and Relations
9.3.1.2. Additional Constraints
9.3.2. OVERVIEW OF THE ALGORITHM
9.3.3. REACTION-PROCESSING PHASE
9.3.3.1. Reaction Reversibility
9.3.3.2. Excluded Reactions
9.3.4. METABOLITE-PROCESSING PHASE
9.3.4.1. State of the Design
9.3.4.2. Pathway Expansion
9.3.4.2.1. Pathways Constructed and Deleted
9.3.4.2.2. Requirements that Remain
9.3.4.2.3. Redundancy Checks
9.3.4.3. Updating the State of the Design
9.3.4.3.1. Deletion of an Existing Pathway 361
9.3.4.3.1. Addition of a New Pathway 361
9.3.5. PATHWAY-MARKING PHASE 362
9.3.5.1. Satisfaction of Strict-Inequality Constraints 362
9.3.5.2. Neutral Pathways 364
9.3.6. PROPERTIES OF THE ALGORITHM 365
9.3.6.1. Mathematical Properties 365
9.3.6.1.1. Correctness Theorem 365
9.3.6.1.2. Correctness Proof 366
9.3.6.1.3. Completeness Theorem 367
9.3.6.1.4. Completeness Proof 367
9.3.6.1.5. Polynomial Complexity Theorem 369
9.3.6.1.6. Polynomial Complexity Proof 369
9.3.6.1.7. Corollary on the Complexity with respect to Pathway Size 370
9.3.6.2. Computational Properties 370

9.4. Discussion and Simple Examples 372
9.4.1. DEPENDENCE OF THE COMPUTATIONAL EFFICIENCY ON THE FORMULATION OF THE PROBLEM 372
9.4.1.1. Common Currency Metabolites 372
9.4.1.1. Synthesis of Serine 375
9.4.1.1.1. Initial Formulation 375
9.4.1.1.2. Inclusion of Acetyl-CoA 376
9.4.1.1.3. Inclusion of Malate 377
9.4.1.1.4. Inclusion of Oxaloacetate 386
9.4.1.2. Synthesis of Alanine 390
9.4.1.2.1. Initial Formulation

9.4.1.2.2. Inclusion of Malate and Acetyl-CoA

9.4.1.2.3. Decomposition of the Problem

9.4.1.2.4. Synthesis of Alanine from Pyruvate

9.4.1.2.5. Elimination of Equivalent Bioreactions

9.4.2. TAKING ADVANTAGE OF CHARACTERISTICS OF BIOCHEMICAL NETWORKS

9.5. Summary

CHAPTER 10: A CASE STUDY IN PATHWAY DESIGN

10.1. Introduction

10.1.1. LYSINE

10.1.2. FRAMEWORK OF THE CASE STUDY

10.2. A Basic Pathway for the Production of Lysine

10.2.1. BIOCHEMICAL NETWORK

10.2.2. CONSTRUCTION OF BASIC PATHWAY

10.3. Maximum-Rate Analysis

10.3.1. PARAMETER VALUES

10.3.2. BASE OF REFERENCE FOR CONCENTRATIONS AND RATES

10.3.3. ESTIMATION OF THE ENZYME REQUIREMENTS OF THE BASIC LYSINE PATHWAY
10.3.3.1. Identification of Kinetic Bottlenecks 431
10.3.3.2. Comparison to Experimental Enzyme Requirements 431
10.3.3.2.1. Estimation of Experimental Enzyme Requirements 431
10.3.3.2.2. Factors causing Disparity between Minimum and Actual Requirements 433
10.3.3.2.3. Interpretation of the Relation between Estimated and Experimental Requirements 433

10.4. Alternative Pathways 436
10.4.1. PATHWAYS INVOLVING CARBOXYLATION OF PYRUVATE 436
10.4.2. PATHWAYS RETAINING TCA 438
10.4.2.1. Pathway through Lactate 438
10.4.2.2. Pathways deriving Aspartate directly from Fumarate 440
10.4.2.2.1. Persistent Intermediates 441
10.4.3. ALTERNATIVE ENTRIES OF PYRUVATE INTO TCA 444
10.4.3.1. Carboxylation of PEP or Pyruvate 444
10.4.3.2. Pathways through Acetate 446
10.4.4. COMPLEX PATHWAYS 448

10.5. Fundamental Constraints 450

CHAPTER 11: CONCLUSIONS AND SIGNIFICANCE 451

11.1. Summary 452
11.1.1. MOTIVATION AND GOALS 452
11.1.2. WORK PERFORMED 452
11.1.2.1. Pathway Analysis 453
11.1.2.1.1. Group- Contribution Method 454
11.1.2.1.2. Estimation of Maximum Rates for Enzymatic Reactions 454
11.1.2.1.3. Order-of-Magnitude Reasoning 455
11.1.2.2. Pathway Synthesis 455

11.2. Reiteration of Significant Contributions 457

11.3. Future Directions 459
11.3.1. GENETIC CONSIDERATIONS 459
11.3.2. POSTULATING ENZYMES 459
11.3.3. HOST CELLS 460

12 BIBLIOGRAPHY 461

APPENDIX A: COMPUTER PROGRAMS 479

A.1. Summary 480

A.2. Introduction 482
A.2.1. PROGRAMMING STYLE 482
A.2.1.1. Character of the Problem 482
A.2.1.2. LISP 482
A.2.1.3. Object-Oriented Programming 483
A.2.2. COMPUTING ENVIRONMENT 483
A.2.2.1. Environment and Facilities 484
A.2.2.2. Advantages of the Environment 484
A.2.2.3. Debugging 486
A.2.3. FACILITIES DEVELOPED 486
A.2.4. DATABASE 487
A.2.4.1. Justification of the Database 488
A.2.4.2. Character of the Database 488

A.3. Program Modules 490

A.4. Using the Programs 500
A.4.1. LOADING THE CODE 500
A.4.2. PREPARATION OF THE GRAPHIC INTERFACE 500
A.4.3. WINDOW PANES 503
A.4.4. MODES OF OPERATION 504
A.4.5. COMMANDS 505
A.4.6. CREATING AND EDITING DATABASE OBJECTS 510
A.4.6.1. Data Attributes 510
A.4.6.2. Display of Information on Objects 515
A.4.6.3. Editing of Objects 516
A.4.6.4. Creation of New Objects 519
A.4.7. ORDER-OF-MAGNITUDE REASONING 519

A.5. Inside the Programs 520
A.5.1. BIOCHEMICAL DATABASE OBJECTS 520
A.5.2. ORDER-OF-MAGNITUDE OBJECTS 548
A.5.2.1. Relators 548
A.5.2.2. Landmarks 548
A.5.2.3. Variables 549
A.5.2.4. Links 549
A.5.2.5. Relations 550
A.5.2.6. Rules 551
A.5.2.7. Assignments 551
A.5.2.8. Constraints 552
A.5.2.9. Use of the System 552
A.5.3. IMPLEMENTATION EFFICIENCY ISSUES 555
LIST OF FIGURES

Figure 4.1: Distinction of metabolites into main-reactants and side-reactants (as well as main-products side-products) for three particular bioreactions.

Figure 4.2: In the pathway for the synthesis of proline, the conversion of L-Glutamate-γ-semialdehyde to L-Δ¹-Pyrroline-5-carboxylate occurs spontaneously.

Figure 4.3: A route for the production of lysine via enzymatic transformation of a chemically produced intermediate.

Figure 5.1: Strict interpretation of the relation A rₙ B.

Figure 5.2: Constrained strict interpretation of the relation A rₙ B.

Figure 5.3: Vague interval boundaries for the heuristic interpretation.

Figure 5.4: Intervals used in the heuristic interpretation.

Figure 5.5: Final intervals for the heuristic interpretation of O[M] relations.

Figure 5.6: A countercurrent heat-exchanger, with some of its important parameters: Molar flowrates and molar heat-capacities of the two streams.

Figure 5.7: (a) Important temperature differences. (b) A simple sketch of the temperature profiles of the streams along the length of the device.

Figure 5.8: The temperature profiles of the streams of the heat-exchanger, drawn based on the results of Order-of-Magnitude reasoning.

Figure 6.1: Competition between the pathways of glycolysis and serine synthesis.

Figure 6.2: (a): An abstract linear pathway converting substrate A to product F, through the enzymatic reactions r₁, r₂, ..., rₙ. (b): A section of the glycolytic pathway, which is an instance of such a linear pathway. The section shown converts glucose to 2-phosphoglycerate.
Figure 6.3: The pathway for the synthesis of serine from 3-phosphoglycerate, which is an intermediate of glycolysis.

Figure 6.4: Reasoning that leads to the rejection of the hypothesis that phosphoglycerate dehydrogenase is the rate-limiting step of the serine pathway.

Figure 6.5: Regulation of a linear biosynthetic pathway. The final product of the pathway exerts negative control on an upstream step.

Figure 8.1: The general ordered mechanism for a two-reactant two-product enzymatic reaction.

Figure 8.2: General numerical solution and analytic asymptotic solutions for one-reactant three-product enzymatic reactions, with $u_1 = 1$, $u_2 = 1$, and $u_3 = 1$.

Figure 8.3: Simple maximum-rate estimations for reactions from the serine pathway.

Figure 8.4: A section of the glycolytic pathway.
Figure 9.1: Synthesis of serine from glucose, with recovery of glutamate by
*Glutamate-dehydrogenase*

Figure 9.2: Synthesis of serine from malate, with recovery of glutamate by
*Glutamate-dehydrogenase*

Figure 9.3: Synthesis of serine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate by *Glutamate-dehydrogenase*

Figure 9.4: Synthesis of serine from glucose, with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate

Figure 9.5: Synthesis of serine from malate, with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate

Figure 9.6: Synthesis of serine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate

Figure 9.7: Synthesis of serine from glucose, with recovery of glutamate accompanied by consumption of Malate and production of oxaloacetate.  

Figure 9.8: Synthesis of serine from malate, with recovery of glutamate accompanied by consumption of Malate and production of oxaloacetate.

Figure 9.9: Synthesis of serine from malate and Acetyl-CoA, with recovery of glutamate accompanied by consumption of Malate and production of oxaloacetate.

Figure 9.10: Synthesis of alanine from glucose, with recovery of glutamate by *Glutamate-dehydrogenase*

Figure 9.11: Synthesis of alanine from malate, with recovery of glutamate by *Glutamate-dehydrogenase*

Figure 9.12: Synthesis of alanine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate by *Glutamate-dehydrogenase*

Figure 9.13: Synthesis of alanine from glucose, with recovery of glutamate through a
loop involving oxaloacetate, aspartate, fumarate, and malate

**Figure 9.14:** Synthesis of alanine from malate, with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate

**Figure 9.15:** Synthesis of alanine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate

**Figure 9.16:** Synthesis of alanine from glucose, with *Alanine-dehydrogenase*

**Figure 9.17:** Synthesis of alanine from malate, with *Alanine-dehydrogenase*

**Figure 9.18:** Synthesis of alanine from malate and Acetyl-CoA, without production of carbon dioxide, and with *Alanine-dehydrogenase*

**Figure 9.19:** Synthesis of alanine from pyruvate, with recovery of glutamate by *Glutamate-dehydrogenase*

**Figure 9.20:** Synthesis of alanine from pyruvate, with recovery of glutamate through a loop involving aspartate, fumarate, malate, and oxaloacetate.

**Figure 9.21:** Synthesis of alanine from pyruvate through a single bioreaction, *Alanine dehydrogenase*, which does not require glutamate

**Figure 9.22:** The structured character of biochemical reaction networks is exploited by the synthesis algorithm in early pruning and abstraction

**Figure 10.1:** The basic bioreaction network for the synthesis of lysine.

**Figure 10.2:** Partial pathway of the normal route of lysine production.

**Figure 10.3:** Completion of the basic pathway for the synthesis of lysine.

**Figure 10.4:** Calculation of minimum enzyme requirements for the basic pathway for lysine production.

**Figure 10.5:** Minimum enzyme requirements for a lysine pathway involving carboxylation of pyruvate.
Figure 10.6: Pathway converting malate to oxaloacetate, with lactate and pyruvate as intermediates.

Figure 10.7: The simplest of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate.

Figure 10.8: The kinetically most efficient of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate.

Figure 10.9: A pathway that carboxylates PEP, bypassing pyruvate.

Figure 10.10: A set of pathways that bypass pyruvate dehydrogenase by converting pyruvate to acetate and then to Acetyl-CoA or citrate.

Figure 10.11: The longest pathway converting PEP to pyruvate.

Figure A.1: Interdependencies among modules.

Figure A.2: Initial form of the interface frame.

Figure A.3: Displaying and Editing the reaction Fructose-Diphosphate-Aldolase.

Figure A.4: Displaying and editing the glycolytic pathway.

Figure A.5: Interrelationships among Biochemical Objects.

Figure A.6: Interconnections among Order-of-Magnitude objects.
LIST OF TABLES

Table 3.1: Problem selection criteria for applying Rule-Based Expert Systems.

Table 5.1: Primitive relations of the O[M] formalism.

Table 5.2: Compound relations of the O[M] formalism.

Table 5.3: O[M] relations representing relations that are commonly used in engineering.

Table 5.4: Order-of-Magnitude Analysis of CSTR and PFR isothermal reactors for an irreversible, first-order reaction.

Table 6.1: Order-of-Magnitude analysis of the rate of a biochemical reaction that follows the Michaelis-Menten rate relation.

Table 6.2: Order-of-Magnitude analysis of the rate of a biochemical reaction with competitive inhibition.

Table 6.3: Order-of-Magnitude analysis of the concentration of a competitive inhibitor of an enzyme, given knowledge on the concentration of the substrate and the rate of the reaction.

Table 6.4: Order-of-Magnitude analysis of the rate of a biochemical reaction with uncompetitive inhibition.

Table 6.5: Order-of-Magnitude analysis of the rate of a biochemical reaction with noncompetitive inhibition.

Table 7.1: Contributions of chain (i.e., non-ring) groups to the Gibbs Energy

Table 7.2: Contributions of ring groups to the Gibbs Energy

Table 7.3: Calculation of the Gibbs Energy of crotonate from the contributions of groups.

Table 7.4: Calculation of the Gibbs Energy of fructose diphosphate aldolase from the contributions of groups
Table 7.5: Data points on the Gibbs Energies of Formation of compounds in aqueous solution, used in the estimation of the contributions of functional groups.

Table 7.6: Data points on reactions used in the estimation of the contributions of functional groups.

Table 8.1: Result of dimensionless maximization of the rate for an ordered mechanism with 1 reactant and 1 product.

Table 8.2: Result of dimensionless maximization of the rate for an ordered mechanism with 1 reactant and 2 products.

Table 8.3: Result of dimensionless maximization of the rate for an ordered mechanism with 1 reactant and 3 products.

Table 8.4: Result of dimensionless maximization of the rate for an ordered mechanism with 2 reactants and 1 product.

Table 8.5: Result of dimensionless maximization of the rate for an ordered mechanism with 2 reactants and 2 products.

Table 8.6: Result of dimensionless maximization of the rate for an ordered mechanism with 2 reactants and 3 products.

Table 8.7: Result of dimensionless maximization of the rate for an ordered mechanism with 3 reactants and 1 product.

Table 8.8: Result of dimensionless maximization of the rate for an ordered mechanism with 3 reactants and 2 products.

Table 8.9: Result of dimensionless maximization of the rate for an ordered mechanism with 3 reactants and 3 products.

Table 8.10: Ranges for parameters pertinent in the maximum-rate methodology.

Table 8.11: Estimation of important parameters for intermediates of the glycolytic pathway.
Table 8.12: Estimation of maximum-rate parameters for some steps of the glycolytic pathway.

Table 9.1: Transformation of the stoichiometric constraint placed on a reversible reaction, into stoichiometric constraints on the forward and backward portions into which the reaction is decomposed

Table 9.2: Abbreviations of the names of metabolites that will be used in the pathways for serine and alanine

Table 10.1: Abbreviations for metabolic intermediates used in bioreaction networks

Table A.1: Description of files in the DATA MANIPULATION module, for editing the database.

Table A.2: Description of files in the DATA APPLICATION module, for estimation of maximum rates of bioreactions and synthesis of biochemical pathways.

Table A.3: Description of files in the DATA STORAGE module, for saving the database.

Table A.4: Description of files in the DRAWING module, for drawing pathways.

Table A.5: Description of files in the MAGNITUDE module, for Order-of-Magnitude reasoning.

Table A.6: Commands for the manipulation of the database

Table A.7: Data Stored for Groups.

Table A.8: Data Stored for Molecules.

Table A.9: Data Stored for Reactions.

Table A.10: Data Stored for Pathways.

Table A.11: Flavor Instance Variables for Groups.

Table A.12: Flavor Instance Variables for Molecules.

Table A.13: Flavor Instance Variables for Reactions.

Table A.14: Flavor Instance Variables for Pathways.

Table A.15: Flavor Methods for Groups.
Table A.16: Flavor Methods for Molecules.

Table A.17: Flavor Methods for Reactions.

Table A.18: Flavor Methods for Pathways.

Table A.19: Example of an instantiated molecule. Instance variable values of FRUCTOSE-DIPHOSPHATE.

Table A.20: Example of an instantiated reaction. Instance variable values of FRUCTOSE-DIPHOSPHATE-ALDOLASE.

Table A.21: Example of an instantiated pathway. Instance variable values of GLC→PYR, which represents the glycolytic pathway.
1

PREFACE

This preface serves as a simple reader's guide to the thesis, giving an overview of the material covered chapter by chapter.

Chapter 2 is a general introduction to the problem of design of biochemical pathways. Sections 2.2, 2.3, and 2.4 are important; they describe the problem attacked, the philosophy of the approach followed, and the goal of the work.

Chapter 3 gives a summary of the work performed, which is described in detail in Chapters 5 through 10. Chapters 5 through 8 focus on methodologies developed for the analysis of biochemical pathways, while Chapters 9 and 10 focus on the synthesis of biochemical pathways.

Chapter 4 attempts to describe the bigger picture of which this research is only a small part. Thus, it discusses issues related to a set methodologies (and their computer implementation) that would achieve effective and efficient design of
bioprocesses at the biochemical level. This chapter is useful to put the completed work in perspective, but it is not necessary for understanding the rest of the thesis.

Chapters 5 and 6 describe the O[M] formal system for reasoning with orders of magnitude and approximate relations. This system was developed to achieve efficient representation of, and reasoning with, semi-quantitative and qualitative knowledge. Chapter 5 focuses on the formalism itself and some simple applications to non-biochemical process systems, while Chapter 6 discusses the formalism's applications in the analysis of biochemical systems.

Chapter 7 provides a group-contribution method which was developed for the estimation of equilibrium constants of biochemical reactions.

Chapter 8 details a methodology developed for the estimation of the maximum rate of any enzymatic reaction. The method is based on the physicochemical rate of encounter between the enzyme and each substrate or product of the reaction.

Chapter 9 describes an algorithm which synthesizes all biochemical pathways satisfying a set of stoichiometric specifications. This algorithm is the primary tool for the synthetic part of pathway design.

Chapter 10 involves a case study on the production of lysine from glucose. A framework for iterative pathway design, utilizing the methods of Chapters 7 through 9, is outlined in the context of this case study.

Chapter 11 briefly reiterates the conclusions of this work, the significance of the work as a whole, and the significance of the individual methodologies developed. It also gives some of the important future directions that should be followed.
CHAPTER 2

OVERVIEW
2.1. SUMMARY

This work consists of a combination of methods from Artificial Intelligence, Mathematics, and Physical Chemistry, addressing the problem of design of metabolic pathways, covering both the development of a new pathway for the production of a bioproduct and the evaluation of an existing pathway.

A database for metabolites and bioreactions contains information needed by various methodologies. Candidate pathways can be generated by a constrained search algorithm, which is based on the iterative satisfaction of constraints through the transformation of a base-set of pathways (Chapter 9). Generated pathways can be screened by heuristic rules and detailed analytical examination. Semiquantitative knowledge can be used within the O[M] formalism for Order-of-Magnitude reasoning, which captures engineering commonsense concepts about the relative orders of magnitude of parameters (Chapters 5 and 6).

Additional methodologies make up for the absence of critical data. They include the estimation of maximum rates for biochemical reactions based on collision limitations (Chapter 8), and a group-contribution method for the estimation of equilibrium constants (Chapter 7). Using these parameters, thermodynamic and kinetic feasibility analysis of pathways can be accomplished, and the bottlenecks of the pathway can be identified. The bottlenecks can be bypassed by synthesizing appropriate modifications to the pathway (Chapter 10).
2.2. THE PROBLEM OF PATHWAY DESIGN

2.2.1. MOTIVATION

When investigating the production of a bioproduct through a biochemical process, one has to select among various metabolic pathways, which may or may not be possessed by existing microorganism strains. One must evaluate each pathway's merits based on its intrinsic characteristics and the interactions of the pathway with the rest of the cell metabolism.

2.2.2. GENERAL AIMS

This project has two general aims, which correspond to the analysis part and the synthesis part of the design task at hand. We would like to develop systematic methodologies that:

(a) Analyze given biochemical pathways, determining potential production rates and yields, as well as production bottlenecks.

(b) Synthesize alternative pathways, including completely new pathways, which have the potential to satisfy process specifications.

2.2.3. SPECIFIC GOALS

These general analysis and synthesis aims, along with other characteristics of the problem and the sparse and qualitative nature of the available knowledge (Section 2.3.1), set the specific goals of this work:
(a) Development of methodologies which can exploit qualitative or sparse knowledge, since such knowledge is usually available about biochemical systems. Artificial intelligence methods often offer this capability.

(b) Development of methodologies for the approximate estimation of important pathway parameters, through extrapolation from existing data and/or based on first principles.

(c) Development of synthesis methodologies which can generate all metabolic pathways that lead from specified substrates to specified products and satisfy other constraints.

(d) Integration of the above methodologies into a unified framework for systematic engineering of biochemical pathways.

The goals of the project have been met, yielding several general methodologies, which will be discussed in more detail below. These methodologies do not model just a particular pathway (or biochemical network), nor do they design a particular biochemical system. They are applicable in the analysis and synthesis of metabolic pathways in general.
2.3. PATHWAY ANALYSIS

2.3.1. NATURE OF THE PROBLEM

Pathway analysis entails examining a given biochemical pathway and producing information about the pathway's potential performance, critical parameters, and bottlenecks.

Analysis is the basis for an accurate evaluation of the pathway, especially in comparison to other candidate pathways. It also leads to recommendations on how an existing or designed pathway could be improved, and what other information about the pathway should be acquired.

2.3.2. CHARACTER OF THE KNOWLEDGE

The analysis of a pathway must be based on relevant biochemical data and knowledge. However, a lot of the data one would like to have on a biochemical system are simply not available in practical applications. The information that is available is often qualitative in nature. Traditional approaches for the modelling and analysis of biochemical systems, based on quantitative mathematical analysis, are not applicable because of these characteristics of the knowledge.

Thus, the main obstacle in pathway analysis is the sparseness and qualitiveness of information on biochemical systems, requiring special efforts in the development of methods for:

- Extension of incomplete knowledge
- Representation and use of qualitative knowledge
2.3.3. GROUP-CONTRIBUTION METHOD FOR THE ESTIMATION OF EQUILIBRIUM CONSTANTS OF BIOCHEMICAL REACTIONS

In order to evaluate the feasibility and reversibility of a biochemical pathway, one must know the equilibrium constant of each bioreaction. Unfortunately, the equilibrium constants of most bioreactions are not known.

To overcome this limitation, we developed a group-contribution method for the estimation of Gibbs Energies of Formation of biochemical compounds in aqueous solution. The equilibrium constant of a bioreaction can be calculated from the Gibbs Energies of its substrates and products.

The contributions of all functional chemical groups of interest were computed, through linear regression, from estimated and experimental data on Gibbs Energies and Equilibrium Constants, collected from the literature.

The equilibrium constants are used in the evaluation of the thermodynamic feasibility and reversibility of a biotransformation (which could be a metabolic pathway or a single bioreaction), providing a rough characterization of the transformation as favorable, marginally favorable, or unfavorable.

2.3.4. COLLISION-LIMIT BASED ESTIMATION OF MAXIMUM RATES FOR ENZYMATIC REACTIONS

To evaluate the potential production rate of a pathway and to identify its kinetic bottlenecks, one needs to know the kinetics of each enzyme in the pathway. Detailed, reliable enzyme kinetics are rarely available, and very few enzymes have been examined in more than one microorganisms. As the properties of an enzyme depend on the source,
the estimation of the actual rate of a given pathway is, for all practical purposes, impossible.

To overcome this formidable obstacle, we developed a methodology for the estimation of the maximum allowable rate for any enzymatic reaction (discussed in detail in Chapter 8).

The strategy of the method consists of using a typical fast mechanism for the bioreaction and stating constraints stemming from equilibrium and the fact that steps which involve bimolecular reactions cannot take place faster than the two species collide in the solution. The method uses these constraints on the steps of the mechanism to estimate a maximum rate for the overall bioreaction; this maximum is tighter (i.e., lower) than the rate which could be obtained by looking at a single collision-rate constant.

For example, for the bioreaction \( A + B \rightarrow P + Q \), with equilibrium constant \( K_{eq} \), an ordered mechanism catalyzed by the enzyme \( E \) has the steps:

\[
\begin{align*}
E + A & \rightarrow EA, \quad \text{with rate constants } k_1 \text{ and } k_{-1} \\
EA + B & \rightarrow EAB, \quad \text{with rate constants } k_2 \text{ and } k_{-2} \\
EAB & \rightarrow EQ + P, \quad \text{with rate constants } k_3 \text{ and } k_{-3} \\
EQ & \rightarrow E + Q, \quad \text{with rate constants } k_4 \text{ and } k_{-4}
\end{align*}
\]

The equilibrium constraint has the form:

\[
K_{eq} = \frac{k_1 k_2 k_3 k_4}{k_{-1} k_{-2} k_{-3} k_{-4}} \quad (2.1)
\]
The collision-limitation constraints refer to the bimolecular steps only; the parameters \( k_1, k_2, k_3, \) and \( k_4 \) must be smaller than the collision constants between the species \( E \) and \( A \), \( EA \) and \( B \), \( EQ \) and \( P \), and \( E \) and \( Q \) respectively.

The rate expression, which can be derived as a function of the kinetic parameters and the concentrations, is then maximized subject to the above equality and inequality constraints, yielding the maximum allowable rate of reaction. The required mathematical analysis has been carried out for a large number of bioreaction mechanisms of interest.

With the results of this method, the kinetic efficiency of a whole pathway can be evaluated, based on the minimum amount of total enzyme required divided by the achieved rate. The methodology can thus be applied to reject proposed mechanisms or pathways as unable to meet rate requirements. The maximum rates of the steps of a pathway also point to kinetic bottlenecks (Chapter 10) which must be removed if the pathway is to be improved.

While in its basic form the method estimates a rate, bounds on other parameters (such as concentrations) can also be estimated if the rate is known or constrained by design decisions.

### 2.3.5. ORDER-OF-MAGNITUDE REASONING

One of the stated goals of the work is to take advantage of available qualitative or semiquantitative knowledge. To achieve the use of such knowledge by a computer, a formal system that represents semiquantitative concepts is needed. The O[M] Order-of-Magnitude formalism (discussed in Chapters 5 and 6) was developed to meet this need.
O[M] reasons with approximate relations among parameters. It is based on seven primitive relations that can hold between two quantities, A and B (which have the same physical dimensions):

- A is much less than B
- A is moderately less than B
- A is slightly less than B
- A is exactly equal to B
- A is slightly larger than B
- A is moderately larger than B
- A is much larger than B

The formalism can perform inferences by manipulating these and other order-of-magnitude relations, algebraic constraints, and if-then rules. The formalism also provides a mechanism for stating goals and evaluating alternative assumptions.

O[M] facilitates the acquisition and use of semiquantitative knowledge about biochemical systems (Chapter 6) and the formalization of important tasks, such as the identification of rate-limiting steps (Section 6.2). O[M] can also assist in other tasks (such as the analysis of fluxes in a metabolic network, or the qualitative analysis of enzyme inhibition, as discussed in Section 6.1) and is useful in reasoning about other process engineering systems as well (such as chemical reactors and heat-exchangers, as discussed in Section 5.10).
2.4. PATHWAY SYNTHESIS

In the synthesis of biochemical pathways, the goal is to generate metabolic pathways that can achieve a desired transformation (Chapter 9). For synthesis purposes, metabolic pathways are simply linear combinations of bioreactions, not restricted to correspond to established or existing pathways functional in an integrated fashion.

2.4.1. IMPORTANT CHARACTERISTICS OF THE PROBLEM FORMULATION

There are some important points to note on the way the synthesis problem is formulated:

(a) We do not restrict ourselves to pathways that are already known to exist in some microorganism. Instead, we allow the construction of pathways even from bioreactions taken from different microorganisms, since genetic engineering can in fact realize such pathways.

(b) All bioreactions are assumed to be reversible, unless there is quantitative information to the contrary (e.g., from our group-contribution method — Section 2.3.3 and Chapter 7). This increases the number of candidate pathways.

(c) In the strictest of the alternative formulations of the problem (Section 9.1.1), a pathway must describe the complete stoichiometry of a bioprocess. It must consume substrates that are fed to the microorganisms, and produce final products and
by-products. It is not sufficient to provide only that specialized section of the pathway which leads from various common metabolic intermediates to the particular bioproduct of interest.

In a more formal statement of the problem (Section 9.2), we want to identify candidate metabolic pathways satisfying requirements which state: (a) which metabolites are necessary as reactants, and which are prohibited; (b) similarly, necessary and prohibited products; (c) similarly, necessary and prohibited intermediates and enzymes; and (d) minimum acceptable thermodynamic driving force for the bioreactions.

2.4.2. ALGORITHM

The problem was solved by the development of an algorithm for the construction of all candidate pathways, based on the recursive satisfaction of requirements (Section 9.3).

At each step, this technique operates on one of the requirements and a set of partial pathways which satisfy the requirements that have already been examined. The set of partial pathways is transformed so that they satisfy the requirement at hand; this is accomplished by combining partial pathways into new pathways and eliminating some of the old pathways. The procedure is recursively repeated until all the requirements are satisfied. The procedure produces minimal, genotypically independent pathways that satisfy all the imposed stoichiometric requirements.
2.4.3. APPLICABILITY

This methodology is useful not only for synthesizing whole metabolic pathways for the production of bioproducts but also for designing modifications to a starting pathway (Chapter 10) to achieve removal of a bottleneck.
2.5. INTEGRATED FRAMEWORK

Each of the methods described above has been shown to work, and has general value independently of all the others, and independently of the exact context of the use of the method in this project. The interactions and the complimentary capabilities of the methods are nonetheless important for their effectiveness in an integrated pathway-design framework (Chapter 10), which is delineated below:

- Given specifications on the bioprocess – its stoichiometry, production rates, thermodynamics, competition among pathways, potential inhibition, etc. – the synthetic methodology (Sections 2.4, and Chapter 9 in more detail) generates alternative pathways satisfying the stoichiometric specifications imposed. If the task is not as stated above, but rather the analysis of a single given pathway, then this step is skipped and all subsequent steps are applied to the pathway at hand, rather than a set of candidate pathways.

- The group-contribution method (Section 2.3.3, and Chapter 7 in more detail) estimates the equilibrium constants of all reactions involved, and discriminates among strongly favorable, marginally favorable, and infeasible pathways. Based on this criterion, as well as simpler ones (e.g., number of reactions and intermediates) promising pathways can be examined in more detail.

- The maximum kinetic efficiency of the conversion (Section 2.3.4) can be evaluated through the estimation of the maximum rate of each enzymatic reaction (Section 2.3.4, and Chapter 8 in more
detail). This allows further reduction of the number of candidates to those that are kinetically more productive.

- Each of the remaining pathways can be analyzed in more detail to find thermodynamic bottlenecks (Section 2.3.3), where bioreactions are marginally feasible, and kinetic bottlenecks (Section 2.3.4, case study in Chapter 10), where the maximum rate of reaction is much lower than in other parts of the pathway.

- The pathways can be examined through Order-of-Magnitude reasoning (Section 2.3.5, Chapters 5 and 6).

It is only in the last phase that the need for specific, detailed knowledge about the particular pathway analyzed becomes pressing. For the previous methods, little more than the stoichiometry of the enzymatic reactions is needed, because of the extensive database from which estimation methods were developed (Sections 2.3.3 and 2.3.4, Chapters 7 and 8).

The results of the Order-of-Magnitude analysis depend very strongly on the specific available knowledge, but the earlier steps of the approach prune out many possibilities and allow us to focus our attention to a much smaller number of reactions and pathways, about which information must be sought.
CHAPTER 3

INTRODUCTION
3.1. THE DOMAIN: BIOTECHNOLOGY

3.1.1. BIOCHEMICAL PATHWAYS

The biomolecules of which living organisms are composed conform to the familiar laws of chemistry, but they also interact with each other in accordance with a set of principles comprising the metabolic logic of the living state [Lehninger, 1982]. Cells function as organized chemical engines because they possess enzymes, macromolecular catalysts capable of great specificity and rates much higher than the rates of non-enzymatic reactions.

The coordination of the extensive network of biochemical reactions is achieved through regulation of the concentrations of enzymes (genetic control) and the specific activities of enzymes (substrate and product inhibition, allosteric control). In genetic control, the production rate of an enzyme is affected by the presence or absence of regulators, which usually are small molecules. In allosteric control, the activity of already produced enzyme is further affected by the presence or absence of other small molecules.

Single enzyme-catalyzed steps in succession form long chains, called pathways, achieving overall transformations of substrates to far removed products. The definition of a biochemical pathway is not exact. There is always interaction among pathways; after all, what is the use of an isolated pathway in a cell? In practice, a pathway is usually construed as a (minimal) set of reactions such that there is a distinct function (purpose) of the pathway, regulation interactions are contained within it, and connections with other pathways occur at a minimal number of intermediate compounds (but freely at its start and end metabolite).
3.1.2. BIOPROCESSES

Short sections of pathways can be employed in bioconversions. These are the bioprocesses in which a substrate is converted by a microorganism to a structurally similar product. The advantages of bioconversions over chemical conversions are [Wang et al., 1979]:

- Great specificity and stereospecificity
- High yield on substrate
- Mild reaction conditions (pH, temperature)
- Activation of normally unreactive positions of a molecule
- Coupling in one step of several chemical transformations

Fermentations, on the other hand, are bioprocesses in which the substrates and the products differ greatly. In a fermentation, the pathway of the conversion normally crosses and interacts with the rest of the organism's metabolism. The product can in fact be a macromolecule produced from small molecules. Fermentations with complex pathways cannot be compared to chemical processes at all, since industrial implementation of such large numbers of steps is practically infeasible.

In the rest of this document we will not make any distinction between fermentations and bioconversions, and we will most often use the latter term to cover all processes carried out by microorganisms.

3.1.3. MUTATION-SELECTION PROGRAMS

Novel bioprocesses or improvements in the productivity of existing bioprocesses are achieved through alterations in the genetic material of microorganisms. These
alterations are called mutations and change drastically the quantity or activity of an enzyme present in the microorganism, under some or all conditions. They can, for example, block a step in a pathway that consumes a desired product, or prevent the regulation of the pathway that produces the desired product and allow higher rates or yields.

Mutations happen spontaneously in nature at very low rates which, for bacteria, are of the order $10^{-9}$ mutations per cell per replication per gene [Snyder et al., 1985]. Since mutations are random events, there is no way of knowing a priori when or where a mutation will occur.

However, with the use of special agents called mutagens, high rates of mutation can be induced. In biotechnology, one can thus embark on mutation-selection programs, in which massive mutation steps are followed by evaluation of the surviving mutants, and selection of the most promising ones for further mutations.

Several years may be required for a mutation-selection program, but improvements by a factor of 100 or more can be achieved [Crueger and Crueger, 1984]. The yield of penicillin, for example, has been improved by a factor of 50,000 [Vanek and Hostalek, 1986]. The speed of a program can be enhanced with modern automatic equipment, such as the Glaser "dumb waiter" which can inoculate 10,000 large agar plates with $10^8$ cells, monitor growth, and assay 85 different parameters [Crueger and Crueger, 1984].

3.1.4. GENETIC ENGINEERING

More systematic improvement of strains can be achieved by a set of new laboratory techniques, collectively called genetic engineering [Glover 1984, Old and
Primrose, 1985]. Genetic engineering allows the introduction and integration of foreign genetic material into a cell, achieving the transfer of bioreactions and pathways from one microorganism to another.

For example, a very slow enzyme limiting a pathway can be replaced by a faster one from another source, and even a totally missing step can be inserted. In the latter case, creation of a totally new biochemical pathway (and hence a new bioprocess) is accomplished.

An important factor determining the success of genetic engineering is the availability of knowledge on the mechanism of expression of the enzymes involved. For amino acids, where such knowledge is available, progress has been made in cloning even entire pathways [Old and Primrose, 1985].

Genetic engineering introduced enormous power in the realization of true synthesis (i.e., construction from pieces) of biochemical pathways and processes. However, from the point of view of biochemical engineering, no methods for the planning or evaluation of such improvements have been applied, and most of this power is not used systematically.

Genetic engineering would be more focused and targeted if there were methods which could set suitable goals, i.e., designs of pathways with good ideal properties. This is exactly the role that this work aims to explore.
3.2. THE PROBLEM: DESIGN OF BIOCHEMICAL PATHWAYS

3.2.1. SYNTHESIS OF BIOCHEMICAL PATHWAYS

When investigating the production of a chemical through a biochemical process, one is faced with a synthesis problem:

Given the desired product,

1. Select an appropriate microorganism strain

2. Design desired genetic modifications to that strain

3. Select a set of substrates to use

4. Estimate good conditions of temperature, pH, and concentrations.

To give a rough estimate of the design space, the number of possible starting strains is anywhere from approximately 10 to several hundreds or thousands. The potential genetic modifications are normally a few dozen, if we include only those that seem to have an effect on the process performance. The potential substrate sets, if we insist on known substrate sets for the strain at hand, could be only a couple or as high as several dozen. The conditions of the process cannot be counted, since they involve continuous variables.

We will not address further the selection of conditions, because it depends heavily on very specific experimental data on the chosen strain grown on the selected substrates.
The first three design subproblems, on the other hand, are not amenable to exhaustive experimental investigation, because the design space is very big and the experimental procedures very costly and lengthy. We consider the experimental investigation useful only to resolve problems at a late stage, when most alternatives have been ruled out, and the nature of the required experimental data has been determined.

It is exactly this preliminary investigation that we intend to tackle. In effect, when a bioprocess with a given goal product is sought, we would like to be able to:

1. Reduce the alternative pathways to as small a number as possible

2. Determine what data are required for the completion of the design of the pathway for the process

3. Suggest preliminary solutions and their merits

In the next sections, we will examine characteristics of this preliminary design problem, the requirements introduced by the character of available knowledge, and the kinds of methods that satisfy the requirements.

### 3.2.2. CRITICAL SYNTHESIS ISSUES

Synthesis is always a creative activity, substantially different from activities that are analytical in nature. We can cast the description of the problem of synthesis of biochemical pathways in terms a more general design framework [Mostow, 1984, Mostow, 1985]. We want here to construct a structure that:

- Satisfies a functional specification, i.e., produces the desired product from the given substrates (if that is required).
• Conforms to the limitations of the domain. For example, if the pathway is composed of segments from different microorganisms, it should be biologically feasible to clone it.

• Meets implicit and explicit requirements on performance and resource usage, such as:
  ◦ Kind of allowable substrates
  ◦ Absence of undesired by-products
  ◦ Specified yield on substrate
  ◦ Specified productivity

• Satisfies implicit or explicit design criteria on the form of the structure. For example, constraints on:
  ◦ The number of bioreactions in the pathway
  ◦ Number of branches
  ◦ Stability of the pathway
  ◦ Presence or absence of certain forms of regulation

• Satisfies restrictions on the design process itself, such as length or cost of the design process below some limits.

As with most design processes, a general framework or model of biochemical pathway synthesis has not been developed — even the design activity itself is very young. The very task of synthesis of biochemical pathways has only been pragmatically infeasible until about 1980, and has never been performed in a truly systematic fashion.
The standard practice today is to simply pick a pathway from a random microorganism possessing it and clone it to another microorganism, without investigating many alternative pathways, microorganisms, or substrates, unless the first attempt fails (e.g., when the enzymes are not expressed).

Hence, design strategies are unavailable, and even case studies are nonexistent. This is the first obstacle we encounter, but the second one is even more severe. Design is normally iterative, requiring the evaluation of partial or completed solutions. We will see in the next sections that the analysis of biochemical pathways under realistic application conditions, i.e., using only the knowledge normally available to the designer, has not been successfully tackled.

3.2.3. ANALYSIS OF BIOCHEMICAL PATHWAYS

In the analysis of biochemical pathways, we are given the microorganism strain, the substrates, and the bioproducts of a process, and we must determine the performance of the process, the key mechanisms and parameters affecting the performance, and potential improvements to the process.

We can refine these goals further:

(1) The performance of the process. This includes:

◊ A qualitative at least characterization of the yield (over substrate, biomass, etc.)

◊ The maximum achievable performance with the same pathway structure, but possibly changed regulation [Eroshin and Minkevich, 1982]
Maximum performance with the same
strain, but possibly changed substrates

(2) Key mechanisms and variables. First the dominant
pathway leading from substrate to product, along with the dominant
competing pathways, must be identified. The rate-limiting steps of
the dominant pathways must be determined so that any unknown
controlling mechanisms may be conjectured. Some qualitative
judgement of the effects of each bottleneck (i.e., how much it
constrains the process) should finally be reached.

(3) Potential improvements. These are possible changes of
substrate and modifications of the microorganism strain. Once the
effects of the bottlenecks are determined above, this step entails more
concrete suggestions for biologically feasible ways to remove the
bottleneck.

3.2.4. CRITICAL ANALYSIS ISSUES

To even consider the bioprocess analysis problem, a working model of the
pathways used and the metabolism of the microorganism in general is needed. It is more
suitable to think of this as a model of our knowledge about the biochemical structures of
interest. Modelling our knowledge well is required in order to use the knowledge
efficiently. As we mentioned before (Section 3.2.2), in the analysis part of a design
process, application of knowledge about the components of the designed structure and
their interactions, is necessary to guide the next synthetic step, i.e., to improve the
design.
Most of the experience in the analysis of complex systems (especially with Artificial Intelligence techniques) comes from the design of electrical circuits [Kelly and Steinberg, 1982, Mitchell et al., 83, Hamscher and Davis, 1984, de Kleer, 1984, Davis, 1984]. There are two fundamental differences between circuits and biochemical pathways — which are in fact differences between electrical circuits and most physical systems of interest in chemical or biochemical process engineering:

(1) Knowledge is always available on all the components of an analyzed circuit device, or for all the components that can be incorporated into a synthesized device. In a biochemical pathway, on the other hand, data on bioreactions are sparse and frequently qualitative in nature.

(2) The behavior of a digital circuit depends heavily on its exact infrastructure; in fact, digital circuits are built that way. Bioreaction networks on the other hand are built to be robust and insensitive to internal details, partly because internal details cannot be accurately controlled.

It is a distinctive characteristic of knowledge about biological systems, that it is seldom complete. In most cases it is in fact very sparse, i.e., most of the relevant factors or data are unknown. It is estimated that we know very little about 95% of the bioreactions [Dixon and Webb, 1979], let alone individual enzymes — the total number of different proteins in nature exceeds $10^{11}$ [Lehninger, 1982].

Furthermore, the knowledge that is available is normally qualitative or at best semiquantitative in nature. Because of the sparseness and qualitativenss of the knowledge, traditional model forms, i.e., differential or algebraic equations [Rapoport et al., 1976, Heinrich et al., 1977, Shuler et al., 1979, Esener et al., 1982, Okamoto and
Hayashi, 1985] will not work, because (as we will see in Section 3.2.5) in practice we do not have the parameters they require. But one cannot wait until all the information is gathered, because important decisions must be made in the design of bioprocesses and are currently made unsystematically.

Fortunately, because of the independence of the behavior of a pathway from many internal details, it is in principle possible to apply a combination of techniques from Artificial Intelligence (AI) and the Biochemistry domain to:

(1) Employ qualitative knowledge in the derivation of qualitative conclusions

(2) Estimate approximate values or bounds for unknown parameters

In summary, the important difficulty in the evaluation of biochemical pathways lies in the incompleteness and qualitatively of the knowledge, requiring some unconventional methodologies to pragmatically tackle the task. In the subsection below we further refine our arguments about conventional biological models. In the section following that, we will discuss the character of the proposed Artificial Intelligence approach.

3.2.5. CONVENTIONAL BIOLOGICAL MODELS

We briefly review here the modelling methodologies for biochemical systems, to examine their potential applicability in pathway design. We will argue that traditional models are in fact inadequate for the evaluation, improvement, and comparison of biochemical pathways.
Because of the complexity of the biochemical mechanisms of a cell, a key choice that must be made in modelling cell metabolism is the selection of the appropriate level of cell-activity lumping. Several other pairs of choices occur in the selection of a modelling approach:

- Qualitative vs. Quantitative
- Descriptive (black box) vs. Predictive (grey box)
- Structured vs. Unstructured
- Deterministic vs. Stochastic
- Continuous time vs. Discrete time
- Distributed vs. Segregated

Not all the combinations of choices have been attempted and we will in fact classify the models in only five major categories:

1. Fermentation equations
2. Unstructured models
3. Population models
4. Structured culture models
5. Single-cell models

In the paragraphs below we will describe briefly the character of each kind of model, and conclude with a discussion of the reasons these models are not, in practice, suitable for the preliminary design of biochemical pathways.
3.2.5.1. Fermentation Equations

The growth of a culture on a substrate with formula \( CH_mO_l \) and producing products with formulae collectively represented as \( CH_rO_sN_t \) can be represented by a lumped chemical reaction. Under aerobic conditions [Oner and Erickson, 1983] this fermentation reaction is:

\[
CH_mO_l + a\ NH_3 + b\ O_2 \rightarrow \yc\ CH_pO_nN_q + z\ CH_rO_sN_t + c\ H_2O + d\ CO_2 \tag{3.1}
\]

while under anaerobic conditions the \( O_2 \) reactant is absent.

Fermentation equation models aim at interrelating the various parameters of the equation via a (typically underdetermined) set of equations and inequalities. The relations are assumed to hold for any arbitrary time interval during the fermentation, and they are sometimes expressed in terms of rates rather than absolute amounts [Yamane and Shiotani, 1981].

The analysis is based largely on the stoichiometries of the biochemical pathways involved in the transformation and assumptions on the composition of the biomass and the energy efficiency of the microorganism.

The fermentation equations can be used to estimate maximal product yields [Papoutsakis, 1984], when the full pathways for the consumption of substrates and production of products are known.

3.2.5.2. Unstructured Models

Unstructured models ignore completely any variations of the biomass composition in response to environmental changes [Esener et al., 1983]. They relate
the concentration of the biomass, \( X \), to the concentrations of substrates, \( S_i \), and sometimes product or inhibitor concentrations. There are great variations in the complexity of the algebraic expressions used in these models. Typically, the empirical parameters in the expressions admit no significant physical (or mechanistic) interpretation.

The attractive feature of unstructured models is their simplicity and their compatibility with macroscopic balances [Esener et al., 1983] and fermentation equations.

Their disadvantages are their lack of mechanistic interpretation and narrow applicability. Unstructured models can only be used for interpolation or moderate extrapolation of experimental data on the growth rate of microorganisms. They cannot be used, however, in the evaluation of alternative pathways, because they essentially lump all of the metabolism (and all its possible alternative forms) into a single transformation.

3.2.5.3. Cell-Population Models

These are models that describe statistical behavior of sets of cells, in terms of their age and cycle phase distributions. Within the classification of Section 3.2.5, they are quantitative, stochastic, and segregated.

To describe the behavior of cell populations, they attempt to portray, at either an abstract or a simple mechanistic level, the kinetic progression of cells along the cell cycle between one division and the next division. They can adopt either a discrete time approximation [Hagander, 1980] yielding difference equations, or a continuous time approximation [Nishimura and Bailey, 1980] yielding differential equations. It has
been shown that the two approaches are equivalent [Sundareshan and Fundakowski, 1984].

There have been several models proposed [Alberghina and Mariani, 1978, Bertuzzi et al., 1981, Kuzcek, 1984] often very complex mathematically [Kimmel, 1980, Kimmel, 1981]. Experimental verifications were carried out with synchronized cultures [Agar and Bailey, 1982, Grdina et al., 1984, White et al., 1984] or normal populations [Bailey et al., 1979]. The latter is more difficult, but the former may lead to wrong conclusions, since the cultures may be severely affected by the synchronization procedure [Ingram et al., 1983].

Population models combine a simple structured approach to segregated multi-cell populations. In effect, they combine characteristics of structured culture models and single-cell models, that will be examined in later sections. They focus on population dynamics, and their applications in biotechnology are limited, because the distribution of the cells along the cell cycle and other variations among cells do not play an important role in actual fermentations.

3.2.5.4. Structured Culture Models

Severe limitations of unstructured models appear when a transient occurs or when there is interest about a particular fraction of the biomass [Harder and Roels, 1982]. Structured models were introduced to remedy the difficulties, by distinguishing among selected components in the biotic phase and thus taking into account the state and past history of the biomass.

Structured models differ from population models in that their focus is on the behavior of biomass as a whole based on its intrinsic mechanisms, rather than the
statistical combination of individual cells at a more abstract level. However, some of their continuum parameters can be related to distribution moments of population models [Harder and Roels, 1982].

Structured culture models can be further subdivided:

- *Universal compartmentalization* models attempt to split the biomass into a few generally important compartments or components, based on structural and functional similarities.

- *Critical compound distinction* models pick out of the biomass a few individual compounds (macromolecules and/or small molecules) which play a very important role in a particular process despite their small concentration, but lump the rest of the biomass in an unstructured manner.

3.2.5.4.1. Universal compartmentalization

The most frequently distinguished compartments are the DNA, the mRNA, the stable RNA, the ribosomal protein, the remainder of the structural protein, the lipids, the carbohydrates, and the soluble pool (precursors).

Using this approach, after the compartments are decided upon, reactions are set for conversion of substrate to the compartments, interconversion among compartments, and turnover corresponding to maintenance requirements.

The concentrations of the compartments are the biotic state variables, and the substrate and product concentrations are the abiotic state variables. The temporal behavior of the system is described by a set of ordinary differential equations derived from mass balances.
Esener et al. [1982] presented a simple two-compartment model in which the K-compartment consists of carbohydrates, RNA, and the soluble pool, while the G-compartment contains the remainder of the biomass (i.e., protein, DNA, etc.). The external limiting substrate is assumed to be converted to K at a Michaelis-Menten rate. G is created from K at a rate first order in each of K and G. The turnover of G back to K is first order in G. The concentrations participating in kinetic expressions must correspond to the phase of the reaction [Harder and Roels, 1982, Roels and Kossen, 1978], which in this case is the biotic phase.

The resulting equations for substrate and biomass in a chemostat are similar to Monod expressions but they contain growth-rate dependent maintenance. Several other compartment models have been proposed [Roels and Kossen, 1978] having behavior similar to the above model.

Although compartment models represent a step forward from unstructured models, they suffer (to a lesser extent) from the same limitations caused by unrealistically lumped descriptions. Since they do not deal with, and distinguish among, structures of actual biochemical pathways, they cannot be of much use in the analysis and evaluation of pathways.

3.2.5.4.2. Critical compound distinction

Papageorgakopoulou and Maier [1984] proposed a structured model to account for lag phases, which occur when the substrate changes, and bang-bang behavior, occurring when multiple substrates are utilized. In this model, for every carbon source considered two critical enzymes are designated, one for transport and one for metabolism of the substrate. Both enzymes are assumed to follow Michaelis-Menten kinetics. The
production of the enzymes is inducible, associated with the growth-rate, and potentially subject to catabolite repression and/or allosteric inhibition.

Other models similar to the above have yielded good agreement with experiments [Chlam and Harris, 1982, Chiam and Harris, 1983]. An interesting class of cybernetic models [Kompala et al., 1984, Dhurjati et al., 1985] perform parameter selection according to an optimal (for the microorganism) strategy.

In critical compound distinction models there is considerably more realism than in universal compartmentalization models. By focusing on particular cases, they achieve parameters which are physically meaningful, can often be predicted a priori, and can always be tested experimentally.

Although the models distinguish critical compounds, they do not model in detail critical mechanisms, i.e., important biochemical reactions and pathways. Thus, these models still abstain from modelling metabolic pathways in sufficient detail to allow their evaluation. Thus, the improvement of a pathway or the comparison of a set of pathways are not aided by these models.

3.2.5.5. Single-Cell Models

Single-cell models describe the behavior of only one cell through several generations, under varying conditions. The behavior of cultures can be simulated by taking a number of cells with appropriately distributed ages.

Typically, single-cell models use structuring of the biomass into roughly 5 to 30 components and include regulation of cell division. The resulting mass balances are conceptually similar to the balances of structured culture models. The balances yield a system of a few dozen differential and algebraic equations. Most of the parameters of the
equations are given \textit{a priori} values; the remaining (few) parameters are fitted to experimental data.

Apart from some early work in the field [Heinmets, 1966] and some work of very limited breadth [Owe Berg and Ljunggren, 1982], there are two major research streams dealing with single-cell models:


One problem with single-cell models is that, in order to simulate industrial asynchronous cultures, a large number of (independent) cells must be simulated through several generations. This is computationally expensive and restricts the complexity that can be incorporated in single-cell models. The limitations are more severe in general-purpose models that introduce complexity uniformly, while special models need only use complex schemes for a small number of cell activities.

For the purposes of pathway design single-cell models do not substantially differ from structured culture models, because their segregated character is of no significance in pathway design. Single-cell models cannot be used for the comparison
of alternative pathways or for the analysis and improvement of a given pathway, because they do not model the structure of pathways (thus neglecting the important biochemical mechanisms that determine the performance of pathways).

3.2.5.6. Discussion

The usefulness of unstructured models and fermentation equations for interpolations, or simple and quick estimations, must be acknowledged. However, their indiscriminate lumping is only justified under the absence of more detailed knowledge. Clearly, there is such additional knowledge, e.g., providing the qualitative structure of pathways. From the point of view of pathway design, fermentation equations (which in fact take into account the structure of pathways) are always useful in the estimation of the maximum yield of a given pathway. In our work we will use an equivalent approach for estimating the yield.

At the other end, indiscriminate complexity employed by universal compartmentalization models (in their culture or single-cell form) is useful for a general understanding of the metabolism. However, for more detail on particular processes and pathway mechanisms, uniform complexity becomes unmanageable as well as undesired. Even with the distinction of critical compounds, the key mechanisms of interest (i.e., the internal structure of biochemical pathways) are not modelled. Thus, such models cannot be used for pathway evaluation.

One may wonder why models have not been directed at the description and evaluation of pathways. The main obstacle is that the kinetic parameters needed to model, quantitatively, a pathway within the metabolism are not available nor can they be obtained by fitting experimental data. Thus, detailed kinetic modelling of pathways has never been applied in practice.
An aspect neglected by all classes of models is the possibility of using qualitative knowledge along with quantitative data, justified by the fact that a large part of biological knowledge is qualitative in nature. Only primitive approaches to such a combination have been attempted [Dohnal, 1985], aimed at modelling bioreactors and not biochemical pathways.

In summary, the absence of the kind of data that are necessary for traditional modelling approaches has kept them away from detailed modelling and evaluation of biochemical pathways. Instead, models have resorted to lumping cell activities in a way that renders them useless for pathway design. Qualitative knowledge has been largely ignored, as it does not fit the quantitative nature of the models.
3.3. THE APPROACH

3.3.1. ARTIFICIAL INTELLIGENCE METHODS

In order to carry out the analysis of biochemical pathways we have to represent and use available knowledge. We resort partly to methods from Artificial Intelligence [Winston, 1984, Charniak and McDermott, 1985, Bibel, 1985], because:

- We aim to formalize reasoning strategies and structure the tasks of analysis and synthesis of biochemical pathways, moving them closer to automation.

- Most of the available knowledge is incomplete, qualitative or semi-quantitative in nature, well suited to the AI methodology and not traditional, strictly quantitative models.

The latter point is justified by the discussion in Section 3.2.5. While traditional models can use only quantitative knowledge (and in fact require a complete set of parameters to yield results), Artificial Intelligence methods are suitable for dealing with knowledge at other levels, such as boolean, qualitative, or order-of-magnitude knowledge.

However, AI is not a panacea and it does not completely relieve us from the obligation to possess and supply knowledge; other methods (Section 3.3.3) based on physicochemical principles are necessary to make some essential predictions of important parameters.

Caution must also be exercised in the selection of the appropriate AI methods. We will see below that some of the more popular AI methods, namely Rule-Based
Expert Systems (Section 3.3.1.1) and Qualitative Reasoning (Section 3.3.1.2), are in fact not useful at this stage of the project. For the needs of our problem, it is necessary to develop a new AI technique for Order-of-Magnitude reasoning, described in Chapters 5 and 6.

In the development of formal methodologies, one aims for representation schemes and reasoning procedures which possess desired properties. Such properties refer to the ability of the representation and/or procedure to:

- Capture all of a given solution space, or a well defined subspace of the solution space
- Guarantee correct results
- Possess favorable complexity (e.g., require time and/or space that is linear or polynomial in the size of the problem specification or the size of the solution space)

Such properties cannot always be achieved, but they help in the exploration of the characteristics of a problem-solving methodology. Through the investigation of properties of this kind, one can gain insight on the structure of the representation and the procedures. Based on this insight, one can compare the problem-solving methodology to other methodologies, and identify the important bottlenecks of the methodology. This, in turn, can lead to significant improvements or extensions of the methodology.

3.3.1.1. Rule-Based Expert Systems

The technology of Rule-Based Expert Systems (RBES) involves encoding knowledge in experiential (empirical) production rules of the form:
IF <antecedents> THEN <consequents>  \hspace{1cm} (3.2)

An inference engine can then run the rules in either of two directions:

- Forward chaining, from the initially supplied information towards results
- Backward chaining, from goals towards antecedents that are known to hold

RBES lack some of the disadvantages of traditional modelling, but cannot perform satisfactorily in our case. Davis [Davis, 1986] provides a list of criteria for the selection of problems amenable to a RBES approach (Table 3.1). Many important criteria are not met by our problem. Most notably:

- The domain of the application is not sufficiently narrow.
- The task is not currently performed satisfactorily (so that all references to "recognized experts" doing "better than the amateurs", or to the duration of the task "between a few minutes and a few days" are vacuous).
- The skill is not routinely taught.
- There are no case studies available. Case studies are always the basis of development of RBES.

For these reasons, we have not used Rule-Based Expert Systems in any part of the research performed.

However, the RBES technology might be useful in future work, within the framework given in Chapter 4. In that framework, there may be particularly confined
subtasks which do satisfy the criteria, and could be carried out by rule-based software modules.

Subtasks attacked through RBES should be very well specified and only loosely dependent on the rest of the tasks. They must be small enough to be handled by a few dozen rules, because rule-based systems of larger size are hard to manage and would move the focus away from other important tasks.
Table 3.1:

Problem selection criteria

for applying Rule-Based Expert Systems [Davis, 1986]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Traditional technology doesn't work</td>
</tr>
<tr>
<td>2</td>
<td>Narrow domain of application</td>
</tr>
<tr>
<td>3</td>
<td>There are recognized experts</td>
</tr>
<tr>
<td>4</td>
<td>The experts are provably better than the amateurs</td>
</tr>
<tr>
<td>5</td>
<td>The task is primarily cognitive</td>
</tr>
<tr>
<td>6</td>
<td>Task takes a few minutes to a few days</td>
</tr>
<tr>
<td>7</td>
<td>Task is combinatorial</td>
</tr>
<tr>
<td>8</td>
<td>Solution involves chains of reasoning</td>
</tr>
<tr>
<td>9</td>
<td>The skill is routinely taught to neophytes</td>
</tr>
<tr>
<td>10</td>
<td>Specialists agree on the knowledge</td>
</tr>
<tr>
<td>11</td>
<td>Right stage of knowledge formalization</td>
</tr>
<tr>
<td>12</td>
<td>Data and case studies available</td>
</tr>
<tr>
<td>13</td>
<td>Incremental progress is possible</td>
</tr>
<tr>
<td>14</td>
<td>Not time critical</td>
</tr>
<tr>
<td>15</td>
<td>Freedom to fail</td>
</tr>
<tr>
<td>16</td>
<td>Resulting system would have high payoff</td>
</tr>
</tbody>
</table>
3.3.1.2. Qualitative Reasoning

A substantial number of physical systems consist of a physical network which can be described by a set of interacting state variables. Examples of such physical systems include electrical circuits, mechanical flow systems, and regulated biochemical systems. A stream of AI research has recently yielded methods that perform qualitative reasoning about physical systems [Bobrow, 1985].

The flaws of Qualitative Reasoning methods (see also Section 5.1.2) stem from their preoccupation with electrical circuits. They use -, 0, and + as qualitative values that can propagate through well-defined models of device components. But, for biochemical pathways, the models are never well specified, the parameters of the models are often missing, and the orders of magnitude of quantities play a far more important role than there signs. These flaws led us to the development of a formalism for Order-of-Magnitude reasoning, which we will discuss extensively in Chapter 5. Its applications in biochemical systems will be discussed in Chapter 6.

3.3.1.3. Assumption-Based Truth-Maintenance

In Artificial Intelligence systems, we often deal with monotonic reasoning, in which once an assertion is inferred and believed true the assertion will remain true throughout the reasoning. Thus the list of valid results increases monotonically during the reasoning effort.

In other cases, however, we are faced with non-monotonic reasoning: A particular inference result which is believed valid at some point, may be considered invalid at a later point. In effect, results may need to be in some way retracted, and perhaps repeatedly retracted and reasserted. In non-monotonic reasoning a
truth-maintenance mechanism must exist to structure and update the dynamic set of current beliefs of the system.

Of particular significance in this work is one specific truth-maintenance approach, called Assumption-based Truth Maintenance System or ATMS [de Kleer, 1986, Reiter and de Kleer, 1987]. With this mechanism, assertions can be stated as assumptions rather than known facts. They can also be stated as dependent on the validity of some other assumptions.

At each inference step, along with the justification of a new result, ATMS forms and retains the set of assumptions under which the result is true. The assumption set of each inferred assertion is equal to the union of the assumption sets of all the antecedents used in the derivation. Conflicting (e.g., mutually exclusive) assertions are allowed to coexist (and can even be used in parallel) provided that they carry different assumption sets.

With ATMS, whenever a contradiction occurs the assumption set that supported the contradiction is marked as nogood. From that point on, no new assertion is made under that assumption set or any of its supersets. Reasoning with subsets of the nogood set is not affected.

ATMS avoids some serious problems of other truth-maintenance systems that use dependency-directed backtracking. To resolve contradictions, these systems backtrack, tracing the tree that describes how each result is derived from previous results. In ATMS there is no backtracking involved, and important assumption sets that will determine the final assumptions and beliefs of the system can be constructed after the main problem-solving effort.
ATMS will be the truth-maintenance and hypothesis-testing mechanism that we will use in the O[M] system for Order-of-Magnitude reasoning, discussed in Chapters 5 and 6.

3.3.1.4. Search

A very common subproblem within Artificial Intelligence methods is that of search [Winston, 1984, Charniak and McDermott, 1985]. In search, we have:

- A potentially infinite set of states, which are described by any number and kind of attributes
- An initial state and a target state from the set of states
- A potentially infinite set of operators, each of which conditionally leads from one state to another.

With states corresponding to nodes and operators corresponding to arcs in a directed graph, the search problem entails finding a path from the initial state to the target state.

We normally maintain a whole set of active states (graph nodes) to which we have found paths from the initial state. This set of active states of the problem is originally the singleton of the initial state and we seek to apply operators to incorporate the target state into the set of active states. At each step, then, one state is transformed through the application of operators to yield a set of alternative states, some of which are retained for further exploration.

The search can have many different forms, depending on how the next state is chosen for expansion, and how its children states are incorporated in the set of active
states. Search is a generic problem in Artificial Intelligence; it occurs whenever there is a set of alternative actions or paths but insufficient knowledge to pick the "correct" path without some exploration.

We will end the general discussion of search here, but we will outline elsewhere (in Section 5.8.2 for Order-of-Magnitude reasoning, and Section 9.4.2 for the synthesis of biochemical pathways) how search is carried out within particular procedural methodologies.

3.3.2. COMPUTING ENVIRONMENT AND PROGRAMMING STYLE

The nature of the problem, just as it requires non-conventional techniques, also introduces needs on the computing environment. Conventional environments and programming styles are inefficient for prototyping and execution in the context of pathway design. In our problem, we want the development of the methodological flowchart to proceed dynamically, in small chunks, and in parallel to implementation and experimentation with partial methods.

The point we are emphasizing here is that we must start solving parts of the problem, perhaps crudely, before we even finalize the overall problem formulation and solution strategy.

In pathway synthesis, for instance, it is not a priori clear what exact formulation of the synthesis problem is useful. The kind of specifications one ought to impose on the synthesized pathways can only be decided upon by trying alternative specifications and inspecting the results. Thus, we need a rough working implementation before we are sure about the exact formulation of the problem.
The initial implementations must be especially flexible. In the case of the synthesis problem, the implementation must represent biochemical reaction networks and pathways in a way that corresponds closely to our perception of these concepts — something clearly not possible with traditional programming paradigms. For a particular initial formulation it might be sufficient to represent a whole bioreaction network as an array and a pathway as a vector. However, upon implementing and testing the formulation, it might be decided that rules looking at the internal structure of the pathway must be used. The array-vector implementation would be inadequate and a whole new implementation would be needed.

What is necessary here is an environment and paradigm that allows us to implement pathways so flexibly that any operation we can conceptualize on pathways can be transparently implemented without necessitating a major overhaul of the implementation.

We use the LISP programming language, which offers amazing flexibility [Charniak and McDermott, 1985]. In LISP, users have total control over what goes on (they could change, for example, the syntax of the language if they did not like it!). LISP is oriented towards the manipulation of symbols, rather than numbers, and allows very easy construction of complicated and flexible data structures. These attributes of LISP make it the appropriate vehicle through which we can meet the demands that we placed in the previous paragraphs. In addition, storage allocation in LISP is done dynamically and garbage collection (reclaiming allocated storage) is automatic and transparent to the programmer.

Another important part of our paradigm is Object Oriented Programming (OOP). In OOP chunks of information distinguishable as conceptual entities are represented (implemented) as objects, which can be constructed, removed, and mutated in ways that
may differ from one class of objects to another. Thus, we can implement biochemical pathways, metabolites, and reactions as objects. The combination of LISP and OOP yields a nearly ideal programming paradigm that satisfies all the needs of our problem.

We use Symbolics 3640 and 3650 LISP computers, supporting the Symbolics Common LISP language [Symbolics, 1986], which is a dialect of LISP [Steele, 1984], the Genera software environment, and other software tools. A bottom-up style (i.e., starting from more primitive operations and advancing to higher-level operations), which fits the way we think about the exploration of the problem, and the way we program partial solutions, is pragmatically feasible in this environment. Some of the important differences between conventional environments and the environment that we are using are the following:

- The usual editing-compiling-linking-loading-running cycle is not needed, because all the activities can be carried out incrementally.

- The environment is uniform, with all the software written in the same language (LISP).

- A number of other tools and capabilities are available, such as means for inspection and modification of data structures, maintenance of software systems, special editing capabilities for LISP, etc.

Symbolics Common LISP provides Flavors [Symbolics, 1986], a powerful Object-Oriented Programming facility.

The characteristics of a series of programs that were developed within this work are described in Appendix A.
3.3.3. PREDICTIVE METHODS

While Artificial Intelligence provides the correct philosophy for handling logical or qualitative knowledge, there is still need for methods that estimate approximate values or bounds for important parameters. We will discuss, in Chapters 7 and 8, two such methods we developed based on physicochemical principles:

- A group-contribution method (Chapter 7) for the estimation of equilibrium constants of biochemical reactions

- A methodology for the estimation of the maximum rate of any enzymatic reaction based on the rate of encounter of species in aqueous solution (Chapter 8).

These methods are based on rather routine mathematical manipulation of physicochemical principles and assumptions, with an engineering philosophy guiding the whole approach. This philosophy entails the derivation of bounds and rough estimates of parameters when exact values cannot be found. We believe that such an approach is very valuable and has not been sufficiently exploited in biotechnology.
3.3. THE GOAL: STUDY OF PATHWAY DESIGN

It should be clear that no easy or general solutions can emerge here. The current computing technology and (more importantly) the state of the knowledge on biological systems do not allow that. Our realistic overall goals are thus:

- A framework within which analysis and synthesis of biochemical pathways can be attacked in some cases and studied in other cases

- Methods that can use more of the available qualitative knowledge, or perform educated guesses, and help point the direction in which to seek more knowledge

- Methods that allow systematization of information and shaping of crude pieces of information into principled knowledge.

We believe that the methods we have developed (discussed in Chapters 5 through 10) meet these goals.
CHAPTER 4

RESEARCH FRAMEWORK
4.1. INTRODUCTION

4.1.1. CHARACTER OF THE BIG PICTURE

The research conducted in this project addresses a number of important issues which are critical for the successful design of biochemical pathways. However, the research is largely exploratory, since it is the first concerted effort aiming for systematic and general methodologies for pathway design. Only a fraction of the important issues encountered in systematic pathway design have been concretely addressed and resolved through this research.

It is, thus, important to give the overall picture, i.e., the author's vision of a set of computer-based methods and tools that would allow the biochemist and the biochemical engineer to quickly and effectively synthesize and evaluate biochemical pathways. The purpose of this chapter is to give a glimpse of the bigger picture that motivated this research.

We will describe a research framework, in which a variety of interrelated conceptual methodologies tackle all the needs of a package for computer-aided design of biochemical pathways. We will discuss ideas for further theoretical breakthroughs along with issues regarding the graphic interface of a computer program and the tools that should be accessible through the interface.

It would be presumptuous to claim that the breakthroughs required are feasible in the near future, or that there already exist insightful ideas about how the breakthroughs will be achieved. Thus, the big picture that will be described in this chapter is not a detailed and feasible itinerary, but merely a vague map on which future research efforts in the field can be designed and coordinated.
4.1.2. CONTEXTS

The envisioned system can be organized around some basic tasks which we will call Contexts. They correspond to distinct but interrelated conceptual subgoals, which, in a system implemented in a computer, will also comprise distinct software modules. The proposed Contexts are:

(a) Biochemical Database Context
(b) Pathway Synthesis Context
(c) Evaluation and Prediction Context
(d) Pathway Recommendation Context
(e) Learning Context

In a computer-based package, these Contexts should be tightly connected through common schemes for Knowledge Representation or easily interconvertible representation structures, so that:

- Each Context can use conceptual methods and display features that reside in other Contexts
- The user can freely leap from Context to Context without interrupting or transforming previous work or having to carefully contemplate and plan the transition.

For example, constructs created or examined in the Biochemical Database Context, should be easily transferable and usable in the Evaluation Context for refinement.
In a computerized system encompassing these contexts, it is also useful to have a graphic interface for all the contexts, in order to make the manipulation of knowledge and methods easier, and present results in a graphical, transparent fashion to the user of the system. We will discuss first some features we would like such a graphic interface to possess. Each of the contexts will then be examined in subsequent sections. In Appendix A, we describe computer programs that meet some of the desiderata raised in this chapter.
4.2. GRAPHIC INTERFACE

The graphic interface for the analysis and synthesis of biochemical pathways is needed to:

- Facilitate storage and retrieval of information in a Biochemical Database
- Allow transparent construction and modification of biochemical pathways through graphical operations
- Automatically formulate problems, and invoke methods requiring different knowledge representation schemes

The last point cannot be overemphasized. The flexible AI methodologies, and programming styles like LISP procedures and Object-Oriented Programming (Section 3.3.2), are required exactly to allow radically different actions, methods, and strategies to be applied to the same original representation of a biochemical pathway (with additional representations automatically created and updated as dictated by efficiency considerations). These programming styles excel at the implementation of very flexible representation schemes and the automatic transformation of one representation scheme to another.

Aiming at providing maximum flexibility in the use of the graphical and conceptual tools, several options should be provided at the interface. These options should be accessible through a hierarchy of menus and should include a variety of operations:

- Display operations, such as:
  - Fonts and sizes of text displayed
◊ Colors or reverse video display
◊ Window creation and manipulation
◊ Manipulation of entities displayed graphically (e.g., zooming in or out, moving a node of a graph, etc.)

• Invocation of operating system actions
• Saving the current state of a problem to a file
• Retrieving a problem from a file
• Invocation of tools that are available in all Contexts, such as:
  ◊ Display of information on an entity from the database (e.g., a metabolite, a reaction, or a microorganism)
  ◊ Display of relations and classification of entities
  ◊ Application of available simple estimation methods, such as a group-contribution estimation of an equilibrium constant

• Selection among the available Contexts
• Context-specific operations, detailing current task in progress, available tasks with options for each task, etc.

The last set of menus, describing modes of operation, are, of course, different for each Context but also vary according to the particular operation within the Context. For Chaining in the Biochemical Database Context (which, as described in Section 4.3, allows
the creation and expansion of a graph of concepts based on some relation), the interface should offer displays and options about:

- The particular form of chaining
- Pruning criteria
- Information to be shown for each node.

The utility of the Graphic Interface will become clearer in the remaining sections of this chapter, in the discussion of various Contexts. For example, in the Intelligent Biochemical Database (Section 4.3.2) information can be stored into and retrieved from the database through graphical operations. In the Interactive Pathway Synthesis activities (Section 4.4.1), the synthesis and modification of pathways is similarly to be carried out through creation and graphical manipulation of icons. This last capability is probably the most important feature of the graphic interface, and at the same time the most difficult, as will be discussed in Section 4.4.1. In Appendix A, we describe an interface for editing a database and applying other analytical techniques.
4.3. BIOCHEMICAL DATABASE CONTEXT

In this Context, creation, modification, storage, and retrieval of objects is carried out in a menu-driven fashion. The database and interface developed in this work is described in Appendix A. Here we will describe a much broader system.

4.3.1. OBJECT TYPES

For each type of object, a different set of data must be available. The basic object types are:

- **Metabolites** of small molecular weight, which serve as substrates, products, intermediates, or cofactors in enzymatic reactions
- Stolchiometrically defined biochemical reactions
- **Enzymes** from defined sources
- **Biochemical pathways** which consist of bioreactions
- **Microorganisms**.

In this document we will often use the term reaction to refer to an enzymatic reaction (also termed bioreaction or biochemical reaction), unless we explicitly state that the reaction is not enzymatic.
4.3.1.1. Distinction between Bioreactions and Enzymes

The terms *biochemical reaction* and *enzyme* are often used interchangeably in the literature. We should thus clarify the difference between the two terms, because it is important in this document, particularly in the Database Context.

4.3.1.1.1. Bioreactions

A biochemical reaction is a *transformation* uniquely and fully defined by:

- Its reactants and products, along with their stoichiometric coefficients.
- Any cofactors which are *modified* by the reaction, since such cofactors ought to be considered reactants or products of the reaction.
- Optionally, any other metabolites which participate in the reaction but are recovered, such as 2,3-diphosphoglycerate in the reaction catalyzed by *phosphoglycerate phosphomutase*:

\[
3\text{-phosphoglycerate} + 2,3\text{-diphosphoglycerate} \rightarrow 2\text{-phosphoglycerate} + 2,3\text{-diphosphoglycerate} \tag{4.1}
\]

4.3.1.1.2. Enzymes

An enzyme is a particular macromolecule (a protein) with specified primary structure, which implies specified molecular weight, and in practice specified secondary and tertiary structure as well.
Although it is in principle possible for a given enzyme to be present in many independent sources, in practice one expects enzymes to display at least small differences (if not major ones) among enzymes from different sources.

An enzyme from a given source may catalyze many biochemical reactions, but there is normally one particular reaction thought of as the main or nominal reaction catalyzed by the enzyme. Thus, for our purposes, we can consider an enzyme uniquely defined by:

- A particular source, such as a bacterium
- A nominal biochemical reaction

4.3.1.1.3. Distinctions and Usage

The definitions provided above should make it clear that a term like phosphoglycerate phosphomutase, which is normally thought of as the name of an enzyme, is in reality the name of a bioreaction (Reaction 4.1).

There are many different enzymes, with different molecular weights, all catalyzing that same reaction. Each of the enzymes can have different kinetics, although the equilibrium constant is the same for all enzymes as it is uniquely specified by the net reaction stoichiometry. Each enzyme will in general be inhibited by different inhibitors.

Each of the enzymes can also catalyze any number of other reactions. In fact the nominal reactions of the enzymes may differ, and not all of them will have Reaction (4.1) as their nominal reaction.
In the literature the terms *enzyme* and *bioreaction* (or *biochemical reaction*) are often used interchangeably. Having made the distinction clear, we may occasionally use the term enzyme when we really refer to a bioreaction, because it is convenient to use the names of enzymes as the names of bioreactions.

The important point one should keep in mind throughout this document is that, as long as we do not specify the source, we are only referring to a bioreaction and not to a specific enzyme.

### 4.3.2. DESIRED DATA

We will, in this section, outline the kind of data we would like to have for each type of entity. Because in some cases multiple conflicting values may be available for an attribute of an object, we must provide the ability to store more than one values in each slot.

For each value it is imperative to maintain appropriate bookkeeping and reference information, such as:

- Bibliographical reference for a datum
- Reported accuracy
- Temperature, pH, and other conditions
- User-supplied tags that may indicate:
  - The user’s confidence (or lack thereof) in the particular datum
  - Possible preferences of the user among alternative values
Although sources of collected data exist in the literature [Altman and Dittmer, 1972, Thauer et al., 1977, Byng et al., 1982, Mendelson, 1982, Neidhart et al., 1983, Sletay and Robinson, 1984, Thauer and Morris, 1984], it is expected that much of the data will not be available a priori. A lot of the knowledge needed for a particular task will be specifically acquired and added to the database to carry out the task at hand. The user-added data must be distinguished from the system's own data, since they are more likely to be saved in a different file, discarded, or modified.

The most useful data are the ones that can be used as building blocks in combinatorial constructs. Thus, data on the simplest building blocks, such as data on groups within a group-contribution estimation technique, are the most valuable. We have developed, for example, a group-contribution method for Gibbs Free Energies of Formation of compounds in aqueous solution (Chapter 7); we will see that from a few group parameters one can estimate the corresponding parameters for any molecule or bioreaction.

With respect to data that cannot be combined in multiple ways, it is not wise to spend effort in collecting them, but it is important that means of expressing such data exist, so that they can be acquired and used in the context of a specific problem that requires them.

We outline below the kinds of data that are desired for the classes of objects listed in Section 4.3.1.
4.3.2.1. Small Molecules

4.3.2.1.1. Structural Data

There are three alternative levels of detail in representing the structure of a molecule:

◊ Some representation of the (three-dimensional, in general) structure of the molecule in terms of elements and bonds. This would be ideal, especially if there are methods for converting such representation to the next level of detail.

◊ The set of groups that comprise the molecule, possibly with their interconnections. Knowledge of the set of groups allows the application of group-contribution methods.

◊ The molecular formula. This is the least informative description, but there is some use for it. For example, it allows element balances to be performed in the development of fermentation equations (Section 3.2.5.1), or in checking the internal consistency of biochemical reactions.

4.3.2.1.2. Physicochemical Data

For example:

◊ Molecular Weight

◊ Gibbs Energy and Enthalpy of formation

◊ Gibbs Energy and Enthalpy of solution
Solubility in water

Solubility in other solvents. Such data would be important in the evaluation of separation schemes (when the by-products or yield of different pathways is not the same)

4.3.2.1.3. Physiological Data

For example:

Role in the metabolism

Physiological effect on cells

Regulatory effects on reactions and pathways

Prevalent pathways for the production and consumption of the molecule in different cells

4.3.2.1.4. Other Data

- Technical and empirical data, such as:
  - Typical concentrations in different cells or bioprocesses
  - Pathways used industrially for the production of the compound

- Economic data, such as the cost of those molecules which might be raw materials or products (or by-products) of bioprocesses.
4.3.2.2. Bioreactions

- Stolchiometry of the reaction.

- Physicochemical data, such as:
  - Expressions for maximum rate estimation,
    with the methodology discussed in Chapter 8
  - Gibbs Free Energy of Reaction
  - Equilibrium constant
  - Enthalpy of Reaction

- Empirical data, such as:
  - Regulation of the bioreaction (if it does not vary among different enzyme sources).
  - Reversibility of the reaction
  - Physiological role

4.3.2.3. Enzymes

- Bioreactions catalyzed

- Physicochemical data, such as:
  - Molecular Weight
  - Number and nature of subunits, reactive sites, and allosteric sites

- Activity data, such as:
  - Michaelis kinetic parameters
  - Data on inhibition by substrate, by analogue, or by product, including both the kind of inhibition
(competitive, uncompetitive, or non-competitive)
and the inhibition constants

- Post-transcriptional regulation, such as:
  - Allosteric inhibition agents and their effect
  - Activation
  - Covalent modification

- Genetic information, such as:
  - Correlations between the level of the regulator and the concentration or production rate of the enzyme
  - Effects of other growth variables on enzyme activity
  - Rigorous models of enzyme synthesis and activity

### 4.3.2.4. Pathways

- Physicochemical data, such as:
  - Overall equilibrium constant
  - Maximum rate

- Critical mechanisms, such as:
  - Rate-limiting steps
  - Regulatory structures

- Information characterizing other aspects of the pathway as an integrally-functioning process. Such information must be
useful in pathway synthesis and the evaluation of competing pathways within the metabolism.

4.3.2.5. Microorganisms

- Composition in terms of:
  - Elements
  - Classes of compounds (such as amino acids, carbohydrates, lipids, etc.)
  - Concentrations of key metabolites.
- Pathways that are present or absent from the metabolism.
- Enzymes that have been studied from the microorganism.
- Macroscopic experimental data, such as:
  - Growth requirements.
  - Growth rates
  - Product yields

4.3.3. OBJECT CONNECTIONS AND CLASSIFICATION

4.3.3.1. Relations among Objects

Wherever in our description of the desired data we described attributes whose values might be other objects from the database (implicitly defining in this manner relations among objects), we require that the reverse relations, and thus the reverse pointers, also be present. The fact that the objects are connected in all possible ways and in both directions means that intelligent features of the Database, which will be discussed in Section 4.3.4, can avoid unnecessary search.
We give just a few examples of relations among objects, in the in-text tables of the following paragraphs.

4.3.3.1.1. Relations for Molecules

A metabolite (small molecule) ought to be connected to:

◊ Reactions that consume or produce it

◊ Pathways that consume or produce it

◊ Pathways in which it is involved as an intermediate (although this particular connection can be easily derived from the reactions that consume or produce the metabolite and the pathways those reactions participate in)

◊ Enzymes which it inhibits

◊ Enzymes whose synthesis it regulates

4.3.3.1.2. Relations for Bioreactions

A bioreaction ought to be connected to:

◊ Metabolites it consumes

◊ Metabolites it produces

◊ Enzymes that catalyze it

◊ Pathways that use it
4.3.3.1.3. Relations for Enzymes

An enzyme ought to be connected to:

◊ Bioreactions it catalyzes
◊ Compounds that inhibit it
◊ Compounds that regulate its synthesis
◊ Microorganisms that contain it

4.3.3.1.4. Relations for Pathways

A pathway ought to be connected to:

◊ Its bioreactions
◊ Metabolites it consumes or produces
◊ Its intermediate metabolites
◊ Microorganisms in which it is active
◊ Microorganisms in which it is inactive
◊ Potential rate-limiting steps
◊ Steps that may be regulated, and the metabolites responsible for the regulation
◊ Intermediates that are important points of branching or competition with other pathways
◊ Related pathways that may substitute it
4.3.3.2. Classification of Objects

In addition to connections among simple objects, there are additional relations that we would like the database to contain. We want objects in each type-class to be further classified in a way that facilitates search and abstraction. This implies the existence of additional classes (which are also self-standing objects) related to the simple objects through the classification scheme.

We discuss possible classifications of the basic types of objects below.

4.3.3.2.1. Metabolites

Small molecules can be classified either chemically (i.e., based on their chemical structure) or physiologically (i.e., based on their common role in the metabolism of organisms).

- Chemically, we can have classes such as hydrocarbons, carbohydrates, carboxylic acids, lipids, alcohols, α-keto-acids, amino acids, etc.

- Physiologically, we can divide metabolites into classes like building blocks (subdivided into more specific classes), currency metabolites (such as NAD, ATP, etc.), cofactors, intermediates of special pathways, etc.

4.3.3.2.2. Bioreactions

The classification of bioreactions is similar to that of metabolites, in that it can be distinguished into chemical classification (based on the character of the chemical transformation involved) and physiological classification (based on the section of the
metabolism in which a reaction belongs and any special physiological role of the reaction).

- Chemically, we can have classes such as decarboxylations, dehydrogenations, hydrations, etc. Clearly this classification corresponds to the one used by the Enzyme Commission in its Enzyme Nomenclature [Webb, 1984]. Thus, the Enzyme Commission (EC) number of a biochemical reaction already provides a good classification.

- Physiologically, we can distinguish between fueling, biosynthetic, and intermediary reactions. The individual pathways (as well as classes of pathways) to which reactions belong are also an implicit classification.

4.3.3.2.3. Enzymes

Enzymes can be classified based on properties like:

- Number of subunits

- Functionality; multi-functional enzymes catalyze several reactions

- Regulation; allosteric enzymes possess sites that bind inhibitors

- Presence in particular classes of microorganisms, when the same enzyme (or very similar enzymes) are present in a whole class of sources.
- Attributes of their reaction mechanism; ordered and random-order mechanisms are the simplest kinds.

4.3.3.2.4. Pathways

Pathways can be classified in a manner similar to the classification of metabolites and bioreactions. Thus, we have chemical and physiological modes of classification:

- Chemically, the classification is based on the transformation achieved, and the classes would be similar to those for reactions. They could even be exactly the same classes as for reactions; after all, pathways and bioreactions both describe transformation and they only differ in the number of enzymes they require.

- Physiologically, we can distinguish between pathways from fueling, biosynthetic, and intermediary metabolism. The microorganism classes in which a pathway functions are also an implicit classification (It is common, for example, to talk about yeast pathways vs. bacterial pathways).

4.3.3.2.5. Microorganisms

A good classification of microorganisms can be obtained based solely on their phylogeny. Additional basis of classification is provided by:

- The activity and regulation of key pathways

- Ability to grow on certain substrates
• Ability to survive under different conditions. For example, thermophile microorganisms grow at temperatures above 50°C, mesophile microorganisms grow optimally around 37°C, and psychrophile microorganisms can grow under 5°C [Ingraham et al., 1983].

4.3.4. INTELLIGENT FEATURES OF THE DATABASE

The intelligence of the Database consists mainly of ways to cooperate with the users more tightly, and allow the users to use their intelligence unimpeded by bookkeeping or other burdensome but unimaginative tasks. There are also areas in which it is possible to make the database behave intelligently by performing certain open-ended and ill-defined tasks.

Many of the desired features of the database are discussed, implicitly or explicitly, in the discussion of the other Contexts (in other sections of this chapter). We will discuss here only some of the features, involving the system's ability to perform:

• Consistency checks to cross-verify data

• Construction of graphs of relations by chaining (i.e., recursively chasing pointers)

• Intelligent guesses about what information the user might like to have, but has not explicitly asked for, in answering a particular query. Let us not forget that computer users, and especially computer-illiterate users, always expect the machine to give them what they want, and not what they asked for.
- Abstraction of complicated structures, such as transformation a large detailed metabolic network into a simpler network which is easier to understand but retains the important features of the initial network.

4.3.4.1. Consistency Checks

As we discussed already, the database must be extendable, and the interface for acquiring and storing data must be an integral part of the database. One expects that information on biochemical systems, because it is often qualitative, informal, and of limited accuracy, will have many more inconsistencies than knowledge on other engineering systems. Thus, it is useful to have, within the Biochemical Database Context, extensive built-in abilities to check acquired data for consistency (self-consistency, consistency against pre-existing data, and consistency against estimation or prediction methods).

Implementational consistencies, such as the correct maintenance of reverse relations as the database is modified, should be taken care of at the level of the implementation and will not be discussed here.

We will provide in this section some examples of the kind of consistency checks that can be performed, counterexamples involving apparent inconsistencies that should not be resolved, and a discussion on how inconsistencies ought to be handled by the database and the system's interface.
4.3.4.1.1. Examples of Inconsistencies that can be Checked for

- If there are different representations of the metabolite's structure, then they must be consistent. For example, the molecular formula should agree with the decomposition of the molecule into groups.

- Reactions must be thermodynamically feasible at least in their forward direction. Thus, exorbitantly low equilibrium constants (which are frequently found in the literature) are incorrect.

- Numerical parameters must lie in predetermined ranges.

For example:

◊ Molecula: weights of metabolites must be in the range 1 to 2000

◊ Molecular weights of enzymes between must be higher than 5000

4.3.4.1.2. Apparent Inconsistencies that cannot be Resolved

- A characterization of a bioreaction as irreversible by biologists may be in disagreement with a measured or estimated equilibrium constant. However, the characterization of a reaction as irreversible often means that a specific enzyme could not catalyze the reverse reaction due to inhibition effects. Thus, there is no true conflict:

◊ The reverse reaction is thermodynamically feasible
The reverse reaction has been observed to be kinetically infeasible because of inhibition.

- Kinetics of enzymes can differ drastically (for a given enzyme source), depending on the method of purification, because the final purity and the extent to which the enzyme has been deactivated depend on the purification method.

4.3.4.1.3. How to handle Inconsistencies

Data on biochemical systems are much more unreliable than data for most other engineering systems, but, as we saw in Section 4.3.4.1.2 above, an apparent inconsistency could in reality be a source of additional refinement of the knowledge, rather than a true conflict.

Thus, a computerized database should make no assumptions about the possible source of inconsistency and should never attempt to remedy an inconsistency automatically. The multitude of special factors that must be considered in each case necessitate resorting to a human expert for the resolution of any contradiction in the database.

4.3.4.2. Chaining

As chaining we term here the creation of internal digraphs (or graphs) whose arcs and nodes can be Database objects. The digraph can have the more specific form of a tree or a loop. It must have a graphic representation that the user can understand at a glance. Information can be displayed for every node (database object) and arc (relation between objects).
4.3.4.2.1. Search Modes

The search in chaining can be:

- Complete (all possible expansions)
- User-directed at every step
- Best-first, beam search, or any other search type [Winston, 1984] with an objective function provided by the user
- Best-first, beam, etc. but based on the system's objective function, provided that the user's general motive is known
- Partial, with pruning criteria

4.3.4.2.2. Applications

A generic chaining framework can be applicable in many different cases:

- Interactive search for a good route for the production of a bioproduct. Examples of pruning criteria can be the requirements for oxidation-reduction or ATP balances.
- Full expansion of the routes for consumption of substrate, similar to above.
- Causality or dependency networks. For example, the rate of a reaction depends on the concentrations of its reactants, products, enzyme, and allosteric regulators. In turn, these concentrations depend on yet other parameters to which the system can chain backwards, and so on.
4.3.4.3. Informative Answers to Queries

In the effort to make the Database more intelligent, it is desirable to provide the user with additional information, not explicitly requested by the user's query, and foresee mistakes that can result from possible exceptions and special cases. This idea can be best demonstrated through hypothetical "dialogues":

Q: What is the equilibrium constant of this reaction?

A: 200, but the reference dates back to 1930.

Q: What is the average of the reported growth rates for this microorganism?

A: 0.5 h⁻¹, but there is only one measurement.

Q: How many reactions produce S?

A: Four, but they only differ in the cofactors they use.

Q: What is the overall equilibrium constant for this pathway?

A: 1, because the reactants and the products are the same.

Q: Does microorganism Z consume X?

A: Yes, but it converts X extracellularly to Y first (This is important when many microorganisms are to compete for X and Y).

There is no predefined scheme for these interventions. They can be expanded as the system is put in practical use and related experience is acquired. This feature,
which is virtually open-ended, can be best implemented incrementally in a rule-based fashion.

4.3.4.4. ABSTRACTION

Abstraction must provide the ability to view a network of biochemical reactions as consisting of smaller ones, which in turn are subdivided (recursively) all the way to simple reactions or even steps in an enzymatic reaction mechanism. The standard approach of viewing whole-cell metabolic networks as consisting of biochemical pathways (rather than individual biochemical reactions) involves this kind of abstraction.

A desirable interface feature is the use of similar levels of abstraction in the display of metabolic networks, to encourage the user to examine the problem one level at a time and neglect lower level details.

The issue is also a pragmatic one, because the graphic display would be rendered unreadable if too much information were displayed. The abstractions in the graphic display, however, should not be simple-minded schemes, such as substituting black boxes for arbitrary portions of the display; instead, they should correspond closely to clean conceptual abstractions.
4.4. PATHWAY SYNTHESIS CONTEXT

Synthesis of biochemical pathways entails construction of pathways from available or conjectured biochemical reactions. The selection of the bioreactions that are used in the construction of a pathway can be done manually (Interactive synthesis) or automatically (automatic synthesis) based on some specifications placed on characteristics of the pathway being synthesized. We will discuss each kind of synthesis below.

4.4.1. INTERACTIVE SYNTHESIS

The construction of complex metabolic structures can be achieved by retrieving, chopping, and connecting pathways (or individual bioreactions), through easy graphic operations. Use can be made of the chaining feature described earlier (Section 4.3.4.1), to display alternatives at each step. Although the synthesis is in nature interactive, there may be specifications placed that allow the interface to guide the user through the construction step by step and prevent errors.

There are some semantic issues arising at the interface level during interactive synthesis. We discuss these issues in the following subsections.

4.4.1.1. Disconnection and Deletion of Objects

When someone disconnects or kills objects at the graphic interface, it is not clear whether the "real" dynamic objects underlying the icons should suffer the same treatment. Note that when an object is killed, its connections must first be destroyed. The user may or may not mean to destroy the connections permanently, but if a killed
object is created again, the connections are simply not there, and they must be explicitly established, one by one. There are two solutions to this issue:

- Limit the operations the user can perform, so that it is always clear what a graphic operation is intended to accomplish. This would severely limit the flexibility of the interface.

- Inhibit creation and updating of the real objects (i.e., objects other than purely graphical icons), until a particular tool or Context requiring them is invoked. As the data-structures of a partially completed pathway are not available, we must either prevent the user from seeking information on partial pathways, or prepare and display information equivalent to the missing data-structures at the user's request (in the latter case, the whole problem is simply carried over to the construction of a consistent and meaningful such display).

Each approach limits the flexibility of the system in some way. This is an issue that can be resolved only when a very powerful system is built and used extensively for practical applications.

4.4.1.2. Side Metabolites in Bioreactions

It is common for reactions to be drawn with some of their reactants designated as side-reactants and some products designated as side-products, as shown for three reactions in Figure 4.1.

Note in particular that glutamate, which is a side-reactant for Phosphoserine transaminase (Figure 4.1 b), and α-ketoglutarate, which is a side-product, become
main reactant and product in *Glutamate dehydrogenase* (Figure 4.1 c). In effect, a metabolite is not restricted to play the same role in all reactions in which it participates.

Reactions are very often (but not always) drawn with only one main reactant and one main product. This is the case with the reactions (a) and (b) of Figure 4.1, but not with reaction (c).

This representation is a manifestation of the perception that the side-metabolites are, at least for that particular reaction, *currency metabolites*, and one does not worry as much about their consumption and production, since they are produced and consumed by many reactions. For example, NAD and NADH (Figure 4.1 a, c) are invariably considered currency metabolites for oxidation and reduction. Thus, in synthesizing pathways, one might decide to deal with side-metabolites at the very end, or neglect the presence of side-metabolites altogether.

However, the distinction of main and side metabolites is not even fixed for a particular reaction. It depends on the particular use of the reaction in a pathway, as well as subjective preferences (dependent both on the strategy of construction of the pathway and the perception of the pathway’s functionality). For example, for *Glutamate dehydrogenase* in Figure 4.1 (c), ammonia can be perceived as either a main reactant or side reactant, depending on whether we wish to emphasize the assimilation of nitrogen or just the regeneration of glutamate.

While there are many cases in which the proper characterization of the metabolites in a reaction or pathway is obvious, it is not possible to come up with general procedures for the task. Thus, some guesswork is necessary to arrive at some defaults which the user should have the option to change.
Figure 4.1:

Distinction of metabolites into main-reactants and side-reactants (as well as main-products side-products) for three particular bioreactions.
4.4.2. AUTOMATIC SYNTHESIS

For the automatic synthesis of biochemical pathways, we would like to have methodologies which, given a set of specifications on the pathway (or even specifications on the whole bioprocess), can generate pathways from existing or conjectured reactions satisfying all (or most) specifications.

4.4.2.1. Specifications

The constraints imposed on the synthesized pathways can have many different forms and can refer to different attributes of the pathway or the process. We give below some examples:

- The desired product is produced
- Only substrates from a prespecified set are used
- Only by-products from a prespecified set are produced
- Oxidation-reduction balances hold
- The reactions of the pathway are thermodynamically favorable
- The pathway does not entail extreme enthalpy changes
- The pathway competes with few other pathways
- All enzymes except at most two should already be present in the same microorganism, so that only two genes will need to be cloned
- The pathway should have high yield
- The kinetics of the pathway should have the potential to achieve high production rates
• The regulation of the pathway should be weak, or limited to at most one enzyme

• All undesired pathways competing with the desired one should have regulated steps early after the branching points, so that they can be eliminated by tuning the regulation

4.4.2.2. Methodology Developed

In Chapter 9 of this thesis we present a methodology we developed for the synthesis of biochemical pathways satisfying stoichiometric constraints. The effective incorporation of other kinds of constraints is limited because:

• Constructing an appropriate algorithm is generally difficult. The existing algorithm rests heavily on the linear character of stoichiometric specifications; non-linear specifications cannot be incorporated in it.

• For many specifications, the information (parameter values) that would make them usable is not available in practical applications.

4.4.3. NON-ENZYMATIC REACTIONS

A complication in both interactive and automatic synthesis is that two types of reactions are not taken into account:

• Intracellular spontaneous reactions, taking place under the normal mild conditions of the cell
• Extracellular chemical reactions, which can be used in a bioprocess for preprocessing of raw materials or conversion of the product of a fermentation to the final desired product

4.4.3.1. Spontaneous Reactions

Although most reactions taking place in a cell must be catalyzed by enzymes and cannot occur non-enzymatically, there are some reactions that occur spontaneously. For example, as shown in Figure 4.2 in the pathway for the synthesis of proline, conversion of L-\(\text{Glutamate-}\gamma\text{-semialdehyde}\) to \(\text{L-}\Delta^1\text{-Pyrroline-5-carboxylate}\) occurs non-enzymatically [Raw, 1983].

In interactive synthesis, it is left to the user to decide that a spontaneous reaction will occur and create the reaction.

In automatic synthesis we can have spontaneous reactions included in the reaction Database, but in this fashion we can only cover special reactions of known importance in biochemical pathways. In general, we need a whole system that can predict spontaneous conversions of compounds in an aqueous environment.

In this work, we choose to deal throughout only with known, specific biochemical reactions, because introducing additional details to include the exceptions would be a major undertaking and would clobber our main effort. We accept the limitations outlined, as inherent to our approach.
Figure 4.2:

In the pathway for the synthesis of proline, the conversion of L-Glutamate-γ-semialdehyde to L-$\Delta^1$-Pyrroline-5-carboxylate occurs spontaneously.
4.4.3.2. Conventional Chemical Reactions

When the conversion of a substrate $S$ to a desired product $P$ is not easy through biochemical processes alone, there may be easier processes if we employ some non-biochemical steps:

(1) $S$ can be preprocessed chemically to obtain $A$, which is biochemically convertible to $P$. For example, as shown in Figure 4.3, lysine can be derived enzymatically from DL-$\alpha$-Aminocapro lactam, which is produced chemically (from chemical raw materials).

(2) $S$ can be converted to $B$ biochemically, then $B$ converted to $P$ through a couple of easy chemical steps.

(3) As a combination of the first two cases, $S$ is chemically preprocessed to $A$, $A$ is biochemically converted to $B$, and $B$ is chemically converted to $P$.

The industrial production of Phenylalanine can be carried out either through fermentation, or through a combination process involving both chemical and biochemical steps [Crueger and Crueger, 1984].

In interactive synthesis the burden rests again with the user. The comments from the discussion of spontaneous reactions (Section 4.4.3.1), on the limitations introduced and the reason it is generally hard to tackle this problem in automatic synthesis, apply here as well. Reactions of this type, however, occur much more frequently in industrial processes, so the limitations are more severe. To solve the problem one needs a system like those developed for laboratory organic synthesis [Blower and Whitlock, 1976] adjusted to industrial processes.
Figure 4.3:
A route for the production of lysine
via enzymatic transformation
of a chemically produced intermediate
[Crueger and Crueger, 1984]
4.5. EVALUATION AND PREDICTION CONTEXT

Given a particular biochemical pathway, which might be an existing pathway already used in a bioprocess or a pathway synthesized automatically, we would like to evaluate how good the pathway is and how it can be improved.

The task at hand is not classical simulation. It is a more general task of reasoning about quantities, and it involves handling qualitative and quantitative information. The goal is to examine important quantities of a pathway, like concentrations, fluxes, mass action ratios, equilibrium constants, kinetic constants, yield, and selectivity, and determine approximate relations or value ranges, so that we can conclude what major factors affect the process, and what bottlenecks must be bypassed to achieve improvement of the pathway.

It is also desirable to determine the consequences of alternative hypotheses. For example, one may examine alternative regulation schemes of a pathway using heuristics about concentrations, displacement of reactions from equilibrium, and extent of inhibition.

We will now examine some characteristics of the knowledge that is to be used in the evaluation of pathways, and the kind of methods (some of which we have already developed) that can address the respective difficulties. Then we will discuss in more detail some particular tasks from this Context.

4.5.1. DEALING WITH INCOMPLETE KNOWLEDGE

We have repeatedly mentioned that much of the desired data on a biochemical system are in general not available. A key factor in successful pathway design is an
effective set of methodologies for the estimation of approximate values (or bounds) for missing but necessary parameters.

In this work we have developed two key estimation methods:

- A group-contribution method for the estimation of equilibrium constants of biochemical reactions, discussed in Chapter 7.

- A method for the estimation of the maximum rate of any enzymatic reaction, based on collision limitations, discussed in Chapter 8. This method allows the kinetic evaluation of pathways in the absence of information on actual enzyme kinetics.

In Section 4.5.3 we will discuss the development of other prediction methods that would be useful in pathway evaluation. The engineering philosophy (discussed in Section 3.3.3) underlying the development of such methods should be kept in mind.

4.5.2. DEALING WITH QUALITATIVE KNOWLEDGE

We mentioned in the previous subsection that knowledge on biochemical systems is in practice incomplete. Another important characteristic of the knowledge is that it is often qualitative in nature.

To capture qualitative knowledge, we have developed the O[M] formal system for Order-of-Magnitude reasoning which can represent and use approximate relations among parameters. This formal system will be discussed in Chapters 5 (and its applications in Chapter 6). While the developed formal system can tackle a number of tasks, there is much to be desired (and a lot of Artificial Intelligence research is taking
place) in methods for generalized reasoning with parameters at both the qualitative and quantitative level.

4.5.3. DETERMINATION OF RATE-LIMITING STEPS

The problem of determining potential rate-limiting steps of a pathway inherently involves reasoning about orders of magnitude. Even the definition of rate-limiting steps is that, in some sense, those steps are "much slower" than the other steps of the pathway. An O[M] formulation of this task will be discussed in Section 6.2.2.

Knowledge useful in the determining rate-limiting steps comes from:

- Equilibrium constants
- Kinetics of enzymatic reactions
- Enzyme activities or concentrations
- Overall fluxes through relevant pathways
- Concentrations of intermediates
- Hypotheses, goals, or other clues

The main difficulty in the identification of rate limiting steps is not the lack of general methodologies but the lack of data. Specifically, enzyme concentrations and kinetics play a major role in rate limitations but are generally not known (even for widely used industrial processes).

Any approach that could, at least heuristically, estimate ranges for these important kinetic parameters would substantially ease most subtasks of pathway
evaluation. However, the feasibility of such an approach can be seriously doubted even in principle.

4.5.4. CONJECTURING REGULATORY MECHANISMS

When the steps of a pathway are known, but the regulation of the pathway is not, it is desirable to conjecture a likely regulatory structure.

We confine the discussion to allosteric inhibition. The regulation of enzyme synthesis and the covalent modification of enzymes are mechanisms with more important effects, but it is simply too hard to conjecture them without detailed a priori knowledge (as we will discuss again in Section 4.6.2).

Useful knowledge for this task can come from

- Rate-limiting steps either known in advance, or identified by reasoning

- Generalized heuristics

The generalized heuristics, referring to the existence or location of regulatory structures, depend heavily on whether the microorganism is wild-type, mutant, or recombinant.

The heuristics, serving mostly as candidate generators, can be translated into algorithms (possibly involving Order-of-Magnitude knowledge) or production rules. After candidate mechanisms are generated, they can be tested using the Order-of-Magnitude formalism (Sections 4.5.4.3 and 6.2.4). We will discuss in the following subsections the relevant knowledge for wild-type organisms, for mutants, and for the testing of candidate mechanisms.
4.5.4.1. Heuristics for Wild-Type Regulation

(1) Steps that follow a rate-limiting step cannot exert any regulatory effect on the pathway.

(2) When connected pathways are functionally distinct (e.g., energy-production and biosynthesis), they are not coupled through regulation.

(3) A pathway from a start-metabolite to an end metabolite must have regulatory structure. A start-metabolite is one which is used for several purposes in quantities comparable to or larger than the purpose at hand. An end-metabolite is one which is produced mainly by the pathway at hand and is either:

   (a) Lacking a dominant use, i.e., it is used in many pathways in comparable quantities, or

   (b) Used in a transformation different from normal biochemical pathways, such as construction of macromolecules or cell membrane.

The underlying argument is teleologic. Since there must be some regulation and no significant indirect regulation seems to take place, direct regulation must occur.

(4) The first rate-limiting step of a pathway after any branching is a likely target for regulation. This step is often called the committing step.

(5) In biosynthetic pathways, the metabolite effecting the regulation lies downstream from the reaction that serves as the target of the regulation.
4.5.4.2. From Wild-Type to Mutant Regulation

Important parts of the knowledge used for the wild-type strains rely on biological teleology, and are therefore inapplicable to mutants.

One way to tackle the problem of regulation in mutant strains is to proceed incrementally, the way biologists interpret mutation experiments on pathway regulation. First, reasoning is done on the wild-type. Then, one attempts to interpret the differences between the wild-type and the mutant in terms of a small number of structural changes (mutations). Knowledge is required on the differences between the strains:

- Nutrition requirements of the mutant
- Changes in yields, rates, and other macroscopic parameters
- New by-products
- Procedure employed to select for the mutant

Reasoning involves Order-of-Magnitude analysis but also reasoning reminiscent of the diagnosis of faults (such as faults in chemical processes or electrical circuits). The task is extremely difficult to automate, because it is further complicated by other phenomena:

- Multiple mutations
- Mutations with multiple effects
- Leaky mutations
The user's expertise may be the main drive, with only bookkeeping and hypothesis-testing support provided automatically. Two approaches can be followed for the automation of this task:

- One can proceed backwards from the evidence (symptoms) to the mutations (cause of the symptoms), using empirical biological rules. The resulting hypotheses can subsequently be tested forward from the mutation to its effects.

- One can generate all possible hypotheses through some algorithm, and then test them. To reduce the search space, assumptions must be made about the nature and the number of mutations.

4.5.4.3. Testing of Regulatory Mechanisms

A proposed regulatory mechanism must be checked for consistency against other available knowledge. A useful reasoning formalism is Order-of-Magnitude reasoning, which can be used in a manner similar to the identification of rate-limiting steps. We can use here all the knowledge employed in rate-limiting step analysis, in addition to two basic assumptions:

1. Inhibition alone can achieve all necessary rate changes. Genetic effects have as their sole purpose the preservation of cell resources.

2. The rate of an inhibited reaction is affected by a factor of roughly 10 to 200 when the inhibition is fully effective, and by a factor of 2 to 10 when the inhibition is fully ineffective.
In Section 6.2.4, we construct an O[M] regulation hypothesis similar to a rate-limiting step hypothesis, based on these two assumptions. Testing of candidate regulatory mechanisms is not independent from their generation, and some of the knowledge mentioned in the generation of candidates should also be used in testing each individual candidate.

4.5.4.4. Difficulties

The diagnosis of mutations is a very hard problem. Typically, in a full mutation-selection program, there are many cycles of mutation and screening. Since even the initial microorganism is usually little studied (except for its yield of the product of interest), the intermediate and final mutants are virtually black boxes. We need to know a lot about their internal processes before we can do the educated guessing implied in the previous paragraphs.

Much of the discussion in this section is given with wild-type E. Coli in mind, because it is the best-studied microorganism — much Biology research has focused on E. Coli. But in Industrial practice it is mostly multiple mutants of other strains that are used in bioprocesses. They were selected experimentally through screening of strains and mutation-selection programs, without any detailed study in the direction of understanding their mechanisms. Thus, for the industrially used microorganisms, even less knowledge is available for the starting strains, and the diagnosis problem is even harder.
4.6. PATHWAY RECOMMENDATION CONTEXT

The Recommendation Context should wrap up the results of the Evaluation Context, with some importance classification and confidence-analysis of the results from the point of view of biochemical engineering. It must also compute other quantities of interest, such as maximum yield or productivity. Finally, it must be able to compare alternative candidate pathways (or a given pathway to its hypothetical improved versions) giving high-level results.

The recommendations of improvements to the pathway, based on all the results, is the main goal. This interpretation of results and the recommendation for improvements could be achieved by a rule-based module.

The improvements to a pathway are based on the simple idea that a bottleneck must either be broken (if it stems from regulation) or bypassed (if it stems from kinetics or thermodynamics). To bypass a bottleneck, one only needs to synthesize pathways achieving the transformation of intermediates upstream from the bottleneck to intermediates downstream from the bottleneck. This idea is demonstrated in the case study of Sections 10.3 and 10.4.

4.6.1. INDUSTRIAL ASPECTS

When we find a pathway leading to a product P from reasonable starting materials, and evaluate it with respect to bottlenecks and regulation, we can conclude, at best, that the pathway is promising, but we cannot claim that P can actually be industrially produced. There are hundreds of compounds in a cell occurring in biochemical pathways that "look suitable," but most of them cannot really be produced industrially.
The characteristics that determine the industrial feasibility of a pathway lie mostly outside the pathway. A number of processes greatly influence the industrial feasibility of a bioprocess, yet we do not account for them at all in our framework. These processes include:

- Transfer of the substrates through the membrane into the cell [Muir et al., 1985, Nikaido and Vaara, 1985, Scarborough, 1985]
- Transfer of the products out of the cell
- Separation of the desired product from the fermentation broth
- Separation and utilization of by-products
- Separation and recycling of unused raw materials

It is only preliminary analysis, of aspects purely biochemical, that the whole approach discussed in this thesis is aimed at. However, each of the factors listed above could be taken into consideration by collecting appropriate knowledge and methods. It is fair to say that each of these aspects is probably easier to take into account than the purely biochemical view addressed in this work. The multitude of tasks would, of course, require considerable effort.

4.6.2. GENETICS

It is extremely hard to reason on the genetic control, because there is usually almost no knowledge available on its structure and parameters. Unlike the aspects of Section 4.6, genetic aspects are much harder than biochemical ones, because we lack the necessary knowledge and models.

It is, however, reasonable to expect that as a more substantial body of knowledge on the regulation of enzyme synthesis accumulates, it will be possible to develop methodologies that include genetic considerations in the evaluation of biochemical pathways.
4.7. LEARNING CONTEXT

With a learning mechanism, the results of the Evaluation and Prediction Context, as well as the refinements of the Recommendation Context, could be intelligently analyzed and summarized, taking into account experience and analogies.

Certain learning methods [Langley, 1979, Langley et al., 1983, Dietterich, 1984, Dietterich and Michalski, 1985, Mitchell et al., 1983], appear applicable to biochemical systems, but they must be converted to work with the semi-quantitative knowledge most frequently available. It should be noted that biochemical pathways behave approximately like systems without state when the enzyme concentrations are constant, i.e., in a short to medium time-scale. In a long time-scale when enzymes can be produced and degraded, the behavior of a pathway is strongly affected by its state, and thus by its past history.

Although most of the learning can be procedure-based, at the very top level a rule-based module may be the best way to express the results in the form most familiar to the user.

4.7.1. SEARCH FOR BIOLOGICAL REGULARITIES

The regularities that often occur in biological systems, are extremely useful in biochemical engineering [Erickson, 1979], but cannot be normally predicted in advance. For example, the elemental composition of bacterial biomass is approximately constant [Frankena et al., 1985] and, although a priori arguments could be made to suggest that such a regularity might exist, only experimental confirmation can show that the regularity is present.
Thus, the discovery of biological regularities requires iterative application of these steps:

- Postulation of potential regularities
- Checking the assumed regularities against available
- Confirmation, rejection, or correction of the assumption

This search for regularities could be undertaken by a learning system that knows how to use the Biochemical Database.

The search involves both continuous and discrete variables, so a combination of AI techniques for continuous variables [Langley et al., 1983] and techniques for discrete variables [Mitchell et al., 1983] would be needed. Such a combination might be achieved using interval arithmetic.
CHAPTER 5

FORMAL REASONING

WITH ORDERS OF MAGNITUDE

AND APPROXIMATE RELATIONS
5.1. INTRODUCTION

5.1.1. HIERARCHY OF MODELS

Models that are used in the description of engineering systems can be classified, according to their level of abstraction or degree of quantitativity, in the following hierarchy:

- Boolean models identify the system parameters or factors. The relations that are represented indicate the existence of constraints among parameters. Thus, information about which parameters may affect other parameters is available, but no information about the direction or magnitude of such effects is present. There is similarly no information available on the sign and magnitude of the parameters themselves. The incidence matrix of constraints and variables is an example of a Boolean model. Information indicating what metabolites participate in each bioreaction can also be characterized as boolean.

- Qualitative models represent not only the existence of parameters and interrelations, but also the signs of variables and the direction in which each variable affects other variables. However, information about magnitudes, or even relative orders of magnitude is not represented. A signed digraph showing how parameters affect each other, used frequently in Fault Diagnosis, is an example of a Qualitative model.
• In Order-of-Magnitude models the rough (absolute or relative) magnitudes of parameters (as well as the magnitudes of the effects of a parameter on other parameters) come into play. Until recently, these models have been only implicit mental models, because a formalism suitable for capturing Order-of-Magnitude commonsense did not exist. The O[M] system that we present here aims to formalize such intuitive engineering models.

• Quantitative models employ the most detailed numerical and algebraic representations, such as systems of (algebraic or differential) equations and numerical values of parameters. Quantitative models have traditionally been the norm in engineering analysis.

5.1.2. QUALITATIVE REASONING

Many efforts have been recently made to apply Qualitative Reasoning (Section 3.3.1.2) to engineering problems. Qualitative Reasoning [Bobrow, 1985] is a well-understood Artificial Intelligence approach for dealing with qualitative models; the approach involves the use of the signs of parameters as qualitative values, and constraints among signs (qualitative constraints) to describe the behavior of a physical system.

Substantial difficulties that are encountered in Qualitative Reasoning about Physical Systems, particularly in engineering applications, stem from the ambiguities inherent [de Kleer and Brown, 1984] in the qualitative values (-, 0, +) used. The incorporation of inequality relations through the quantity-space notion [Forbus, 1984] only partially resolves the ambiguities. Furthermore, Qualitative Reasoning gives too much attention to values of single parameters and neglects the importance of relations among parameters.

In engineering problems, apart from signs of quantities there is additional partial knowledge available about absolute value ranges, relative orders of magnitude of quantities, even approximate numerical values. General Qualitative Reasoning systems intentionally neglect most of this knowledge about parameters, because they target reasoning at a very abstract level.

Abstracting algebraic constraints all the way to qualitative constraints among signs, also degrades the knowledge that the constraints originally contained. Most of the inferences that can be made with these abstracted constraints correspond to "layman" reasoning and fail to capture the expert engineer's rules, or aid the engineer within a computer-aided paradigm. Qualitative Reasoning also excludes completely the manipulation of numbers, and any information more accurate than signs remains thus unused.

All these points are not shortcomings of the original Qualitative Reasoning efforts, which in fact intentionally focused on naive physics rather than engineering applications. But when employing qualitative or semiquantitative reasoning in engineering applications, it is desirable to introduce a level of abstraction more accurate than the qualitative values, in order to achieve more expressive power and effective reasoning.
5.1.3. ORDER-OF-MAGNITUDE CONCEPTS IN THE ENGINEERING ANALYSIS OF BIOCHEMICAL SYSTEMS

If one attempts to classify the knowledge available on biochemical systems (as well as implicit mental models and reasoning), in terms of the modelling hierarchy outlined at the beginning of this chapter, one finds that they are often Order-of-Magnitude in nature.

5.1.3.1. Knowledge

In the engineering analysis of biochemical systems, which are the focus of this work, one can examine the relative orders of magnitude of parameters and use verbal descriptions denoting approximate relations among parameters, as in the following examples:

- In an exponentially growing culture of bacteria, [ATP] is much higher than [ADP]; when [ATP] becomes approximately equal to [ADP], bacteria die due to lack of energy. These particular relations may appear contrived because it is customary to use the Energy Charge [Ingraham et al., 1983] to represent the [ATP]/[ADP] proportion (the above relations do reflect the accepted ranges for the Energy Charge). However, similar rough relations (i.e., correlations between parameters) can be produced for the concentrations of many species and, since it is not practical to define special ratios and terms to represent them, one must resort to using these relations directly.
• For anaerobic growth, the flux through the biosynthetic part of the metabolism is much lower than the flux through the energy metabolism.

• A reaction is near equilibrium when its mass action ratio is only slightly lower than its equilibrium constant. If the mass action ratio is higher than the equilibrium constant, the reaction can only take place in the reverse direction.

5.1.3.2. Reasoning

The approximate relations are not necessarily known a priori. They are frequently derived from other approximate relations, algebraic equations, inequalities, or empirical rules.

If, for example, it is known that [ATP] is much higher than [ADP], but approximately equal to [GTP], it can be concluded that [GTP] is much higher than [ADP]. Considering that the total phosphorous concentration is higher than the sum of [ATP], [GTP], and [ADP], it can be further reasoned that [ATP] and [GTP] are significantly lower than the total phosphorous concentration.

Approximate relations can also be used to infer other characteristics of the system. For example, knowing that a reaction's mass action ratio is much lower than its equilibrium constant, it can be concluded that the reaction is irreversible and a potential target for regulation.
5.1.4. NEED FOR FORMALIZATION

We saw that the use of verbal expressions describing approximate relations can bring important knowledge to bear in the analysis of biochemical systems. However, the vague and informal character of such expressions can cause problems:

- There is always the risk of misunderstanding or misusing a relation. What appears "slightly smaller" in one instance, may be "significantly smaller" in another.

- Many complex biological systems, such as large metabolic networks, can only be analyzed by computer, but there is no convenient way to use verbal, informal relations in traditional computer programs. Even in rule-based expert systems, it would be quite cumbersome to encode all the implicit rules necessary for reasoning with the verbal approximate relations, especially since arbitrary algebraic expressions may arise in the models of biochemical systems.

To overcome the above difficulties, we are thus led into Formal Order-of-Magnitude Reasoning, a novel Artificial Intelligence approach that offers exactly the necessary intermediate abstraction level between Qualitative Reasoning and precise quantitative models. This approach is based on the representation of the relative orders of magnitude of the parameters of a system, through the definition of relations among quantities.

Order-of-Magnitude relations formally capture semiquantitative concepts that are implied by engineering statements of the kind "A is much smaller than B," "C is of the same order as D," or "E is slightly larger than A." Inferences on such statements
stem from common sense of the kind "if A is much smaller than B, and E is only slightly larger than A, then E must be much smaller than B," or "if B is much larger than E, and F is larger than B, then F is much larger than E."

The problem of formal Order-of-Magnitude reasoning about Biochemical systems is addressed here through the O[M] formalism. This chapter aims to provide the basics of the formalism and demonstrate how informal knowledge, notions, and inferences can be formalized in O[M]. Applications of Order-of-Magnitude reasoning in (non-biochemical) Process Engineering systems and activities will also be discussed in this chapter (Section 5.10).

Applications of the formalism in biochemical systems, including the Order-of-Magnitude modelling of enzyme inhibition, analysis of fluxes in a biochemical network, and identification of rate-limiting steps of pathways, will be discussed in Chapter 6.
5.2. PREVIOUS WORK

Several forms of Order-of-Magnitude ideas have already been used in Artificial Intelligence. Most are only indirectly related to the philosophy of the formalism we present here, but one particular effort, the FOG system, has the same goals as our formalism, O[M].

5.2.1. AI CONCEPTS INDIRECTLY RELATED TO ORDERS OF MAGNITUDE

Order-of-Magnitude concepts have been examined as means for algebraic simplification [McAllester, 1981]. The aim of the work was to simplify expressions by deriving the order, with respect to each variable, of each term of an expression.

In a diagnostic system for digital circuits [Davis, 1984, Hamscher, 1984], hierarchic representation of time uses several time granularities; the longest delay until quiescence at the finer level determines how many fine-grain units correspond to one coarse-grain unit.

A similar concept in qualitative reasoning is the concept of mythical time [de Kleer, 1984], a finer time granularity that can distinguish cause and effect among simultaneous events; events that occur simultaneously in "real" time are allowed to occur sequentially in mythical time, the cause temporally preceding its effects.

Underlying time granularities and mythical time, is the notion of different orders of magnitude in time scales. More recently, abstraction by time-scale was explicitly proposed as a way to reduce ambiguities in Qualitative Simulation [Kuipers, 1987].
5.2.2. THE FOG SYSTEM

In the FOG Formal Order of MaGnitude System [Raiman, 1986, Dague et al., 1987], explicit Order-of-Magnitude inferences were first introduced not just for time-scales but for any kind of variables. The FOG system is based on three relations:

- A Ne B: A is negligible in relation to B.
- A Vo B: A is close to B (and has the same sign as B).
- A Co B: A has the same sign and order of magnitude as B.

The FOG system has 30 rules of reasoning [Raiman, 1986] involving its basic relations, classical qualitative values, addition, and multiplication. For example a FOG rule states:

\[(A \ Co \ B) \text{ and } (B \ Co \ D) \Rightarrow A \ Co \ D\]  \hspace{1cm} (5.1)

5.2.2.1. Deficiencies of FOG

Unfortunately FOG has several disadvantages particularly important in engineering applications:

(a) It does not provide concrete interpretation of its relations. If one does not intuitively understand what "A Co B" means, there is no further explanation available. This difficulty is particularly bothersome because the Order-of-Magnitude relations have different meanings for different engineering activities. The relation "Negligible" does not mean the same thing in a preliminary-design context and a detailed-design context.
(b) The set of rules that FOG employs appears arbitrary, and it is not clear how it can be extended, e.g., to exponentials or trigonometric functions.

(c) It does not allow incorporation of partial quantitative information. Such information is usually available and very useful in the solution of engineering problems. FOG cannot use such information because it does not relate numbers to Order-of-Magnitude relations. For example, if FOG is told that "A Vo 0.1" and "B Vo 1000" it is unable to infer the obvious "A Ne B".

(d) FOG lumps signs and magnitudes in the same relations. The relation "A Co B" carries unnecessary sign connotations, contrary to the engineer's intuitive Order-of-Magnitude notion. In many engineering situations, the signs of parameters depend on conventions rather than fundamental physical reasons. For example, pressure differences along pipes or around process equipment may have positive or negative signs (depending on case-specific conventions), but the engineer may nevertheless have knowledge about their magnitudes.

(e) FOG requires the explicit use of negation and disjunction, along with inequality relations, to express even very simple and frequently used relations, such as > and <. In fact, in the application of FOG in troubleshooting [Dague et al., 1987] two additional relations were introduced (named ~+ and ~-) expressed in terms of the original FOG relations and inequalities. It is desirable to have such
relations be present a priori in a formalism and avoid combinatorially explosive means of combination.

### 5.2.2.2. Need for a New Formal System

The previous discussion shows that there is a need for a formalism that can improve qualitative reasoning by capturing Order-of-Magnitude notions that arise in engineering, but the deficiencies of FOG do not allow it to fulfill this role. In this work, we address the problem of applying qualitative and semiquantitative reasoning in engineering with the O[M] formalism [Mavrovouniotis and Stephanopoulos, 1987] for reasoning about orders of magnitudes and approximations.

The O[M] formalism lacks the basic deficiencies of FOG, and has the potential to be used in many engineering applications.

Before presenting O[M], we will discuss the advantages of Order-of-Magnitude reasoning in engineering tasks and activities. This discussion should point out what characteristics of O[M] the reader should look for in the description of the formalism and its applications. In the rest of the chapter, we will provide a description of the formalism and an overall evaluation of the method's expressiveness and potential. In this chapter the discussion will be confined to O[M] as a general formal system; detailed examples involving applications of O[M] will be presented in the next chapter of the thesis.
5.3. ADVANTAGES OF FORMAL ORDER-OF-MAGNITUDE REASONING

With a formalization and automation of Order-of-Magnitude Reasoning, many advantages are realized:

- Previously informal notions and methods can be described and communicated concretely. This will be demonstrated in the next chapter, where we will formalize the definition of a rate-limiting step. The O[M] formal system provides a vocabulary for the formalization of expert knowledge, such as knowledge on the relative magnitudes of fluxes through a biochemical network, as will be discussed in the example on pathway competition.

- Available quantitative knowledge can be used along with semiquantitative knowledge. One of the motives for using Artificial Intelligence methods is the desire to apply as much of the available knowledge as possible, despite the disparity in the forms of knowledge involved.

- Order-of-Magnitude Reasoning can be performed by a computer capable to handle complex systems such as large biochemical networks, whose analysis is too tedious and time-consuming for people.

- In the application of other Artificial Intelligence methods, such as the technology of Knowledge-Based Expert Systems, formal Order-of-Magnitude Reasoning achieves the representation and use of semiquantitative knowledge on which experts rely, and can
Incorporate deep, mathematical models in reasoning systems that are primarily empirical.

We must caution however that Order-of-Magnitude Reasoning should not be viewed as a substitute for numerical simulation currently used in various detailed-design activities. If full numerical data and models are available and full numerical results are required, classical quantitative simulation is the preferred alternative. Neither Order-of-Magnitude Reasoning nor any form of qualitative reasoning is of much value in such a case.

To identify the exact domains where the O[M] formalism may be applied, we need only look for domains in which Order-of-Magnitude arguments are employed very frequently, even if this is done only implicitly. Order-of-Magnitude knowledge is particularly relevant in the analysis of biochemical systems, because quantitative knowledge is harder to obtain while qualitative knowledge is used extensively by the experts. Furthermore, some concepts that are very important in pathway analysis, such as the notion of a rate-limiting step or the notion of a feedback regulatory mechanism, involve Order-of-Magnitude ideas.

In delineating the O[M] formalism, we begin with the representation of quantities and Order-of-Magnitude relations that can relate them. We discuss in detail the strict and heuristic interpretation of Order-of-Magnitude relations, as well as other properties of the relations. After we mention some additional knowledge-representation elements, namely assignments, constraints, and rules, we will discuss how inferences in O[M] are guided, constrained, and maintained, with emphasis on the options available in the resolution of contradictions.
5.4. QUANTITIES AND LINKS

Quantities in the O[M] formalism are all the specific parameters of a particular physical system. Below, we will distinguish between variables, whose value is unknown, and landmarks, which have known constant values.

5.4.1. VARIABLES AND SIGN SPECS

A variable in O[M] refers to a specific physical quantity, with known physical dimensions but unknown numerical value.

Knowledge about the sign (-, 0, +) of the variable is kept as assertions termed sign specs, stored within the variable. In biological systems the signs of most variables are known from their definition. For the few parameters whose sign is unknown, the mechanisms of O[M] provide the ability to derive the sign from approximate relations and algebraic constraints.

5.4.2. LANDMARKS

A landmark is similar to a variable, but it has known (constant) sign and value. Two quantities are compatible if they have the same physical dimensions. To facilitate definitions of quantities and the search for compatible quantities we let physical dimensions and units be separate entities.

5.4.3. LINKS

Within each quantity, there are links, each representing a compatible pair of quantities (which cannot both be landmarks) that can be interrelated.
A link contains all the Order-of-Magnitude relations asserted between the two quantities, and information on where such relations can be obtained from and where they can be used (e.g., relevant constraints and rules, entities we will describe later).
5.5. PRIMITIVE AND COMPOUND RELATIONS

Order-of-Magnitude relations relate the absolute magnitudes of quantities, without reference to their sign. Thus, there is no interference between signs and magnitudes, so that reasoning with signs can be performed under the usual Qualitative Reasoning principles.

5.5.1. PRIMITIVE RELATIONS

We introduce seven primitive binary relations among quantities, as shown in Table 5.1.

Some of the informal notions discussed earlier can now be expressed through primitive relations. It can be stated, for example, that [ATP] \( \gg \) [ADP] (i.e., [ATP] is much higher than [ADP]) and that for a near-equilibrium reaction \( Q \sim K_e (\text{i.e.}, Q \text{ is slightly smaller than } K_e) \) where \( Q \) is the mass action ratio and \( K_e \) the equilibrium constant of the reaction.
**Table 5.1:**
Primitive relations of the O[M] formalism.

<table>
<thead>
<tr>
<th>O[M]-RELATION</th>
<th>VERBAL EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_1$:</td>
<td>$A &lt;&lt; B$</td>
</tr>
<tr>
<td>$r_2$:</td>
<td>$A -&lt; B$</td>
</tr>
<tr>
<td>$r_3$:</td>
<td>$A ~&lt; B$</td>
</tr>
<tr>
<td>$r_4$:</td>
<td>$A == B$</td>
</tr>
<tr>
<td>$r_5$:</td>
<td>$A &gt;~ B$</td>
</tr>
<tr>
<td>$r_6$:</td>
<td>$A &gt;- B$</td>
</tr>
<tr>
<td>$r_7$:</td>
<td>$A &gt;&gt; B$</td>
</tr>
<tr>
<td></td>
<td>A is much smaller than B</td>
</tr>
<tr>
<td></td>
<td>A is moderately smaller than B</td>
</tr>
<tr>
<td></td>
<td>A is slightly smaller than B</td>
</tr>
<tr>
<td></td>
<td>A is exactly equal to B</td>
</tr>
<tr>
<td></td>
<td>A is slightly larger than B</td>
</tr>
<tr>
<td></td>
<td>A is moderately larger than B</td>
</tr>
<tr>
<td></td>
<td>A is much larger than B</td>
</tr>
</tbody>
</table>
5.5.2. COMPOUND RELATIONS

The approximate relation that holds between two parameters may be known with less precision than the primitive relations imply. It may be known, for example, that [ATP] is approximately equal to [GTP], without information on which of the primitives \( \sim< \) (slightly smaller), \( == \) (exactly equal), and \( >\sim \) (slightly larger) may hold.

Such knowledge can be expressed through compound relations, formed as implicit disjunctions of any two or more consecutive primitive relations. The relation "[ATP] is approximately equal to [GTP]" can be expressed by the compound relation [ATP] \( \sim<\ldots\sim \) [GTP], i.e., "[ATP] is from slightly lower up to slightly higher than [GTP]." The relation \( \sim<\ldots\sim \) combines the primitives \( \sim<,==, \) and \( >\sim \).

In general, we accept as a compound relation any implicit disjunction of two or more consecutive primitive relations. The primitive relations combined must be consecutive; \( <, \sim<, \) and \( \sim< \), for example, can be combined to yield the equivalent of the "less than" \( (\sim<) \) relation; as a counterexample, \( >\sim \) and \( \sim< \) cannot be combined. The notation for a compound relation produced from all primitives from \( r_n \) through \( r_m \) is \( r_n..r_m \), where the order of primitives must always be as indicated in Table 5.1. It should be emphasized that this restricted disjunction refers mainly to the semantics of the relations, and there are no provisions for explicit syntactic disjunction in the formalism. There are in total 21 compound relations, shown in Table 5.2.

5.5.3. RESTRICTIONS

As stated above, this notation does not allow the combination of primitives that are not consecutive. It is not possible to state, for example, that [ATP] is either much lower or slightly higher than [GTP], because the primitives implied here (\( \sim< \) and \( >\sim \))
are not consecutive. This restriction is justified by the fact that such relations are counterintuitive and in most cases useless. Prohibiting them altogether guards the formalism against computational complexity.

The 7 primitive relations and the 21 compound relations give a set \( L \) with a total of 28 legitimate relations \( r_1, \ldots, r_{28} \). To keep complexity manageable, disjunction and negation are not present in the formalism. Thus, in \( Q[M] \) you cannot say that \( A \) is "either much smaller or slightly larger" than \( B \), and you cannot say that \( A \) is "not slightly smaller" than \( B \). These restrictions assure the system a simple constant-length form for all the assertions it is handling.
**Table 5.2: Compound relations of the O[M] formalism.**

<table>
<thead>
<tr>
<th>COMPOUND RELATION</th>
<th>O[M]</th>
<th>PRIMITIVES</th>
</tr>
</thead>
<tbody>
<tr>
<td>moderately to much less than</td>
<td>&lt;&lt;...&lt;&lt;</td>
<td>&lt;&lt;...&lt;&lt;</td>
</tr>
<tr>
<td>less than</td>
<td>&lt;&lt;...&lt;</td>
<td>&lt;&lt;...&lt;</td>
</tr>
<tr>
<td>less than or equal to</td>
<td>&lt;&lt;...==</td>
<td>&lt;&lt;...&lt;...&lt;==</td>
</tr>
<tr>
<td>less than or approximate equal to</td>
<td>&lt;&lt;...&gt;~</td>
<td>&lt;&lt;...&lt;...&lt;==...&gt;~</td>
</tr>
<tr>
<td>less than or comparable to</td>
<td>&lt;&lt;...&gt;~</td>
<td>&lt;&lt;...&lt;...&lt;==...&gt;~</td>
</tr>
<tr>
<td>may bear any relation to (trivial relation)</td>
<td>&lt;&lt;...&gt;~</td>
<td>&lt;&lt;...&lt;...&lt;==...&gt;~</td>
</tr>
<tr>
<td>moderately or slightly less than</td>
<td>&lt;=...&lt;</td>
<td>&lt;=...&lt;</td>
</tr>
<tr>
<td>from moderately less up to equal to</td>
<td>&lt;=...==</td>
<td>&lt;=...&lt;...==</td>
</tr>
<tr>
<td>moderately less than or approximately equal to</td>
<td>&lt;=...&gt;~</td>
<td>&lt;=...&lt;...==</td>
</tr>
<tr>
<td>comparable to</td>
<td>&lt;=...&gt;~</td>
<td>&lt;=...&lt;...==...&gt;~</td>
</tr>
<tr>
<td>greater than or comparable to</td>
<td>&lt;=...&gt;~</td>
<td>&lt;=...&lt;...==...&gt;~</td>
</tr>
<tr>
<td>slightly less than or equal to</td>
<td>=~...==</td>
<td>=~...==</td>
</tr>
<tr>
<td>approximately equal to</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>moderately larger than or approximately equal to</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>larger than or approximately equal to</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>slightly larger than or equal to</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>from equal to up to moderately larger than</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>greater than or equal to</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>slightly or moderately larger than</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>greater than</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
<tr>
<td>moderately or much greater than</td>
<td>=<del>...&gt;</del></td>
<td>=<del>...&gt;</del></td>
</tr>
</tbody>
</table>

---
5.5.4. EXPRESSIVENESS

Despite the restrictions, this relation set provides substantial expressive power. The inverse of every legitimate relation is always a legitimate relation. The negation of a legitimate relation is a legitimate relation if and only if that relation includes either of $\gg$ or $\ll$.

The relations are powerful enough to express quantity-space partial ordering, all of FOG's relations, and other relations that engineers use in Order-of-Magnitude arguments.

For example the relation "less than or approximately equal to" ($\preccurlyeq$), frequently used in engineering, is expressed as $\ll\ldots\ll$, and its negation as $\gg\ldots\gg$. The relation A "roughly equal to" is expressed as $\sim\ll\ldots\ll$, but its negation cannot be expressed. This is not truly limiting in practice, because we would seldom know that A and B are "not roughly equal" without also knowing which of the two variables is smaller than the other; once we know this kind of inequality information the relation becomes expressible (if A<B then A $\ll\ldots\ll$ B; if A>B then A $\gg\ldots\gg$ B). These types of arguments underlie our choice to sanction compound relations, as we described them, but exclude other forms of combining relations.

Table 5.3 shows the correspondence of some relations to commonsense relations frequently used informally. FOG cannot express many of the relations in the table without the use of inequality relations, disjunction, and negation.
Table 5.3:

O[M] relations representing relations that are commonly used in engineering.

<table>
<thead>
<tr>
<th>CLASSICAL COMMONSENSE RELATIONS</th>
<th>O[M]</th>
</tr>
</thead>
<tbody>
<tr>
<td>less than (&lt;)</td>
<td>&lt;&lt;..&lt;~&lt;</td>
</tr>
<tr>
<td>less than or equal to (≤)</td>
<td>&lt;&lt;..==</td>
</tr>
<tr>
<td>greater than (&gt;)</td>
<td>&gt;..&lt;&gt;&gt;</td>
</tr>
<tr>
<td>greater than or equal to (≥)</td>
<td>==..&gt;&gt;</td>
</tr>
<tr>
<td>equal to (=)</td>
<td>==</td>
</tr>
<tr>
<td>approximately equal to (=)</td>
<td><del>&lt;..&lt;&gt;</del></td>
</tr>
<tr>
<td>less than or approximately equal to (≤)</td>
<td>&lt;&lt;..&lt;&gt;~</td>
</tr>
<tr>
<td>greater than or approximately equal to (≥)</td>
<td>~&lt;..&lt;&gt;&gt;</td>
</tr>
<tr>
<td>much less than</td>
<td>&lt;&lt;</td>
</tr>
<tr>
<td>much greater than</td>
<td>&gt;&gt;</td>
</tr>
</tbody>
</table>
5.6. SEMANTICS AND PROPERTIES OF O[M] RELATIONS

We will provide here concrete semantics for the Order-of-Magnitude relations we defined in the previous section. There are two possible semantic interpretations of the relations:

- A strict interpretation which guarantees correct results but is too weak to imitate human commonsense reasoning
- A heuristic interpretation which is more aggressive and human-like but cannot exclude errors

5.6.1. STRICT INTERPRETATION

In this first interpretation we will discuss here, each relation $A \preceq B$ is equivalent to the relation $(A/B) \preceq 1$ and signifies an interval for the ratio $(A/B)$, as indicated in Figure 5.1, with $e_1$ through $e_4$ the boundaries of the intervals. So far, the only commitment we have made is that $A \preceq B$ corresponds to $A/B = 1$. 
Figure 5.1:
Strict interpretation of the relation $A_r^n B$. 
5.6.1.1. Determination of the Boundaries of the Intervals

To sanction the symmetry of the relations

\[ A \sim B \iff B \lessdot A \quad (5.2) \]

\[ A \gg B \iff B \ll A \quad (5.3) \]

we impose the restrictions \( e_3 = 1/e_2 \) and \( e_4 = 1/e_1 \) (Figure 5.1 was drawn to a logarithmic scale, taking into account these restrictions). To sanction the intuition that, for \( A > B > 0 \):

\[ A \lessdot B \iff A \sim B \quad (5.4) \]

we further impose the restriction \( e_3 \cdot 1 = e_1 \).

Under this strict interpretation, the above constraints leave only one degree of freedom for the semantics of our relations, as depicted in Figure 5.2. With respect to the openness of the intervals (i.e., assigning the boundaries to intervals) the requirements we placed allow only two solutions, but the choice has no practical effect. We choose to make only the intervals \(-<, >-, \) and \(=\) closed. The end result is a set of 7
exhaustive and mutually disjoint intervals for the 7 primitive relations.
Figure 5.2:

Constrained strict interpretation of the relation $A \rhd_{n} B$. 
5.6.1.2. Remaining Degree of Freedom

We leave the tolerance parameter $e$ unspecified because it depends on the particular engineering activity. The value of $e$ denotes the largest amount which we are willing to regard as "much smaller than 1." In the preliminary design of chemical processes for example, the designer tends to think of $e$ between 0.05 and 0.20. On the other hand during detailed design the designer would require $e$ to be less than 0.05.

In the examples in the rest of this chapter, unless we specifically indicate otherwise, the value of $e$ will be assumed 0.1, which corresponds to the common idea that one order of magnitude denotes roughly a factor of 10.

5.6.1.3. Advantages and Disadvantages of the Interpretation

With this clear semantics there is no need for prespecified rules since they can be derived from the intervals, which moreover allow incorporation of quantitative information. We named this interpretation strict because its solid intervals support only accurate correct inferences.

Since the semantics of a relation are based on the ratio of the two operands, the property $A \sim B \Leftrightarrow (KA) \sim (KB)$, for $K \neq 0$, which was implied earlier, is preserved by the semantics. Note that adding (rather than multiplying) a constant into both sides of relation does not necessarily preserve the validity of the relation. In other words, it is not, in general, the case that: $A \sim B \Leftrightarrow (K+A) \sim (K+B)$.

The strict interpretation is accurate, but too strict compared to human reasoning. For example, from the relations $A >\sim B$ and $B >\sim C$ the strict interpretation can only conclude $A >\sim_2 >\sim C$ while human commonsense would aggressively conclude $A$
Clearly the latter result is heuristic. It is not guaranteed correct, but it is correct
often enough for an engineer to be happy with. Any mechanism that can accommodate this
result will have to accept the risk of wrong conclusions as the price for more aggressive
focused inferences. Note that FOG sanctions the even more aggressive inference:

\[(A \text{ Co } B \text{ and } B \text{ Co } C) \Rightarrow A \text{ Co } C\]  \hspace{1cm} (5.5)

a subcase of which in our notation would be

\[(A \Rightarrow B \text{ and } B \Rightarrow C) \Rightarrow A \Rightarrow C\]  \hspace{1cm} (5.6)

We feel this inference is too aggressive and error prone. We do not want the
O[M] formalism to be this risky, so we choose not to sanction this inference.

5.6.2. HEURISTIC INTERPRETATION

In order to remedy the deficiencies of the strict interpretation, we adopt a
heuristic interpretation that replaces the boundary points of the intervals with regions,
as shown in Figure 5.3.
Figure 5.3:
Vague interval boundaries for the heuristic interpretation.
5.6.2.1. **Construction**

We construct two sets of primitive intervals: A set of non-exhaustive but mutually disjoint intervals and a set of exhaustive but overlapping intervals, as shown in Figure 5.4. The following heuristic inference convention is adopted: For every inference step, assume the antecedent relations to denote non-exhaustive intervals, but allow the consequent relations to denote overlapping intervals. Thus, when the consequents are used as antecedents at a later step their intervals are "shrunk" and therein lies the power and the risk. Note that for compound relations this mechanism refers only to the end points of the compound intervals. In other words, the compound intervals expand and shrink at their ends, but they never have holes.

The good properties that were mentioned for the strict interpretation are preserved by this transformation, with one exception: The accuracy of the inferences, guaranteed by the strict interpretation, is sacrificed to make the heuristic inference procedure closely resemble human reasoning. In the previous example it would aggressively infer:

\[ A \sim B \text{ and } B \sim C \Rightarrow A \sim C \] (5.7)

Once an inference is made, people use the consequent without reconsidering its uncertainty and would use the result in (5.7) to infer further:

\[ A \sim C \text{ and } C \sim D \Rightarrow A \sim D \] (5.8)

Hence when a consequent is used further the expanded intervals must be contracted.
Figure 5.4:

Intervals used in the heuristic interpretation.

Top:  Non-exhaustive, mutually-exclusive intervals.

Bottom: Exhaustive, overlapping intervals.
5.6.2.2. **Determination of the Boundaries**

To choose the new interval boundaries, we sanction the symmetry of the relations, as before, and the following inferences:

\[ A >- B \Rightarrow A - B < B \quad (5.9) \]

\[ A >- B \text{ and } B >- C \Rightarrow A >- C \quad (5.10) \]

\[ A >- B \text{ and } A >> C \Rightarrow B >> C. \quad (5.11) \]

The interval boundary regions in the final form are shown in Figure 5.5. Note that the interval boundaries include the boundaries of the strict interpretation, so that strict inferences are also valid heuristic inferences. In order to maintain the ordering of the boundaries as shown, there is a restriction on the value of \( e \):

\[ e - 1 > (1 + e)^2 \Rightarrow e < 0.4656 \quad (5.12) \]

5.6.2.3. **Other Inferences**

A very large number of inferences are valid regardless of the exact value of \( e \). Apart from inferences based on addition and subtraction, this group also includes inferences with other functions, e.g.

\[ x << 1 \Rightarrow \exp(x) >- 1 + x \quad (5.13) \]

\[ x << 1 \Rightarrow \sin(x) >- x. \quad (5.14) \]
Figure 5.5:
Final intervals for the heuristic interpretation of O[M] relations.
5.6.3. SECONDARY PROPERTIES OF ORDER-OF-MAGNITUDE RELATIONS

In this section we will discuss some secondary properties of Order-of-Magnitude relations, which stem from the (strict or heuristic) interpretation of the relations but do not depend on the accuracy parameter e.

If P is the set of primitive relations, then these are ordered in P as shown in Table 5.1. Thus, \((r_1 \ b \ r_2)\) and \((r_2 \ b \ r_3)\), while both imply that \((r_1 \ b \ r_3)\), where "b" is the "before" relation defined in P. Let C be the set of compound relations, \(r_n \in C\), and \(p(r_n)\) be the ordered set of primitives corresponding to the relation \(r_n\). Then:

\[
p(r_n) = \{ r_i | [r_{\min}(r_n) \ b \ r_i] \land [r_i \ b \ r_{\max}(r_n)] \}
\]

(5. 15)

where \(r_{\min}(r_n)\) and \(r_{\max}(r_n)\) are the lower and upper limits delineating the ordered set \(p(r_n)\), with \((r_{\min}(r_n), r_{\max}(r_n)) \in P^2\) and \((r_{\min}(r_n) \ b \ r_{\max}(r_n))\). Obviously, if \(r_n \in P\) then \(p(r_n) = \{r_n\}\). Thus, \(p\) is defined over the whole set of legitimate relations \(L = P \cup C\).

5.6.3.1. Inverse Relations

Let \(\text{Inv}(r_n)\) be the inverse of relation \(r_n\). For any relations \(r_i, r_j \in L:\n
\[
\text{Inv}(r_i) = r_j \iff \text{Inv}(r_j) = r_i
\]

(5. 16)

For primitive relations, referring to Table 5.1, it is easy to see that \(\text{Inv}(r_1) = r_7\), \(\text{Inv}(r_2) = r_6\), \(\text{Inv}(r_3) = r_5\), and \(\text{Inv}(r_4) = r_4\).

If \(r_n \in C\), then:
\[ \text{Inv}(r_n) = p^{-1}(\{ r_i \mid \text{Inv}(r_i) \in p(r_n) \}) \] (5.17)

which allows the inverse of compound relations to be reduced to the inverse of primitives. Furthermore, it can be shown that we only need consider the two limits of a compound relation:

\[ \text{Inv}(r_n) = g^{-1}(\text{Inv}[r_{\text{max}}(r_n)], \text{Inv}[r_{\text{min}}(r_n)]) \] (5.18)

where:

\[ g(r) = (r_{\text{min}}(r), r_{\text{max}}(r)) \] (5.19)

### 5.6.3.2. Transitivity of Relations

Let \( x, y, z \) be real numbers. If \((x \; r_n \; y)\) and \((y \; r_j \; z)\), then \((x \; r_k \; z)\), where 
\( r_k = \text{Trans}(r_n, r_j) \) is the transitivity combination, \( \text{Trans}: L \to L \). In terms of the corresponding sets of primitives:

\[ \text{Trans}(r_n, r_j) = p^{-1}(\{ r_i \mid \\
\quad r_i = \text{Trans}(r_{p_n}, r_{p_j}), r_{p_n} \in p(r_n), r_{p_j} \in p(r_j) \}) \] (5.20)

which allows the transitive combination of compound relations to be reduced to the combination of primitives. As with inversion of relations, we only need consider the limits:

\[ \text{Trans}(r_n, r_j) = g^{-1}(r_{\text{min}}[\text{Trans}(r_{\text{min}}(r_n), r_{\text{min}}(r_j))] , \\
\quad r_{\text{max}}[\text{Trans}(r_{\text{max}}(r_n), r_{\text{max}}(r_j))]) \] (5.21)

The following properties also hold:

\[ \text{Trans}(r_n, r_j) = \text{Trans}(r_j, r_n) \] (5.22)

\[ \text{Trans}(\text{Inv}(r_j), \text{Inv}(r_n)) = \text{Inv}(\text{Trans}(r_n, r_j)) \] (5.23)
If \( r_n \in P \) and \((r_n \in \text{Inv}(r_n))\)

then \( \text{Trans}(r_n, \text{Inv}(r_n)) = g^{-1}(r_n, \text{Inv}(r_n)) \) \hspace{1cm} (5.24)

If \( \text{Inv}(r_n) = r_n, \text{Inv}(r_j) = r_j, \) and \( r_k = \text{Trans}(r_n, r_j) \)

then \( \text{Inv}(r_k) = r_k \) \hspace{1cm} (5.25)
5.7. OTHER MEANS OF KNOWLEDGE REPRESENTATION

While Order-of-Magnitude relations are the basis of the formalism, we must provide for additional constructs to represent algebraic or empirical knowledge and use it together with the relations.

5.7.1. ASSIGNMENTS

Assignments are "solved" algebraic relations that allow some quantities to produce relations among other quantities. For example:

\[ A = B + 3.5 \ C \ \exp (D/E) - 2 \ F \quad (5.26) \]

\[ (A/B) = 1 + 3.5 \ (C/B) \ \exp (D/E) - 2 \ (F/B) \quad (5.27) \]

The left hand side, called destination, can be a link or a variable. In Equation (5.27) the destination is a link, and the assignment can be used to produce relations for that link, i.e., relations between A and B. In Equation (5.26) the destination is a variable, so the assignment can produce a range for that variable (i.e., A) and, from that, relations between the variable and any compatible landmark.

The right hand expression of an assignment can be used most effectively if it involves only links, landmarks and numerical constants (as in Equation 5.27, but not Equation 5.26). Thus, algebraic expressions, must be converted to the most effective form, by mapping ratios of compatible quantities to links. For example, starting from the algebraic form (5.26), the assignment (5.27) can be derived, and used in parallel with the initial version. The success of the transformation may depend on the form of the algebraic expression and specific knowledge about the manipulation of algebraic expressions and functions. Between two algebraically equivalent forms one
may be much harder to transform than another. To illustrate this point, consider the assignment

$$G = F + A B^2 \ (1 - \sin^2(A/B) - \cos^2(A/B)) \ . \quad (5.28)$$

In order to transform this assignment to the mathematically equivalent

$$\frac{G}{F} = 1 \quad (5.29)$$

it is not sufficient to know how to compute the sine and cosine functions; knowledge on the existence and application of a particular trigonometric identity (or an equivalent set of identities) is needed. In the absence of such knowledge, relation (5.28) can still be used within O[M], but the resulting inferences will not be as effective.

5.7.2. CONSTRAINTS

Constraints are "unsolved" algebraic relations among quantities, such as

$$A + 2 F - B = 3.5 C \ \exp (D/E) \ . \quad (5.30)$$

As with assignments, there are requirements on the form of the algebraic expressions, and the system attempts automatic conversion. There are two ways in which constraints can be used within the O[M] formalism:

- The first way, is to simply substitute values for variables and links, and test the constraint. If the constraint is not satisfied, a contradiction has been found, and the assumptions on which the values were based can be rejected as inconsistent, as we will explain in the Truth-Maintenance section.

- The second way is to form a set of assignments by solving the constraint in all obvious ways. By "obvious" solutions we mean
simply getting hold of one occurrence of a variable (or link) in the expression and solving with respect to that, regardless of any other occurrences of the same variable. In this fashion the constraint (5.30) would yield the assignments (5.26) and (5.27). Of course, solving the constraint in more clever ways (specifically, solving it *completely* with respect to each variable) is desirable and, if achieved, will yield better results than the naive solution procedure outlined above.

5.7.3. RULES

Knowledge of highly empirical nature often cannot be expressed in algebraic form. Rules in O[M] allow such unstructured knowledge to be expressed in simple if-then statements. For example, for some specific variables A, B, and D, we may state the rule:

\[
\text{if } (A \ll B \text{ and } D \ll B) \text{ then } (A \sim D) \tag{5.31}
\]

Whenever the antecedents of the rule are found to hold, the consequents can be automatically inferred. In the current formalism, global quantification is not allowed in rule statements, so rules are not allowed to have free variables. The use of explicit empirical rules is only dictated by considerations of style and efficiency. One can equivalently assert the consequents of a rule as "true under the assumption of the antecedents" through the mechanisms of the Assumption-based Truth-Maintenance System, which we will be discussed in a later section.
5.8. REASONING

The reasoning carried out within O[M] is based on the use of Order-of-Magnitude relations along with other representational entities outlined above. Through the precise interpretation of the relations, O[M] offers mechanisms that exploit Order-of-Magnitude relations, algebraic constraints, and empirical rules.

5.8.1. EXAMPLES

- Given the relations

\[ [\text{GTP}] \sim \sim [\text{ATP}] \quad (5.32) \]

and

\[ [\text{ATP}] \gg [\text{ADP}], \quad (5.33) \]

the formalism concludes, from the transitivity properties of the Order-of-Magnitude relations:

\[ [\text{GTP}] \gg [\text{ADP}] \quad (5.34) \]

- In an inference similar to one discussed informally in the introduction, if the formalism is given the relation

\[ [\text{ATP}] + [\text{ADP}] + [\text{GTP}] \ll \ll \ll [\text{total P}], \quad (5.35) \]

which states that \([\text{ATP}] + [\text{ADP}] + [\text{GTP}]\) is smaller than the total phosphorous concentration, and the relation

\[ [\text{ATP}] \lessapprox [\text{GTP}], \quad (5.36) \]

where the compound relation \(\lessapprox\) means "comparable to" or "of the same order as," it can conclude the relation
\[
[\text{ATP}] \ll \text{[total P]}, \quad (5.37)
\]
i.e., [ATP] is from much smaller to moderately smaller than [total P].

- Consider the reaction catalyzed by fructose diphosphate aldolase:

\[
\text{FruDP} \rightarrow \text{DHAP} + \text{GAP} \quad (5.38)
\]

where FruDP is fructose diphosphate, DHAP is dihydroxyacetone phosphate, and GAP is glyceraldehyde phosphate. If the reaction is near equilibrium, i.e.,

\[
Q \sim K_e, \quad (5.39)
\]

and the concentrations of fructose diphosphate and dihydroxyacetone phosphate are related by

\[
[\text{FruDP}] \gg [\text{DHAP}], \quad (5.40)
\]

relations (5.39) and (5.40) can be used in the algebraic equation defining the mass action ratio

\[
Q = [\text{DHAP}] \cdot [\text{GAP}] / [\text{FruDP}] \quad (5.41)
\]

to infer the relations

\[
[\text{GAP}] \gg Q, \quad (5.42)
\]

\[
[\text{GAP}] \gg K_e. \quad (5.43)
\]

Note that, for this reaction, Q and Ke have concentration units and can be related to metabolite concentrations. To carry out reasoning with equations, some algebraic manipulation must be performed to form constraints among ratios of quantities, since an O[M] relation implies an interval for the ratio of the parameters involved. In the derivation
of relation (5.42) for example, the constraint (5.41) must be transformed to

\[
([\text{GAP}] / Q) \cdot ([\text{DHAP}] / [\text{FruDP}]) = 1 \quad (5.44)
\]

• An empirical rule can state

\[
\text{if } ([\text{ATP}] \gg [\text{ADP}]) \text{ then } ([\text{GTP}] \gg [\text{GDP}]), \quad (5.45)
\]
i.e., the energy state of the cells, indicated by ATP, is reflected, to the same extent, in the level of GTP as well. Given this rule, whenever the antecedent relation

\[
[\text{ATP}] \gg [\text{ADP}] \quad (5.46)
\]
is known to hold, the consequent relation can be automatically inferred:

\[
[\text{GTP}] \gg [\text{GDP}] \quad (5.47)
\]

5.8.2. INFEERENCE METHODS

We will briefly describe here how the inference tree is expanded and pruned. The basic strategy within O[M] is depth-first data-driven reasoning. In effect, any potential new fact is first checked for redundancy, and then created and used in more inferences. The use of a new fact begins immediately even when the use of its parent antecedent fact has not been completed. Within O[M] all possible scenarios can be invoked for further reasoning:

(a) From the conjunction of relations within the same link, i.e., relations between the same two quantities, new relations are inferred. Redundant relations of the same link are retracted. Hence, if it is already known that
A \langle\ldots\rangle B \quad (5.48) \\

and the new relation

A \langle\ldots\rangle B \quad (5.49) \\

is derived, the two can be combined to yield

A \langle\ldots\rangle B . \quad (5.50) \\

Upon employing this new relation, relations (5.48) and (5.49) become redundant and are retracted.

(b) Employing the symmetry and transitivitiy properties of O[M]-relations, new relations are inferred. The symmetry inference mechanism, allows relation (5.50) to be inverted to produce

B \langle\ldots\rangle A \quad (5.51) \\

If, in addition to relation (24), it is also known that

A \triangleright C . \quad (5.52) \\

then the transitivitiy mechanism can infer

B \langle\ldots\rangle C . \quad (5.53) \\

(c) For relations between a variable and a landmark, numeric transitivitiy is applied. The idea is that if we find another variable related to another landmark compatible to the original one, we can infer a relation between the two (non-landmark) variables. Assuming that the variables A and B represent mass, this mechanism allows the relations

A \ll 10 \text{ kg} \quad (5.54) \\

and
B \geq\cdots\geq 0.5 \text{ kg} \quad (5.55)

to be combined and yield:

A \ll\cdots\ll B \quad (5.56)

This mechanism is possible because of the clear semantics of O[M] relations, and the procedural character of the inference strategy. If our approach were based on rules, it would be particularly cumbersome to tackle these inferences.

(d) When a relation (actually its link) can serve as the antecedent of rules, applicable rules are invoked. If rule (5.31) and the antecedent relations A\ll B and D\ll B are all known, then the consequent A \ll D can be immediately inferred.

(e) When a relation (actually its link) participates in the expression of assignments and constraints, these are invoked as described in Section 5.7. For example, assuming that all variables in assignment (5.27) are positive, the set of relations

\begin{align*}
C &\ll B \\
D &\ll E \\
F &\ll\cdots\ll B
\end{align*} \quad (5.57, 5.58, 5.59)

is used in assignment (5.27) to yield

\begin{equation}
A \geq B. \quad (5.60)
\end{equation}

Applying an assignment can yield knowledge about the magnitude as well as the sign of a variable. Sign specs are thus created, which may depend on the assumption of other sign specs.
The most important feature in constraining the inference search is the physical
dimensionality of quantities. There are always many different kinds of variables present
in applications of any size: temperatures, pressures, volumes, flowrates, masses,
concentrations, etc. The requirement for compatibility to form links or relations cuts
down the number of alternative links or quantities that have to be examined for any
inference scenario. The requirement also reduces the number of possible derivations
from constraints and assignments.

5.8.3. TRUTH-MAINTENANCE AND RESOLUTION OF
CONTRADICTIONS

Assertions can be stated as assumptions rather than known facts. They can also
be stated as dependent on the validity of some other assumptions. At each inference step,
along with the justification of an assertion we form and retain the set of assumptions
under which the assertion is true. The assumption set of an inferred relation is equal to
the union of the assumption sets of all the antecedent relations.

Several different relations between two quantities are allowed to coexist and can
be used in parallel, provided that they carry different assumption sets.

The assumption mechanism is useful, for example, in examining alternative
rate-limiting steps of a biochemical pathway. Candidates can be stated as assumptions;
O[M] finds the results each set of assumptions leads to, and determines which limiting
steps are consistent with other information about the pathway.

The assumption-driven multiple worlds are created and maintained using an
adaptation of de Kleer's Assumption-based Truth-Maintenance System (ATMS)
methodology [de Kleer, 1986, Reiter and de Kleer, 1987]). ATMS avoids some serious
problems of other truth-maintenance systems that use dependency-directed backtracking. In ATMS there is no backtracking involved, and important assumption sets can be parsed after the main problem-solving effort.

ATMS had to be adjusted to the characteristics of our formalism. The consistent use of assumptions dictates that, in the inference mechanisms we describe, the assumption sets of the relations satisfy certain requirements. In an example we used before, involving relations (5.48), (5.49), and (5.50), we described how the conjunction of

\[ A \langle\ldots\rangle \ B \]  
(5.48)

and

\[ A \langle\ldots\rangle \rightarrow B \]  
(5.49)

yields

\[ A \langle\ldots\rangle \rightarrow B \]  
(5.50)

which then renders (5.48) and (5.49) redundant and O[M] can discard them. Let S_{48}, S_{49}, and S_{50} be the assumption sets of relations (5.48), (5.49), and (5.50) respectively. If relation (3.48) is to be discarded as redundant in light of relation (5.50), S_{50} must be a subset of S_{48}. That can only be true if S_{49} is a subset of S_{48}, because S_{50} is the union of S_{48} and S_{49}.

If S_{49} is not a subset of S_{48}, the relation (5.50) will still be inferred, but none of the three relations will be retracted. Note that the validity of an assertion decreases with the size of its assumption set. Thus, if S_{50} is a subset of S_{48} the relation (5.50) is true in more occasions (more "alternative worlds," in the truth-maintenance terminology) than relation (5.48).
With ATMS, whenever a contradiction occurs the assumption set that supported the contradiction is marked as \textit{nogood}. From that point on, no new assertion is made under that assumption set or any of its supersets. Reasoning with subsets of the nogood set is not affected.

The resolution of contradictions however requires particular attention in O[M] because, with the heuristic interpretation, neighboring relations that apparently conflict may actually be both valid heuristically (since neighboring heuristic intervals are overlapping).

For example, the compound relations $\ll<\ldots<\ll$ and $\sim<\ldots>\sim$ are formally disjoint, but the primitives $\ll<\ll$ (part of $\ll<\ldots<\ll$) and $\sim<\sim$ (part of $\sim<\ldots>\sim$) are neighboring and their heuristic intervals overlap. In effect, discovering that

$$A \ll<\ldots<\ll B \quad (5.6.1)$$

$$A \sim<\sim B \quad (5.6.2)$$

for a particular assumption set, does not necessarily mean that the assumption set is contradictory. We will delineate here the alternative ways of handling these apparent contradictions.

- The first way is to avoid any special treatment of neighboring conflicting relations. This would create big problems, since all kinds of assumption sets and eventually the whole problem (i.e., the empty assumption set) would be marked inconsistent, without being truly so.

- The second way is to simply allow neighboring relations to coexist, and mark them in a special way as non-conflicting. Since they will both propagate, this strategy amounts to implicitly
asserting that indeed the overlapping part of the two neighboring intervals represents the "true" relation. This is thus an aggressive strategy.

- The third and most conservative way is to disclaim both relations (and mark them to avoid recurrence of the problem) and replace them by the compound relation representing their disjunction.

If the initial relations are compound one need only consider the two primitive components (one from each initial relation) that are neighboring and take their disjunction, because those are the primitives that are actually caused the overlap in the first place. In the example we discussed above, the disjunction of relations (5.61) and (5.62) would be

\[ A <<..>> B \quad (5.63) \]

but taking only the primitives that cause the problem, we obtain

\[ A -<..<~< B \quad (5.64) \]

and we accept this last tighter relation instead of the original relations. This approach represents a reasonable compromise that corresponds to the way human commonsense resolves conflicts of this type.

Any of the above strategies could be followed within O[M], but the last one is the preferred default.

When the conflict involves a variable and a landmark, one can be yet more aggressive and take advantage of the contradiction to pinpoint the range of the variable
more accurately. We believe, however, that such a strategy is rather risky and does not follow the more conservative general philosophy of our formalism.

5.8.4. GOAL DIRECTION

The search mode for O[M] is opportunistic forward chaining, but there are ways to focus the system and devote special attention to the search for a particular relations. To activate these goal-driven mechanisms, the user must specify that a link between two variables is a target link.

One way to work towards a target is to invoke additional ways to use constraints and assignments. Whenever one of the two target quantities occur, the system introduces the other one as well (for example, it divides both sides of the constraint by that variable). As a result, links between the two goal quantities occur in more places and relations between them are produced and used more frequently.

For example, if the link between quantities A and G is a target, and constraint (5.30) is given, a transformed constraint is derived:

\[(A/G) + 2 (F/G) - (B/G) = 3.5 (C/G) \exp(D/E) \quad (5.65)\]

in which the target link A/G occurs. The constraint in turn yields the assignment

\[(A/G) = (B/G) + 3.5 (C/G) \exp(D/E) - 2 (F/G) \quad (5.66)\]

which is capable of producing relations between A and G.

Another way the system can examine different (alternative) relations between two target quantities, involves the creation of seven assumptions, one for each of the seven primitive relations.
With $A$ and $G$ as the target, for example, the assumptions $A << G$, $A \prec G$, $A \sim G$, $A \equiv G$, $A \succ G$, and $A \gg G$ are created. As the reasoning proceeds some of the assumptions will be disproved by contradictions, and the disjunction of the remaining primitives yields the desired relation between the target quantities.
5.9. DISCUSSION

In the real world, there are always many positive and negative effects on any aggregate result. We believe that any intelligent approach in dealing with them, must concentrate on deciding which of the effects are important and which can be neglected. Only then should a system attempt to determine the sign of the overall result. Even when this final step could not be achieved, at least it would be clear what further information is needed — only information on the dominant effects. The $O[M]$ formalism is aimed exactly at sorting out dominant effects.

There is recently another similar effort in this direction, involving absolute Order-of-Magnitude values (as opposed to binary relations) in the resolution of ambiguities within Qualitative Process Theory [D'Ambrosio, 1987].

Even in quantitative reasoning people use Order-of-Magnitude arguments to reduce algebraic complexity. This is often done systematically: As terms are dropped from equations, a term of the form $O(x)$ is used to for bookkeeping. This $O(x)$ term signifies that the largest term that was neglected is "of the order of $x$". Numerical constants are not introduced in the $O(x)$ term. This type of reasoning resembles the $O[M]$ formalism with the understanding that we keep track of orders $O(e)$, and we additionally distinguish between $O(e)$ and $O(-e)$, but terms of order $O(e^2)$ or higher are neglected.

Using either strict or heuristic interpretation, the acquisition of knowledge from a human expert can be greatly enhanced by the $O[M]$ formalism. Instead of asking the expert to give hard numbers and exact relations, a system can ask the expert questions of the type:
• Is A much larger than B?

• Which of the relations \( A << B, \ A \sim < B, \ A \sim B \) could hold between A and B?

We believe that in many ways the O[M] formalism bridges the gap between traditional qualitative reasoning with signs and full quantitative reasoning with exact numbers. It is suitable for many engineering activities where extensive knowledge is naturally expressible in Order-of-Magnitude relations, especially since it is capable of handling numerical and algebraic knowledge as well.

In conclusion, the O[M] system for reasoning with Orders of Magnitude and approximate relations captures engineering commonsense about parameter sizes, and offers a vocabulary for formalizing concepts that deal with rough parameter magnitudes. O[M] provides for handling diverse forms of knowledge, including constraints, Order-of-Magnitude relations, rules, assumptions, goals, and even exact numerical knowledge.

In Section 5.10 below, we discuss the application of O[M] in the analysis of Process Engineering systems. Applications in Biochemical systems will be the sole topic of the next chapter (Chapter 6).
5.10. APPLICATION OF O[M] IN PROCESS ENGINEERING SYSTEMS AND ACTIVITIES

The O[M] formalism was specifically developed to satisfy needs arising from the semiquantitative nature of knowledge about biochemical systems. However, it appears that many other systems and activities in Process Engineering are characterized by similar semiquantitative knowledge and reasoning.

Thus, it is useful to discuss the scope of the application of the O[M] formalism in Process Engineering at large and provide some examples of Order-of-Magnitude analysis of process systems. The analysis of biochemical systems will be discussed in more detail in Chapter 6.

5.10.1. ENGINEERING ACTIVITIES IN WHICH ORDER-OF-MAGNITUDE ANALYSIS IS RELEVANT

In many design or diagnostic activities in engineering, engineers already use Order-of-Magnitude relations implicitly. In the analysis part of a design task, inferring Order-of-Magnitude relations involving the parameters of process systems (simple units, sections of flowsheets, control systems, reaction pathways, etc.) is a major part of the analysis as it is performed by human designers. The results of this part can be used by other analytical steps operating at a more abstract level. The results can also be used in subsequent synthetic tasks, to select a rigorous synthetic methodology, or apply empirical rules.

Potential applications of O[M] in Process Engineering center around activities characterized by the use of rough models, and the effort to discriminate among alternatives or discern dominant factors. The preliminary design of process flowsheets,
design of control structures, planning of process operations, process trend analysis, fault diagnosis, and analysis of reaction pathways are among the tasks in which Order-of-Magnitude analysis may be useful. We will briefly examine below some of these engineering activities in which O[M] is a methodology capable of formalizing and automating Order-of-Magnitude arguments.

5.10.1.1. Preliminary Design of Process Flowsheets

In the design of process flowsheets it is useful to have shortcut design procedures, which rely on Order-of-Magnitude arguments [Douglas, 1987, Douglas et al., 1985]. Based on absolute and relative orders of magnitude of process parameters a given design procedure is simplified, giving the human designer a better grasp of the design methodology.

Similarly, the determination of the critical design parameters or factors in a preliminary flowsheet is based on Order-of-Magnitude arguments. The search for improvements to the process flowsheet is guided by the results of this analysis.

5.10.1.2. Preliminary Design of Control Structures

In the design of control structures for chemical plants, Order-of-Magnitude arguments have been used to simplify the requisite process models and specify the control objectives [Douglas, 1982], or to select the best pairings of controlled and manipulated variables to form the control loop [Govind and Powers, 1982].
5.10.1.3. Planning of Process Operations

The models required for dealing with the planning of process operations are necessarily dynamic ones, but quantitative dynamic models are hard to obtain and cumbersome to use. In operations like startup or shutdown, the changes involved are so drastic that the assumptions usually involved in the derivation of process models do not hold. It is nevertheless important to represent the process at least semiquantitatively so that rough reasoning about state changes, time scales, and time delays can be carried out.

5.10.1.4. Process Trend Analysis Fault Diagnosis

In methodologies for the diagnosis of process faults, Qualitative Reasoning alone [Oyeleye and Kramer, 1987, Venkatasubramanian and Rich, 1987] may be an inadequate representation. O[M] can provide the semiquantitative knowledge to eliminate spurious solutions; in the simulation of faults and explanation of process behavior [Dalle Molle et al., 1988], O[M] can remove the ambiguity on the consequences of a fault, or the temporal evolution of process behavior. In the analysis of process trends, Order-of-Magnitude Reasoning is needed to supply a vocabulary for the description of relevant process concepts, as well as the potential for computerization of hybrid qualitative-quantitative reasoning.

5.10.2. DETAILED REASONING EXAMPLES

The utility of the O[M] system in practical applications is most pronounced when analysis of complex systems is performed. Since the purpose of this section, however, is simply to demonstrate the applicability of concepts, we will only discuss some simple examples, whose reasoning steps the reader can easily track and verify. These examples involve analysis of a heat exchanger and comparison of simple chemical reactors.
5.10.2.1. A Heat Exchanger

We will show how the O[M] formalism can be used in reasoning about the
countercurrent heat-exchanger shown in Figure 5.6.
Figure 5.6:

A countercurrent heat-exchanger,

with some of its important parameters:

Molar flowrates and molar heat-capacities of the two streams.
5.10.2.1.1. Important Parameters

The important parameters in the analysis of the device are, as shown in Figure 5.6:

- KH, the molar-heat of the hot stream
- FH, the molar-flowrate of the hot stream
- KC, the molar-heat of the cold stream
- FC, the molar-flowrate of the cold stream

Also four important temperature differences can be defined as shown in Figure 5.7 (a, b):

\[
D_{TH} = Th1 - Th2 \quad (5.67)
\]

\[
D_{TC} = Tc1 - Tc2 \quad (5.68)
\]

\[
DT1 = Th1 - Tc1 \quad (5.69)
\]

\[
DT2 = Th2 - Tc2 \quad (5.70)
\]

DTH is the temperature drop of the hot stream, DTC is the temperature rise of the cold stream, DT1 is the driving force at the left end of the device, and DT2 is the driving force at the right end of the device, as shown in Figure 5.7 (a, b).

We chose to define all the temperature differences so that they are positive, to make the example clearer. In fact O[M] would have no difficulty had we defined them in any other way, as long as enough information was provided about individual temperatures to allow O[M] to infer the correct signs for the temperature differences.
Figure 5.7:

(a) Important temperature differences.

(b) A simple sketch of the temperature profiles of the streams along the length of the device.
5.10.2.1.2. Constraints and Relations

The following fundamental constraints must hold among the parameters of the heat-exchanger:

\[ \text{DTH} - \text{DT1-DTC+DT2} = 0 \quad (5.71) \]

\[ \text{DTH} \cdot \text{KH} \cdot \text{FH} = \text{DTC} \cdot \text{KC} \cdot \text{FC} \quad (5.72) \]

Equation (5.71) is a consequence of the definition of the temperature differences, given by Equations (5.67) to (5.70). Equation (5.72) is the energy balance of the device.

For this countercurrent heat-exchanger we may, additionally, know the following Order-of-Magnitude relations:

\[ \text{DT2} \ll \text{DT1} \quad (5.73) \]

\[ \text{DT1} \ll \text{DTH} \quad (5.74) \]

\[ \text{KH} \gg \text{KC} \quad (5.75) \]

Relations (5.73) and (5.74) are typical of high energy-recovery situations, while Relation (5.75) would only hold if the two fluids have similar or even identical composition.

5.10.2.1.3. Inferences

Given the Order-of-Magnitude relations (5.73), (5.74), and (5.75), and the algebraic constraints (5.71), and (5.72), a number of other relations using
the O[M] formalism. We will delineate only the most important reasoning steps that are involved in the application of O[M].

(a) From symmetry and transitivity properties of the relations (5.73) and (5.74), O[M] infers that:

\[
DT2 \ll DTH 
\]  
(5.76)

(b) The algebraic constraint (5.71) is automatically nondimensionalized to:

\[
1 \cdot \frac{DTC}{DTH} \cdot \frac{DT1}{DTH} + \frac{DT2}{DTH} = 0
\]
(5.77)

This dimensionless constraint is then solved for each dimensionless ratio it contains to yield three assignments, one of which is:

\[
\frac{DTC}{DTH} = 1 \cdot \frac{DT1}{DTH} + \frac{DT2}{DTH}
\]
(5.78)

Employing relations (5.73) and (5.76) into this assignment (5.78) O[M] infers that:

\[
DTC \ll\ll DTH 
\]
(5.79)

(c) Again through properties of O[M]-relations, relations for DT1 are inferred:

\[
DT1 \ll DTC 
\]  
(5.80)

\[
DT1 \ll DTH 
\]  
(5.81)

(d) Similarly to constraint (5.71) in part (b) above, constraint (5.72) is nondimensionalized, and then solved to yield:

\[
\frac{FC}{FH} = \frac{KH}{KC} \frac{DTH}{DTC}
\]
(5.82)

Inserted in this assignment, the relations (5.75) and (5.79) yield:
The last relation indicates that under these conditions the flowrates of the two streams are roughly equal. The results for the temperature differences give a better picture of the temperature profiles, shown in Figure 5.8; DT1 and DT2 are small compared to DTC and DTH; DTC is roughly equal to DTH; and the two temperature profiles are almost parallel.

What O[M] could not infer in this example is the fact that not only $DTC \sim FH$, but in fact $DTC \sim DTH$. O[M] is unable to find this result, because it does not perform generalized algebraic manipulation. To infer this result O[M] would have to define a new quantity equal to $DTC - DTH$ and infer its sign through constraint (5.71). Definition of new arbitrary quantities, or other equivalent manipulation, are beyond the scope of O[M].
Figure 5.8:
The temperature profiles of the streams of the heat-exchanger, drawn based on the results of Order-of-Magnitude reasoning.
5.10.2.2. Stirred-Tank and Plug-Flow Chemical Reactors

We will compare here the Order-of-Magnitude behavior of a Continuous Stirred-Tank Reactor (CSTR) and a Plug-Flow Reactor (PFR), for the irreversible first-order reaction:

\[ A \xrightarrow{k} B \] \hspace{1cm} (5.84)

We assume that the reaction is first-order and irreversible. Thus, its rate, \( r \), is given by:

\[ r = k [A] \] \hspace{1cm} (5.85)

where \( k \) is the rate-constant.

5.10.2.2.1. Parameters and Constraints

Let \( F \) be the flowrate through the reactor, \( V \) the reactor volume, \( C_1 \) the concentration of \( A \) in the feed, and \( C_2 \) the concentrations of \( A \) in the effluent of the reactor. If we define the residence time

\[ T = V/F \] \hspace{1cm} (5.86)

and the reaction's time constant

\[ t = 1/k \] \hspace{1cm} (5.87)

for isothermal operation, we can obtain an algebraic constraint involving \( C_1 \) and \( C_2 \) from the mass-balance for the reactor, dependent on the reactor type:

- For a CSTR:
  \[ C_1 \cdot t - C_2 \cdot t - C_2 \cdot T = 0 \] \hspace{1cm} (5.88)

- For a PFR:
\[ \log \left( \frac{C_1}{C_2} \right) = \frac{T}{t} \] (5.89)

Here, \( \log \) denotes the natural logarithm.

5.10.2.2.2. Order-of-Magnitude Analysis

Given the algebraic constraint and the fact that all parameters are positive, an O[M] relation between the residence time and the reaction time constant can yield an O[M] relation between the feed and effluent concentrations. The Order-of-Magnitude analysis is shown in Table 5.4.

The results capture the basic behavior of the system. The Order-of-Magnitude behavior for the CSTR and the PFR is almost the same. The only difference arises when \( T > t \); for a CSTR that yields \( C_2 < C_1 \), while for PFR \( C_2 << \ldots < C_1 \). This is indeed the important difference between the two reactor types; when the residence time is moderately larger than the reaction time a PFR may achieve very high conversion while a CSTR cannot.
Table 5.4:
Order-of-Magnitude Analysis of CSTR and PFR isothermal reactors
for an irreversible, first-order reaction.

<table>
<thead>
<tr>
<th>Time-constants</th>
<th>Concentrations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CSTR</td>
</tr>
<tr>
<td>T &lt;&lt; t</td>
<td>( C_2 \sim C_1 )</td>
</tr>
<tr>
<td>T -&lt;-&gt; t</td>
<td>( C_2 &lt; C_1 )</td>
</tr>
<tr>
<td>T &gt;&gt; t</td>
<td>( C_2 \sim C_1 )</td>
</tr>
<tr>
<td>T &gt;&gt; t</td>
<td>( C_2 \ll C_1 )</td>
</tr>
</tbody>
</table>
5.11. SUMMARY

The O[M] formal system is aimed at formalizing reasoning with approximate relations among quantities — relations like "much smaller than" or "slightly larger than." O[M] is based on seven primitive relations among quantities, and compound relations formed as implicit disjunctions of consecutive primitives.

In the interpretation of the relations, strict interpretation allows exact conservative inferences, while heuristic interpretation allows inferences more aggressive and human-like, by permitting some slack at each inference step. Inference strategies within O[M] are based on propagation of order-of-magnitude relations through properties of the relations, solved or unsolved algebraic constraints, and if-then rules.

Assumption-based truth-maintenance is used, and the physical dimensions of quantities efficiently constrain the inferences. Statement of goals allows more effective employment of the constraints and focuses the formalism's opportunistic forward reasoning.

O[M] is useful in the analysis of Process Engineering systems, and can aid in a variety of Process Engineering activities. As we will see in Chapter 6, O[M] relations permit Order-of-Magnitude analysis of biochemical systems as well.
CHAPTER 6

APPLICATIONS

OF ORDER-OF-MAGNITUDE REASONING

IN THE ANALYSIS

OF BIOCHEMICAL SYSTEMS
6.1. ANALYSIS OF ISOLATED ENZYMATIC REACTIONS

This section focuses on Order-of-Magnitude reasoning in the kinetic analysis of isolated enzymatic reactions. The intent here is to show the validity of Order-of-Magnitude concepts in small applications whose reasoning is not complex and can be followed manually. However, in the analysis of complex networks of enzymatic reactions, the concepts outlined in this section can be applied to each reaction in the network, leading to much more complex reasoning.

In all the examples in this chapter, unless we specifically indicate otherwise, the value of the formalism's accuracy parameter, $e$, will be assumed to be equal to 0.1. This corresponds to a very common interpretation: One order of magnitude denotes a rough factor of 10.

6.1.1. MICHAELIS-MENTEN KINETICS

This example is the simplest possible test of the methodology. The objective is the derivation of the Order-of-Magnitude kinetic behavior of an irreversible biochemical reaction, assuming that the rate is limited by only one substrate, $A$, and follows the Michaelis-Menten relation:

$$ r = \frac{[A]}{K_m + [A]} V_{\text{max}} $$

(6. 1)

This constraint is transformed into another one, involving ratios of compatible quantities:

$$ \frac{r}{V_{\text{max}}} \left( \frac{K_m}{[A]} + 1 \right) = 1 $$

(6. 2)
With the additional knowledge that all parameters are positive, the O[M] formalism yields the Order-of-Magnitude analysis of this constraint capturing the basic features of Michaelis-Menten kinetics (Table 6.1):

- When the concentration of A is much larger than the Michaelis constant, the enzyme is saturated and the rate approaches the maximum.
- When [A] is much smaller than $K_m$, the enzyme is very far from saturation, so the rate is much smaller than the maximum.
- Throughout the intermediate range, where A has concentration comparable to its Michaelis constant (i.e., [A] is from moderately smaller up to moderately larger than $K_m$), the rate is moderately smaller than $V_{max}$. 
Table 6.1:
Order-of-Magnitude analysis of the rate of a biochemical reaction that follows the Michaelis-Menten rate relation.

<table>
<thead>
<tr>
<th>Relation between ([/\cdot]) and (K_m)</th>
<th>Relation between (r) and (V_{\text{max}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>([A]\ll K_m)</td>
<td>(r\ll V_{\text{max}})</td>
</tr>
<tr>
<td>([A]\approx K_m)</td>
<td>(r\approx V_{\text{max}})</td>
</tr>
<tr>
<td>([A]\gg K_m)</td>
<td>(r\ll V_{\text{max}})</td>
</tr>
</tbody>
</table>
6.1.2. INHIBITION OF AN ENZYMATIC REACTION

The above treatment of biochemical reaction rates can be easily extended to cover inhibition of the enzyme by an inhibitor B. Let $K_I$ be the inhibition constant for the effect of B on the reaction rate. Depending on the nature of the competition between the substrate and the inhibitor, one can distinguish three kinds of inhibition [Rawn, 1983], namely competitive, uncompetitive, and non-competitive, each characterized by a different rate expression.

6.1.2.1. Competitive inhibition

Competitive inhibition will be discussed first. It is described by the equation:

$$r = \frac{V_{\text{max}} K_I [A]}{K_I [A] + K_m (K_I + [B])}$$  \hspace{1cm} (6.3)

This constraint is transformed by O[M] into a form involving ratios:

$$\frac{r}{V_{\text{max}}} \left( 1 + \frac{K_m}{[A]} \left( 1 + \frac{[B]}{K_I} \right) \right) = 1$$  \hspace{1cm} (6.4)

For all possible relations between $[B]$ and $K_I$ and all possible relations between $[A]$ and $K_m$, the resulting relations between $r$ and $V_{\text{max}}$ are determined (Table 6.2).

When $[B]$ is much smaller than $K_I$, the inhibition is not apparent at the Order-of-Magnitude level; therefore, those results reduce to the simple Michaelis-Menten case with no inhibition (Table 6.1).
Table 6.2:

Order-of-Magnitude analysis of the rate of a biochemical reaction with competitive inhibition.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[A] &lt;&lt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
</tr>
<tr>
<td>[A] ~&lt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
</tr>
<tr>
<td>[A] <del>&lt;..</del>&lt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
</tr>
<tr>
<td>[A] ~&lt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
</tr>
<tr>
<td>[A] &gt;&gt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
</tr>
<tr>
<td>[A] &gt;&gt; K_m</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r ~&lt; V_max</td>
<td>r <del>&lt;..</del>&lt; V_max</td>
</tr>
</tbody>
</table>
6.1.2.1.1. Refinement through Additional Knowledge

When the concentrations of both A and B become very high (i.e., \([A] \gg K_m\) and \([B] \gg K_I\)) the result of the analysis is \(r \ll V_{max}\), equivalent to \(r \ll V_{max}\). O[M] cannot produce a tighter relation, because there are two factors acting strongly in opposite directions: acceleration of the reaction by high \([A]\), and deceleration by high \([B]\). There is no knowledge about the relative magnitudes of these factors, so any of the two could become dominant. To resolve this ambiguity, additional knowledge can be introduced, relating the two factors. For example, if:

\[ K_m \gg K_I \]  \hspace{1cm} (6.5)

and

\[ [A] \gg [B] \]  \hspace{1cm} (6.6)

then a less ambiguous result can be concluded:

\[ r \ll V_{max} \]  \hspace{1cm} (6.7)

6.1.2.1.2. Inference of Concentrations from Rates

In the above examples the rate was inferred from the concentrations, but the inverse can also be performed; one can infer concentration relationships from given rate relationships.

For example, starting from knowledge about the rate of the reaction and the concentration of the substrate, results about the concentration of the inhibitor are obtained (Table 6.3). From any given relationship between \([A]\) and \(K_m\) and any relationship between \(r\) and \(V_{max}\), the relationship that must hold between \([B]\) and \(K_I\) is inferred.
In the cases where, instead of a relation, the table entry is "contradiction," the two antecedent relations are shown by $O[M]$ to be inconsistent and no conclusion can be drawn from them. It is, for example, impossible for the rate to be only slightly smaller than $V_{\text{max}}$ when $[A]$ is much smaller than $K_m$ — regardless of the amount of inhibitor present.

In a number of other cases in Table 6.3, the relation inferred is $[B] \ll K_I$ ([B] is from much smaller to much larger than $K_I$), which is the trivial relation that always holds between compatible parameters. This is the case when $[A] \ll K_m$ and $r \ll V_{\text{max}}$, because once the concentration of the substrate is very low, the rate is going to be very low regardless of the presence of the inhibitor; since $O[M]$ has no distinction between "much smaller" and something like "very much smaller," the amount of the inhibitor cannot be inferred. Similarly, when $[A] \gg K_m$ and $r \sim V_{\text{max}}$, since the inhibition is competitive, the effect of the substrate overturned any effect from the inhibitor, and there is no indication of the actual amount of the inhibitor.
Table 6.3:
Order-of-Magnitude analysis of the concentration of a competitive inhibitor of an enzyme, given knowledge on the concentration of the substrate and the rate of the reaction.

<table>
<thead>
<tr>
<th></th>
<th>( r \ll V_{\text{max}} )</th>
<th>( r \ll V_{\text{max}} )</th>
<th>( r \ll V_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>([A] \ll K_m)</td>
<td>([B] \ll..\gg K_l)</td>
<td>contradiction</td>
</tr>
<tr>
<td></td>
<td>([A] \ll K_m)</td>
<td>([B] \ll..\gg K_l)</td>
<td>contradiction</td>
</tr>
<tr>
<td></td>
<td>([A] \ll..\gg K_m)</td>
<td>([B] \gg..\gg K_l)</td>
<td>([B] \ll..\gg K_l)</td>
</tr>
<tr>
<td></td>
<td>([A] \gg K_m)</td>
<td>([B] \gg K_l)</td>
<td>([B] \ll..\gg K_l)</td>
</tr>
<tr>
<td></td>
<td>([A] \gg K_m)</td>
<td>([B] \gg K_l)</td>
<td>([B] \gg..\gg K_l)</td>
</tr>
</tbody>
</table>
6.1.2.2. Uncompetitive Inhibition

Similar Order-of-Magnitude analysis can be performed for the cases of uncompetitive and noncompetitive inhibition of the enzyme. For uncompetitive inhibition, the rate equation

\[
\frac{r}{V_{\text{max}}} = \frac{V_{\text{max}} K_i [A]}{K_m K_i + [A] (K_i + [B])}
\]

(6. 8)

is converted to

\[
\frac{r}{V_{\text{max}}} \left( 1 + \frac{K_m}{[A]} \right) \left( 1 + \frac{[B]}{K_i} \right) = 1
\]

(6. 9)

and yields the results shown in Table 6.4.

6.1.2.3. Noncompetitive Inhibition

For noncompetitive inhibition, the initial constraint

\[
r = \frac{V_{\text{max}} K_i [A]}{(K_i + [B]) (K_m + [A])}
\]

(6. 10)

is transformed to:

\[
\frac{r}{V_{\text{max}}} \left( 1 + \frac{[B]}{K_i} \right) \left( 1 + \frac{K_m}{[A]} \right) = 1
\]

(6. 11)

and yields the results shown in Table 6.5.
Table 6.4:
Order-of-Magnitude analysis of the rate of a biochemical reaction with uncompetitive inhibition.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[A] &lt;&lt; K_m</td>
<td>r &lt;&lt; V_{max}</td>
<td>r &lt;&lt; V_{max}</td>
<td>r &lt;&lt; V_{max}</td>
<td>r &lt;&lt; V_{max}</td>
</tr>
<tr>
<td>[A] ~ K_m</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r <del>-..</del> V_{max}</td>
<td>r <del>-..</del> V_{max}</td>
</tr>
<tr>
<td>[A] <del>-..</del> K_m</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r <del>-..</del> V_{max}</td>
</tr>
<tr>
<td>[A] &gt;&gt; K_m</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
</tr>
<tr>
<td>[A] &gt;&gt; K_m</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
<td>r ~ V_{max}</td>
</tr>
</tbody>
</table>
Table 6.5:

Crude-of-Magnitude analysis of the rate of a biochemical reaction with noncompetitive inhibition.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[A] &lt;&lt; K_m</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
</tr>
<tr>
<td>[A] ~&lt;- K_m</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
</tr>
<tr>
<td>[A] ~&lt;....&lt; K_m</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
</tr>
<tr>
<td>[A] ~-&gt; K_m</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
</tr>
<tr>
<td>[A] ~-&gt; K_m</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
<td>r &lt;&lt;....&lt; V_max</td>
<td>r &lt;&lt; V_max</td>
</tr>
</tbody>
</table>
6.1.3. COMPARISON OF INHIBITION MODES

A comparison of the three kinds of inhibition (from Tables 6.2, 6.4, and 6.5) reveals that in most cases the three types of behavior are actually the same, but some differences do occur:

- When \([A] \gg K_m\) and \([B] \ll K_i\), competitive inhibition yields \(r \ll V_{max}\), while uncompetitive and noncompetitive inhibition yield \(r \ll V_{max}\). This shows that only in competitive inhibition the high concentration of A can outweigh the moderate concentration of B.

- When \([A] \gg K_m\) and \([B] \gg K_i\), uncompetitive inhibition yields the tightest result, \(r \ll V_{max}\), i.e., the two effects combine to yield an intermediate rate. The results for the other two modes are more uncertain, each in a different direction: \(r \ll \cdots \ll V_{max}\) for competitive and \(r \ll \cdots \ll V_{max}\) for noncompetitive inhibition. For competitive inhibition, A has the chance to overpower B, so the result must include the possibility \(r \ll V_{max}\). In noncompetitive inhibition, on the other hand, there is enough uncertainty in the concentration of B (note that \(>\) is a much wider relation than, e.g., \(>\)) to allow the possibility that B dominates and keeps the rate low.

When inferences are made in the reverse direction, deriving knowledge about the concentration of [B] from knowledge about the rate, an ambiguity arises with competitive inhibition, as was described earlier. Let us briefly discuss how the ambiguity is removed for uncompetitive inhibition. When \([A] \gg K_m\) and \(r \ll V_{max}\), the constraint describing uncompetitive inhibition yields
\[ [B] \ll K_i \quad (6.12) \]

as opposed to the most ambiguous trivial relation

\[ [B] \ll \ldots K_i \quad (6.13) \]

that was previously inferred for competitive inhibition (Table 6.3). For uncompetitive inhibition, the high concentration of substrate is not enough to avert the negative effect of the inhibitor. Consequently, if the rate is to be high the concentration of the inhibitor must be low.

6.1.3.1. Inhibition Mode Diagnosis

The differences in behavior among the inhibition modes (as well as the pure Michaelis-Menten kinetics) point to another possibility, which is the diagnosis of the mode of inhibition based on Order-of-Magnitude information. If the inhibition mode is not known, but it is known that \([A] \gg K_m\) and \([B] \gg K_I\), then the inhibition mode could be derived from the rate:

- If \(r \ll V_{\text{max}}\), then the inhibition is noncompetitive.
- If \(r \sim V_{\text{max}}\), then the inhibition is competitive.
- If \(r \sim V_{\text{max}}\), then no conclusion can be drawn about the inhibition mode.

6.1.4. IDENTIFICATION OF MICHAELIS OR INHIBITION CONSTANTS

While, for simplicity, the above discussion was cast as reasoning from concentrations to rates and vice versa, it is very important to remember that the \(O(M)\) relations have no particular directionality. In practice, one may think of the whole
analysis and the resulting relations as providing information about the Michaelis constant $K_m$, the inhibition constant $K_i$, and the maximum enzyme turnover $V_{\text{max}}$. In this context, there are in fact many more scenarios for the analysis and diagnosis of enzyme kinetics and inhibition.

For example, if the inhibition constant is unknown, one may perform a kinetic experiment with roughly known [B] and known relation between [A] and $K_m$. The outcome of the experiment is then a relation between $r$ and $V_{\text{max}}$. From the known relations, one can infer a relation between the unknown $K_i$ and the roughly known [B]. The result may depend on the inhibition mode, but, with more information or more experiments, derivation of the rough magnitude of $K_i$ and diagnosis of the mode of inhibition can be done in parallel.
6.2. ANALYSIS OF BIOCHEMICAL PATHWAYS

In this section, potential contributions of O[M] in the analysis of biochemical pathways are demonstrated through case studies in flux analysis and pathway competition, identification of rate-limiting steps, and testing of regulatory mechanisms of pathways. The examples merely point out potential application areas in which Order-of-Magnitude reasoning is relevant.

6.2.1. COMPETITION OF PATHWAYS

The competition of the reactions of Figure 6.1 determines the relative fluxes of the pathways of glycolysis and serine synthesis. Let $v_1$, $v_2$, and $v_3$ be the rates of the three reactions, and $v_{\text{ser}}$ and $v_{3\text{PG}}$ the net rates of production of serine and 3PG (3-phosphoglycerate), respectively. The rates are expressed in mol/m$^3$/s. All the concentrations and rates refer to total suspension volume rather than cell volume, and the degradation of serine is neglected to simplify the reasoning. It is assumed that no other reactions interfere with the intermediates of the pathways. The following mass balance holds for 3PG:

$$v_{3\text{PG}} + v_1 \cdot v_2 \cdot v_3 = 0 \quad (6.14)$$

The rate $v_3$ is not exactly equal to the production of serine, because it has to provide for the accumulation of the other intermediates of the serine pathway. Thus, it is only slightly larger than the serine production:

$$v_3 \approx v_{\text{ser}} \quad (6.15)$$

Since serine is a building block for biological macromolecules, its production rate must include not only the accumulation of free serine, but also the serine
incorporated in proteins. On the other hand, 3PG is only an intermediate in fueling and biosynthesis, and its production rate is only enough to balance the growth-related dilution effect. Thus, the rate of serine synthesis is much larger than the rate of production of 3PG:

\[ v_{\text{ser}} \gg v_{\text{3PG}} \]  

(6. 16)

For anaerobic growth, fueling metabolism takes place at rates much higher than biosynthesis. Thus, the flux \( v_2 \) through the energy-producing glycolysis (equal to the rate of anaerobic ATP production), is much greater than the rate of serine synthesis:

\[ v_2 \gg v_{\text{ser}} \]  

(6. 17)

Using relations (6. 15), (6. 16), and (6. 17) in the algebraic constraint (6.14) O[M] can correctly infer the relationships that hold among the three fluxes:

\[ v_2 \gg v_3 \]  

(6. 18)

\[ v_2 \ll v_1 \]  

(6. 19)

\[ v_3 \ll v_1 \]  

(6. 20)
Figure 6.1:

Competition between the pathways of glycolysis and serine synthesis. $v_1$, $v_2$, and $v_3$ are the rates (in mol/m$^3$s) of the steps around the branching point, while $v_{3PG}$ and $v_{ser}$ are the rates (in mol/m$^3$s) of accumulation of 3-phosphoglycerate and serine.
6.2.2. ANALYSIS OF FLUXES IN BIOCHEMICAL NETWORKS

This example is an instance of flux analysis in a biochemical network. In fact, the relations (6.15) and (6.17) are not raw knowledge, but rather the result of semiquantitative reasoning that could have been performed by O[M]. In general, in the analysis of large biochemical networks these relations would be inferred automatically from other statements, similar to (6.15) and (6.17). The kinds of relations that can be used in flux analysis are outlined below.

- Mass balances, like the constraint (6.14), can be written for every metabolite in a network.

- The rates of net production of metabolites are related to the metabolic requirements and concentrations of the metabolites. Thus, relations like (6.17) are common knowledge (or can be derived from relations among requirements and concentrations).

- Relations among reaction fluxes, such as (6.18), (6.19), and (6.20), are available, only for well-studied sections of the metabolism.

- Quantitative information on exact values or ranges of particular fluxes can easily be used by O[M], along with the relations.

Given knowledge in the above forms, O[M] derives new relations among fluxes and rates. For a large biochemical network with several dozen parameters, manual commonsense reasoning is cumbersome and it is useful to have a formalization like O[M].
6.2.2. IDENTIFICATION OF RATE-LIMITING STEPS

6.2.2.1. Formulation of the Hypothesis

First, a formalization of the notion of a rate-limiting step must be described. Consider a linear pathway consisting of the reactions $r_1, ..., r_n$, shown in Figure 6.2(a). Let $K_i$ be the equilibrium constant of step $r_i$, and $Q_i$ be the mass-action ratio (i.e., the concentrations of the products divided by the concentrations of the reactants) for step $r_i$. The hypothesis that the step $r_L$ is rate-limiting is formulated below:

- If there were an upstream step far from equilibrium, that step would be rate-limiting instead of $r_L$. Thus, every step upstream from $r_L$ must be near equilibrium:

$$K_i \sim Q_i \quad \text{for } i<L \quad (6.21)$$

- Step $r_L$ itself must be far from equilibrium:

$$K_i \gg Q_i \quad \text{for } i=L \quad (6.22)$$

- Every step downstream from $r_L$ must proceed in the specified direction. Thus, each step must have an equilibrium constant larger than its mass-action ratio:

$$K_i \sim...\gg Q_i \quad \text{for } i>L \quad (6.23)$$

The pathway has $r_L$ as its rate-limiting step if and only if relations (6.21) to (6.23) are true. The underlying assumption is, of course, that only a single rate-limiting step exists in a linear pathway. This is almost always correct, considering that:

- Any attempt to increase the rate by accelerating steps downstream from $r_L$ (through more efficient enzymes or higher
enzyme concentrations) will affect the rate of \( r_L \) only by reducing its mass-action ratio (causing lower concentrations of the products of \( r_L \)). But, since \( r_L \) is far from equilibrium, its rate cannot be manipulated through mass-action effects. Thus, the rate of the overall pathway will not be affected.

- The L-1 near-equilibrium steps upstream from \( r_L \) cannot influence the rate of the pathway either, because they cannot affect the concentrations of the reactants of step L, since the concentration are limited by equilibrium and not by kinetics. In other words, adding higher concentrations of those enzymes will not significantly affect the concentrations and \( r_L \) will proceed with the same as before rate.

In effect, there is usually only one rate-limiting step in a linear pathway. The only case were multiple rate-limiting steps occur is when reactions are neither very close nor very far from equilibrium (e.g., mass-action ratio equal to 0.3 times the equilibrium constant).

To test the hypothesis that step \( r_L \) is rate-limiting, one can state the requisite relations, (6.21) through (6.23), as dependent on an assumption. Then O[M] uses these relations along with other knowledge about the system and checks whether a contradiction arises. Knowledge that can be used here involves mainly relations among concentrations and the effect of concentrations on rates.
(a): An abstract linear pathway converting substrate A to product F, through the enzymatic reactions $r_1, r_2, \ldots, r_n$.

(b): A section of the glycolytic pathway, which is an instance of such a linear pathway.

The section shown converts glucose to 2-phosphoglycerate.
6.2.2.2. Example: Aldolase

To make the example more concrete, consider the section of glycolysis depicted in Figure 6.2(b). The hypothesis that aldolase is the rate-limiting step will be tested. This example also shows how effectively O[M] can handle numerical knowledge.

The assumption that aldolase is a rate-limiting step is expressed as a relation among reactant and product concentrations, with all concentrations in mM:

\[(DHAP \cdot GAP) / FruDP \ll 0.07 \quad (6.24)\]

The next reaction downstream, Triose-P-isomerase, is a very fast enzyme and the reaction it catalyzes will be close to equilibrium:

\[\frac{GAP}{DHAP} \sim 0.05 \quad (6.25)\]

Through semiquantitative inferences based on relations (6.24) and (6.25), O[M] can relate the concentrations of GAP and FruDP, depending on the order of magnitude of the concentration of FruDP:

\[
\begin{align*}
(\text{assuming } FruDP < 1 \text{mM}) & \quad GAP < FruDP \\
(\text{assuming } FruDP -<..>- 1 \text{mM}) & \quad GAP \ll FruDP
\end{align*}
\quad (6.26)\]

Propagating the consequences of relation (6.24) further downstream in the pathway, O[M] eventually reaches the relation

\[DPG \ll 0.1 \mu M \quad (6.28)\]

which is determined to be contradictory, because a rule states that all metabolite concentrations are greater than 0.1 \(\mu M\). The relation (6.28) is based on the
assumption (6.24), which can now be rejected. Thus, the hypothesis that aldolase is a rate-limiting step of the pathway is disproved.

6.2.2.2. Example: Phosphoglycerate Dehydrogenase

We provide here a second example of identification of rate-limiting steps. We examine the short pathway for the synthesis of serine, from 3-phosphoglycerate through 3-phospho-hydroxy-pyruvate and 3-phosphoserine (Figure 6.3).

The hypothesis we are interested in, is whether the first step of the pathway, Phosphoglycerate dehydrogenase, is rate-limiting. There is actually good reason to believe it is, because it is the first step that is unique to the serine pathway; any steps upstream from it belong to glycolysis (or other pathways).

A sketch of the reasoning performed is shown in Figure 6.4. The initial relations include the hypothesis that Phosphoglycerate dehydrogenase is rate-limiting as well as additional knowledge on the equilibrium constants of the three steps and the concentrations of the intermediates of the pathway. The formalism concludes that one of the two relations on the right must hold. Since both of the relations are in conflict with the initial knowledge, the hypothesis can be rejected.
Figure 6.3:
The pathway for the synthesis of serine from 3-phosphoglycerate, which is an intermediate of glycolysis.
Figure 6.4:

Reasoning that leads to the rejection of the hypothesis that Phosphoglycerate dehydrogenase is the rate-limiting step of the serine pathway
6.2.3. TESTING HYPOTHESES ON FEEDBACK REGULATORY MECHANISMS OF PATHWAYS

Once a step in a biosynthetic pathway is determined to be rate-limiting, the next question that arises is whether that step is a target for regulation. One would like to know whether the final product of the pathway, or some intermediate downstream, regulates the flux through the pathway by allosterically inhibiting the enzyme that catalyzes the step in question.

Ultimately, the concentration of the enzyme is affected through genetic control. To turn the pathway off, the production of the enzyme is halted, the enzyme is depleted, and there is finally very little enzyme present to catalyze the reaction. To turn the pathway on, large amounts of enzyme are produced and accumulate in the cell; the resulting high enzyme concentration achieves high flux.

The response of this control mechanism is quite slow, and it is preceded by fast allosteric feedback control. While the main purpose of genetic control is to preserve the cell's resources by avoiding production of unnecessary enzymes, the purpose of the feedback control is the survival of the cell, through maintenance of the correct metabolite pools. The present discussion will be limited to the feedback allosteric control mechanism. It follows from the above discussion that, in order to sustain the correct fluxes and concentrations in the cell, this mechanism must be able to control the key enzyme activities fully, independently of any genetic control.

The specific activity of an enzyme can be physicochemically affected by the presence of almost any biochemical compound. The mere existence of an effect signifies nothing more than physicochemical interactions that do not achieve any concerted teleologic control. Thus, if an interaction is to qualify as a control mechanism, it must
be significant, i.e., it must exert substantial control over the flux of the pathway. Empirically, the levels of a metabolite that turn a control mechanism on and off actually differ significantly.

Throughout this discussion, verbal expressions and implicit Order-of-Magnitude concepts were used. All the ideas can be expressed in a rigorous form by using the $O(M)$ formalism. For a linear biosynthetic pathway of the form shown in Figure 6.5, the hypothesis to be formalized is that the final product, $F$, regulates its level by inhibiting a step, $L$, upstream in the pathway. Under constant substrate and enzyme concentrations but varying product concentration, the activity of the enzyme must be controlled by the product, so that:

- There are drastically different product concentration values $C_{on}$ and $C_{off}$, which turn the inhibition fully on and fully off, respectively. The restriction that they are drastically different means:

$$C_{on} \gg C_{off} \quad (6.29)$$

- The rates imposed on the controlled step, $R_{on}$, corresponding to $C_{on}$, and $R_{off}$, corresponding to $C_{off}$, must satisfy the opposite relation:

$$R_{on} \ll R_{off} \quad (6.30)$$

- Since, even when the inhibition is turned off, there is a significant effect exerted by the product on the rate, the hypothetical rate under the absence of an inhibition mechanism, $R_{null}$, and the corresponding product concentration $C_{null}$, must satisfy the relations:

$$C_{null} \ll C_{off} \quad (6.31)$$
In summary, a feedback mechanism that regulates the product level of a biosynthetic pathway must provide rates \( R_{\text{null}}, \ R_{\text{off}}, \ \text{and} \ R_{\text{on}} \), that correspond to product concentrations \( C_{\text{null}}, \ C_{\text{off}}, \ \text{and} \ C_{\text{on}} \), and satisfy the relations (6.29) through (6.32). When testing the feasibility of different regulatory mechanisms, one can state each mechanism as an assumption on which relations (6.29) through (6.32) depend. O[M] aids in the screening of alternative mechanisms, by reasoning with these relations and locating contradictions to known facts or other assumptions.
Figure 6.5:

Regulation of a linear biosynthetic pathway.
The final product of the pathway exerts negative control on an upstream step.
6.3. CONCLUDING REMARKS

Inferring Order-of-Magnitude relations involving the parameters of biochemical reactions, pathways, and microorganisms is a valuable step in the analysis of these systems, and its results can be used by other analytical steps operating at a more abstract level, such as inferring the state of a microorganism or the bottlenecks of a pathway.

The examples of this chapter indicate that the O[M] formal system is applicable in Biochemical Engineering and Biotechnology, because extensive knowledge is naturally expressible in the form of Order-of-Magnitude relations. O[M] is also capable of handling numerical and algebraic knowledge we may have. The formalism is suitable for a number of activities within the analysis of biochemical systems, such as analysis of enzyme inhibition, analysis of fluxes in a biochemical network, identification of rate-limiting steps of a biochemical pathway, and testing of regulatory mechanisms of pathways.

As was discussed in Section 5.10, the formalism is also suitable for the analysis of other systems from Process Engineering at large.
CHAPTER 7

A GROUP CONTRIBUTION METHOD

FOR THE ESTIMATION

OF EQUILIBRIUM CONSTANTS FOR

BIOCHEMICAL REACTIONS
7.1. INTRODUCTION

Values for equilibrium constants for biochemical reactions, especially in less studied pathways, are rarely found in the literature. A reaction's equilibrium constant, $K$, can be obtained from the corresponding Standard Gibbs Free Energy $\Delta G^\circ$ through the relation:

$$K = \exp \left( -\frac{\Delta G^\circ}{RT} \right) \quad (7.1)$$

$\Delta G^\circ$ is equal to the difference between the Standard Gibbs Free Energies of Formation of the products and the Standard Gibbs Free Energies of Formation of the reactants.

For the rest of this chapter, unless otherwise specified, the term "Gibbs Energy" denotes the Standard Gibbs Free Energy of Formation of a molecule, or the Standard Gibbs Free Energy of a reaction, both in aqueous solution.

7.1.1. GROUP-CONTRIBUTION METHODS

Group-contribution methods have been widely used [e.g., Reid et al., 1977, Joback and Reid, 1987] to estimate a variety of thermodynamic properties of chemical compounds.

Such methods are developed by using literature data to estimate the contributions of different functional groups to the property of interest, so that the errors, i.e., the differences between known and predicted values, are minimized.
In the application of the methods, i.e., to estimate the property for a particular compound, one breaks the compound into functional groups and sums up the contributions of the groups (plus a constant residual value).

Group-contribution methods, however, have focused on ideal-gas properties. In particular, no method exists for properties of compounds in aqueous solutions, which are the only conditions of interest in the thermodynamic analysis of *in vivo* biochemical systems.

### 7.1.2. MOTIVATION FOR THE DEVELOPMENT OF THE METHOD

This chapter describes a group contribution method that allows the estimation of Gibbs Energies for compounds in aqueous solution. With this method, one can estimate the equilibrium constant of a biochemical reaction, using only the reaction stoichiometry and the chemical structure of the reactants and the products. The equilibrium constant is essential for the application of other methodologies. Specifically:

- A method developed for the estimation of the maximum rate of any enzymatic reaction requires that the equilibrium constant of the reaction be known. The method will be described in detail in Chapter 8.

- Another method, which achieves the synthesis of biochemical pathways (described in Chapter 9), requires at least approximate knowledge of the equilibrium constant, so that we can distinguish between reversible and irreversible biotransformations (reactions or pathways).
Besides these specific methods, however, it is important to distinguish among infeasible and reactions (as well as reversible and irreversible reactions), by using a (rough) equilibrium constant.
7.2. ESTIMATION OF THE CONTRIBUTIONS OF GROUPS

The estimation of the contribution of each functional group present in biomolecules entails the use of literature data on Gibbs Energies (or, equivalently, on equilibrium constants). With enough data collected, a least-squares fit can be performed to estimate the contributions that minimize the mean square deviation of estimates from actual data.

7.2.1. DATA USED IN THE ESTIMATION

Extensive data from the literature, referring to both Gibbs Energies of Formation of compounds in aqueous solution and Gibbs Energies of reaction for various transformations in aqueous solution, were collected and used in the estimation of the contributions of groups. The data used are shown in Tables 7.5 and 7.6, at the end of this chapter.

In order to compensate for inaccuracies and uncertainties of some data, a subjective weight was assigned to each datum before performing the fit. Higher weight was given to data for molecules (as opposed to reactions), data recorded at the selected standard conditions, and data claiming higher accuracy.

Particularly inaccurate data were excluded from the estimation, especially when more reliable data were available for the same molecule or reaction.

7.2.1.1. Data on Molecules

Data on Gibbs Energies for a variety of biochemical compounds were found in tabulated form, in several literature sources [Thauer et al., 1977, Wagman et al.,
1982, Burton, 1957], referring to aqueous solutions a temperature of 25°C and pH equal to 0 (i.e., \([H^+] = 1M\)).

In addition, data on Gibbs Energies of formation for amino acids, in aqueous solution and at 25°C, were obtained [Sober 1970].

A number of data points existed in many sources (since some of the literature sources have used others). When the values were not identical, due to (normally undocumented) corrections or improvements applied by the later references, we included all the different values as independent data.

The data points are shown in Table 7.5 at the end of this Chapter.

### 7.2.1.2. Data on reactions

Data on the Gibbs Energy of reaction can be used in the estimation because the Gibbs Energy of a reaction is a linear combination of Gibbs Energies of compounds, and can therefore be linearly related to the contributions of those groups that are consumed or produced (overall) by the reaction. A reaction can thus be viewed as a collection of groups, with the coefficients of produced groups being positive, and those of consumed groups being negative.

Gibbs Energies of reaction in aqueous solution at a temperature of 25°C and pH equal to 7, were obtained for a small number of biochemical reactions [Lehninger, 1975, Hinz, 1986]. Equilibrium constants for a large number of biochemical reactions were obtained from Barman's Enzyme Handbook [Barman, 1969 — Volume 1; Barman, 1969 — Volume 2; Barman, 1974 — Supplement 1], under various pH and temperature conditions.
The data points are shown in Table 7.6 at the end of this Chapter.

All the data were converted into Gibbs Energies, in kJ/mol, under selected standard conditions:

- Dilute aqueous solution
- Temperature 25°C
- pH=7

7.2.2. DECOMPOSITION OF COMPOUNDS TO GROUPS

We used the set of groups shown in Tables 7.1 and 7.2. This set is the one recommended by the most recent group-contribution method for ideal-gas properties [Joback and Reid, 1987]. Some modifications were necessary, due to the character of aqueous biochemical systems:

- Groups not occurring in biochemical compounds were omitted, and others were modified into their ionized form common in aqueous solution (e.g., -COO\(^{-}\) instead of -COOH for the carboxyl group).

- Groups on which a large number of data exist, were partitioned into more specific cases (specifically, distinction among primary, secondary and tertiary hydroxyl groups was introduced).

- Certain small molecules (e.g., formate) were left unbroken, and molecules of special biological significance (e.g., ATP, NAD\(^{+}\)) were described by special groups of their own, whose contributions are known from the start. This kind of treatment makes the method more accurate.
7.2.3. ASSUMPTIONS

In performing the fit to obtain the contributions of groups, we assume that the Gibbs Energy of a molecule is given by the linear combination of contributions from each constituent group multiplied by its number of occurrences in that molecule. In addition, there is a fixed residual contribution for each molecule. All of this is standard practice in group contribution methods [Reid et al., 1977, Joback and Reid, 1987].

As with all group-contribution methods, this first-order approximation neglects any possible group interactions. Also, the existence of a molecule in several forms is disregarded, and the form considered predominant is used instead. Since activity coefficients are not used, a variety of other effects are ignored as well. These neglected effects include:

- Ionic strength
- Presence of impurities
- Effect of buffer solutions
- Effect of other solutes in general

While one would like to account for all these factors, the amount of data that would be required and the potential (mathematical as well as physicochemical) complications that would be introduced render such considerations infeasible for the simple methodology pursued here.
7.2.4. RESULTS

Gibbs Energy contributions of functional groups have been estimated by a least-squares fit. Tables 7.1 and 7.2 give the Gibbs Energy results.
<table>
<thead>
<tr>
<th>group</th>
<th>$\Delta G$ (kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>residual</td>
<td>-93.7</td>
</tr>
<tr>
<td>-CH$_3$</td>
<td>31.0</td>
</tr>
<tr>
<td>-NH$_2$</td>
<td>37.3</td>
</tr>
<tr>
<td>primary OH-</td>
<td>-124.4</td>
</tr>
<tr>
<td>secondary OH-</td>
<td>-135.9</td>
</tr>
<tr>
<td>tertiary OH-</td>
<td>-131.6</td>
</tr>
<tr>
<td>-NH$_3^{1+}$</td>
<td>13.7</td>
</tr>
<tr>
<td>-CH=O</td>
<td>-78.1</td>
</tr>
<tr>
<td>-SH</td>
<td>-53.9</td>
</tr>
<tr>
<td>-COO$_1^-$</td>
<td>-302.3</td>
</tr>
<tr>
<td>=CH$_2$</td>
<td>75.1</td>
</tr>
<tr>
<td>-CH$_2^-$</td>
<td>6.8</td>
</tr>
<tr>
<td>-O-</td>
<td>-94.0</td>
</tr>
<tr>
<td>-S-</td>
<td>39.2</td>
</tr>
<tr>
<td>&gt;C=O</td>
<td>-116.1</td>
</tr>
<tr>
<td>&gt;CH-</td>
<td>-17.8</td>
</tr>
<tr>
<td>-CH=</td>
<td>45.0</td>
</tr>
<tr>
<td>&gt;C=</td>
<td>-50.9</td>
</tr>
<tr>
<td>=C=</td>
<td>20.3</td>
</tr>
</tbody>
</table>
Table 7.2:

Contributions of ring groups to the Gibbs Energy

<table>
<thead>
<tr>
<th>group</th>
<th>$\Delta G$(kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-ring  =C&lt;</td>
<td>73.4</td>
</tr>
<tr>
<td>aromatic &gt;CH</td>
<td>34.5</td>
</tr>
<tr>
<td>aromatic &gt;C-</td>
<td>8.8</td>
</tr>
<tr>
<td>ring =C&lt;</td>
<td>28.2</td>
</tr>
<tr>
<td>ring -CH$_2$</td>
<td>24.7</td>
</tr>
<tr>
<td>ring -O-</td>
<td>-103.1</td>
</tr>
<tr>
<td>ring &gt;CH-</td>
<td>-7.7</td>
</tr>
<tr>
<td>ring -CH=</td>
<td>42.9</td>
</tr>
<tr>
<td>ring &gt;C&lt;</td>
<td>-49.8</td>
</tr>
</tbody>
</table>
7.2.4.1. **Accuracy**

For the data used in the estimation the error was in most cases smaller than 5 kJ/mol. The highest errors, of the order of 20 kJ/mol, occurred for very unreliable equilibrium-constant data of a small number of reactions. For some reactions, there is good reason to believe that the equilibrium constant values are inconsistent with the reported definition of the equilibrium constant (the problem is that species like H\(^+\) may be omitted from the definition and included in the value, or vice versa, introducing very high errors).

Data for reactions are particularly unreliable when they are estimated through the Haldane relationships from kinetic parameters, because the latter are harder to obtain experimentally, and more sensitive to nonidealities.

While the mean error between estimates and initial data is 6.6 kJ/mol, the mean square error is 10 kJ/mol, due to a small number of outliers (the reactions discussed above). The errors for data on compounds only are lower: 5 kJ/mol mean error and 7.8 kJ/mol mean square error.

If we only look at data that were given directly in Gibbs-Energy form, the errors are 5.1 kJ/mol (mean) and 7.5 kJ/mol (mean square). This further reinforces our earlier claim that equilibrium constant measurements (or their estimates from kinetic data) are particularly unreliable because of the assumptions involved and the confusion over definitions and standard states.
7.3. EXAMPLES

7.3.1. GIBBS ENERGIES OF FORMATION OF COMPOUNDS

As an example of using method, consider the estimation of the Gibbs Energy of a metabolite. *Crotonate* (CH\(_3\)-CH=CH-COO\(^{-}\)) can be broken down to the groups -CH\(_3\), -CH= (2 occurrences), -COO\(^{-}\).

The contributions of the groups taken from Table 7.1 are added, along with the residual (which must be added to every molecule), as shown in Table 7.3. The result is 274.9 kJ/mol, which is very close to the value of -277.4 kJ/mol given in the literature [Thauer *et al.*, 1977].
Table 7.3:
Calculation of the Gibbs Energy of crotonate
from the contributions of groups.

<table>
<thead>
<tr>
<th>GROUPS</th>
<th>CONTRIBUTIONS in ΔG</th>
</tr>
</thead>
<tbody>
<tr>
<td>residual</td>
<td>-93.7 kJ/mol</td>
</tr>
<tr>
<td>2 x -CH=</td>
<td>2 x +45.0 kJ/mol</td>
</tr>
<tr>
<td>-COO¹⁻</td>
<td>31.0 kJ/mol</td>
</tr>
<tr>
<td>-CH₃</td>
<td>-302.2 kJ/mol</td>
</tr>
</tbody>
</table>

-274.9 kJ/mol
7.3.2. EQUILIBRIUM CONSTANTS OF REACTIONS

To calculate the Gibbs Energy of a reaction, one should consider the net production and consumption of groups by the reaction. Consider the reaction catalyzed by fructose diphosphate aldolase, which converts 1 mol of fructose diphosphate (FDP) to 1 mol of glyceraldehyde phosphate (GAP) and 1 mol of dihydroxyacetone phosphate (DHAP). FDP will be assumed to have its ring configuration, which is generally considered predominant.

The calculation of the net groups occurring in the reaction and the Gibbs Energy is shown in Table 7.4. The net occurrences of each group (including the residual) in the reaction is derived by summing the occurrences in GAP and DHAP and subtracting that of FDP, multiplied by appropriate stoichiometric coefficients; in this case, the stoichiometric coefficients of the metabolites are all unity.

The result is 20.3 kJ/mol (which is equivalent to an equilibrium constant of $2.8 \times 10^{-4}$ M). The result compares well to the two values provided in the literature:

- [Lehninger, 1975] gives a Gibbs Energy of 24.0 kJ/mol
- [Barman, 1969] gives an equilibrium constant which, converted to Gibbs Energy, amounts to +23.7 kJ/mol.

The observed error of approximately 3.5 kJ/mol is equivalent to a rough factor of 4 in the estimation of the equilibrium constant.
Table 7.4:
Calculation of the Gibbs Energy of *fructose diphosphate aldolase*
from the contributions of groups

<table>
<thead>
<tr>
<th>Occurrences of Group</th>
<th>Group</th>
<th>Contribution</th>
<th>FDP</th>
<th>GAP</th>
<th>DHAP</th>
<th>net</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>residual</td>
<td>-93.7 kJ/mol</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-93.7</td>
<td></td>
</tr>
<tr>
<td>-OPO$_3^-$</td>
<td>not needed</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>ring &gt;CH-</td>
<td>-7.7 kJ/mol</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>+23.1</td>
<td></td>
</tr>
<tr>
<td>ring &gt;C&lt;</td>
<td>-49.8 kJ/mol</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>+49.8</td>
<td></td>
</tr>
<tr>
<td>ring -O-</td>
<td>-103.1 kJ/mol</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>+103.1</td>
<td></td>
</tr>
<tr>
<td>prim. OH-</td>
<td>-124.4 kJ/mol</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-124.4</td>
<td></td>
</tr>
<tr>
<td>sec. OH-</td>
<td>-135.9 kJ/mol</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>+135.9</td>
<td></td>
</tr>
<tr>
<td>tert. OH-</td>
<td>-131.6 kJ/mol</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>+131.6</td>
<td></td>
</tr>
<tr>
<td>-CH=O</td>
<td>-78.1 kJ/mol</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-78.1</td>
<td></td>
</tr>
<tr>
<td>&gt;C=O</td>
<td>-116.1 kJ/mol</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-116.1</td>
<td></td>
</tr>
<tr>
<td>&gt;CH-</td>
<td>-17.8 kJ/mol</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-17.8</td>
<td></td>
</tr>
<tr>
<td>-CH$_2$-</td>
<td>+6.9 kJ/mol</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>+6.9</td>
<td></td>
</tr>
</tbody>
</table>

+20.3 kJ/mol
7.4. SUMMARY AND SIGNIFICANCE

Using Gibbs Energies of compounds, as well as Gibbs Energy changes and
 equilibrium constants of biochemical reactions, the contributions of functional groups to
 the Gibbs Energy (in aqueous solution, temperature 25°C, and pH=7) have been
 estimated. These contributions allow the estimation of the Gibbs Free Energy and the
 equilibrium constant of a biochemical reaction, given only the structures of the
 reactants and products of the reaction.

The value of this method for the estimation of the equilibrium constant of a
 biotransformation lies in the fact that the information it requires consists solely of the
 net stoichiometry of the transformation and the molecular structures of the reactants
 and products. This information can be easily obtained for the vast majority of
 biochemical reactions.

In light of the sparseness of experimental data, the approximate a priori values
 the method provides are extremely valuable. The methodology is, like all
 group-contribution methods, only a first order approximation, but its accuracy is
 sufficient for the preliminary-design purposes of this work.

We will use the numerical estimates of equilibrium constants in the estimation of
 the maximum rates for enzymatic reactions (Chapter 8). In the synthesis of
 biochemical pathways (Chapter 9) we will use only the rough magnitude of the
 equilibrium constant to discriminate among irreversible, potentially reversible, and
 infeasible reactions.
Table 7.5:

Data points on the Gibbs Energies of Formation of compounds
in aqueous solution, used in the estimation of the contributions
of functional groups.

The keywords used for the references are:

"Burton" for [Burton, 1957]
"Lehninger" for [Lehninger, 1975]
"Hinz" for [Hinz, 1986]
"Wagman" for [Wagman et al., 1982]
"Thauer" for [Thauer et al., 1977]

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>XANTHINE</td>
<td>-33.3</td>
<td>kcal/mol</td>
<td>Burton</td>
</tr>
<tr>
<td>BETA-MALTOSE</td>
<td>-1497.04</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ALPHA-LACTOSE</td>
<td>-1515.24</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-TYROSINE</td>
<td>-370.7</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-RIbose-5-Phosphate</td>
<td>-757.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-Phenylalanine</td>
<td>-207.1</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Sorbose</td>
<td>-915.38</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>D-Glucose-Ring</td>
<td>-917.22</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Mannitol-1-Phosphate</td>
<td>-942.61</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-Cystine</td>
<td>-666.93</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>D-Heptose</td>
<td>-1077.</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Sorbitol</td>
<td>-942.7</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Mannitol</td>
<td>-942.61</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Iditol</td>
<td>-942.61</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Creatine-Phosphate</td>
<td>-204.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>D-Gluconate-</td>
<td>-1128.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-Leucine</td>
<td>-343.1</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-Isoleucine</td>
<td>-343.9</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Cis-Aconitate</td>
<td>-922.61</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-Aspartate</td>
<td>-700.4</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>Valerate-</td>
<td>-698.69</td>
<td>kJ/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>continued</strong></td>
</tr>
<tr>
<td>Name</td>
<td>Value</td>
<td>Units</td>
<td>Reference</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>ALPHA-KETOGLUTARATE2-</td>
<td>-797.55</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-CYSTEINE</td>
<td>-339.78</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>D-ERYTHROSE</td>
<td>-598.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-SERINE</td>
<td>-510.87</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-MALATE2-</td>
<td>-845.08</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>BETA-HYDROXYBUTIRATE-</td>
<td>-506.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-ALANINE</td>
<td>-371.54</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-371.3</td>
<td>kJ/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>N-BUTANOL</td>
<td>-171.84</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>TRI-METHYL-AMINE</td>
<td>37.2</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>37.4</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>SUCCINATE2-</td>
<td>-690.23</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>GLYCEROL</td>
<td>-477.06</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>BUTYRATE-</td>
<td>-352.63</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>OXALOACETATE2-</td>
<td>-797.18</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>GLYCERATE-</td>
<td>-658.1</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>BETA-KETOButYRATE-</td>
<td>-493.7</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>FUMARATE2-</td>
<td>-604.21</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-GLYCERALDEHYDE</td>
<td>-437.65</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>DI-HYDROXYACETONE</td>
<td>-445.18</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>LACTATE-</td>
<td>-517.81</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>BETA-HYDROXYPROPIONATE-</td>
<td>-518.4</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>CROTONATE-</td>
<td>-277.4</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>L-GLYCINE</td>
<td>-370.788</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-370.647</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>1-PROPANOL</td>
<td>-175.81</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ISO-PROPANOL</td>
<td>-185.94</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>DI-METHYL-AMINE</td>
<td>-3.3</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-3.22</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>PROPIONATE-</td>
<td>-361.08</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>KETOPYRUVATE</td>
<td>-474.63</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ETHYLENE-GLYCOL</td>
<td>-330.5</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>DIMETHYL-KETONE</td>
<td>-161.17</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>GLYCOLLATE-</td>
<td>-530.95</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ACRYLATE-</td>
<td>-286.19</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value.</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETHYL-ALCOHOL</td>
<td>-181.75</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-181.64</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td></td>
<td>-181.6</td>
<td>kJ/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>METHYL-AMINE</td>
<td>-40.0</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-39.86</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>OXALATE2-</td>
<td>-674.04</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-673.9</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>UREA</td>
<td>-203.76</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>CH3-COO@-1</td>
<td>-369.41</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-369.31</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td></td>
<td>-372.3</td>
<td>kJ/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>ETHANE</td>
<td>-17.01</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>GLYOXALATE-</td>
<td>-468.6</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ACETALDEHYDE</td>
<td>-139.9</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>H2CO3</td>
<td>-623.16</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>HCO3@-1</td>
<td>-586.85</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-586.77</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td></td>
<td>-587.14</td>
<td>kJ/mol</td>
<td>Lehninger</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO$_3^@$-2</td>
<td>-527.9</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-527.81</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>HYDRAZINE</td>
<td>128.03</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td>ETHYLENE</td>
<td>81.36</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>AMMONIUM</td>
<td>-79.37</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-79.31</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>FORMATE-</td>
<td>-351.04</td>
<td>kJ/mol</td>
<td>Thauer</td>
</tr>
<tr>
<td></td>
<td>-351.0</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
<tr>
<td>METHANE</td>
<td>-34.33</td>
<td>kJ/mol</td>
<td>Wagman</td>
</tr>
</tbody>
</table>

Table 7.5 completed
Table 7.6:
Data points on reactions used in the estimation of the contributions of functional groups.
The units of each data point indicate whether it is:
a Gibbs Energy
(with possible units of kJ/mol, J/mol, kcal/mol, or cal/mol) or
an Equilibrium Constant
(with possible units of M, M\(^{-1}\), M\(^2\), M\(^{-2}\), etc.,
or U, which indicates a dimensionless equilibrium constant).
The keywords used for the references are:
"Barman" for [Barman, 1969 - Vol. 1&2, or Barman, 1974 - Supplement 1]
"Lehninger" for [Lehninger, 1975]
"Hinz" for [Hinz, 1986]

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>OXALOACETATE-DECARBOXYLASE</td>
<td>0.00218</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-MANNOSE-6-PHOSPHATEKETOISOMERASE</td>
<td>1.78</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>continued</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Value</td>
<td>Units</td>
<td>Reference</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
</tr>
<tr>
<td>D-MANNOSE-KETOL-ISOMERASE</td>
<td>2.45</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-RIBOSE-5-PHOSPHATE-KETOL-ISOMERASE</td>
<td>0.3</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-XYLOSE-KETOL-ISOMERASE</td>
<td>0.16</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ARABINOSE-KETOL-ISOMERASE</td>
<td>0.11</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-ARABINOSE-KETOL-ISOMERASE</td>
<td>0.179</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ARABINOSE-ISOMERASE[FUCOSE]</td>
<td>0.123</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GLYCERALDEHYDE-3-PHOSPHATE-KETOL-ISOMERASE</td>
<td>0.045</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ADENYLATE-CYCLASE</td>
<td>0.065</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>ADENYLOSUCCINATE-AMP-LYASE</td>
<td>0.0068</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ARGININOSUCCINATE-ARGININE-LYASE</td>
<td>0.0114</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ASPARTATE-AMMONIA-LYASE</td>
<td>0.023</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>CROTONOYL-ACP-HYDRATASE</td>
<td>19.</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>26.8</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-3-HYDROXYACYL-COA-HYDROLYASE</td>
<td>16.2</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-PHOSPHO-D-GLYCERATE-HYDROLYASE</td>
<td>6.3</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>5-DEHYDROQUINATE-DEHYDRATASE</td>
<td>15.</td>
<td>U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value.</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HISTIDINE-AMINOLYASE</td>
<td>3.</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-MALATE-DYDRO-LYASE</td>
<td>4.348</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-[+]CITRAMALATE-PYRUVATE-LYASE</td>
<td>8.3</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-OXO-4-HYDROXYGLUTARATE-GLYOXYLATE-LYASE</td>
<td>10.8</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-CITRATE-OXALOACETATE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[COA-ACETYLATING-AND-ATP-DEPHOSPHORYLATING]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CITRATE-OXALOACETATE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CITRATE-OXALOACETATE-LYASE[COA-ACETYLATING]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-MALATE-GLYOXYLATE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-RHAMNULOSE-1-PHOSPHATE-L-LACTALDEHYDE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L-FUCULOSE-1-PHOSPHATE-L-LACTALDEHYDE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-PHOSPHO-2-KETO-3-DEOXY-D-GLUCONATE-D-GLYCERALDEHYDE-3-PHOSPHATE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FRUCTOSE-1-6-DIPHOSPHATE-D-GLYCERALDEHYDE-3-PHOSPHATE-LYASE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.6, continued

<table>
<thead>
<tr>
<th>Name</th>
<th>Value.</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>KETOSE-1-PHOSPHATE-ALDEHYDE-LYASE</td>
<td>1.7e-6</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ALLOTHREONINE-ACETALDEHYDE-LYASE</td>
<td>0.0178</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-DEOXY-D-RIBOSE-5-PHOSPHATE-ACETALDEHYDE-LYASE</td>
<td>2.0e-4</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>ERYTHRULOSE-1-PHOSPHATE-FORMALDEHYDE-LYASE</td>
<td>0.33</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>PEP-CARBOXYLASE[ITP]</td>
<td>0.373</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>5-10-METHENYL-TETRAHYDROFOLATE-5-HYDROLASE[DECYCLIZING]</td>
<td>2.4e-8</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>4,5-DIHYDROPYRIDINE-AMIDOHYDROLASE</td>
<td>0.67</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-GLUTAMINE-AMIDOHYDROLASE</td>
<td>320.</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>N-PTEROYL-L-GLUTAMATE-L-GLUTAMATE-HYDROLASE</td>
<td>0.064</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>S-ADENOSYL-L-HOMOCYSTEINE-HYDROLASE</td>
<td>1.4e-6</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>SUCCINYL-COA-3-OXOACID-COA-TRANSFERASE</td>
<td>0.0049</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>0.067</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-PYRUVATE-ORTHOPHOSPHATE-PHOSPHOTRANSFERASE</td>
<td>1.0e-17</td>
<td>M2</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTP-1-PHOSPHO-α-D-GLUCURONATE-URIDYLTRANSFERASE</td>
<td>0.34</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>CTP-α-D-GLUCOSE-1-PHOSPHATE-CYTIDYLTRANSFERASE</td>
<td>0.57</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>1.</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>GTP-α-D-HEXOSE-1-PHOSPHATE-GUANYLTRANSFERASE</td>
<td>0.25</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>NUCLEOSIDE-TRIPHOSPHATE-HEXOSE-1-PHOSPHATE-NUCLEOTIDYLTRANSFERASE</td>
<td>0.25</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>GTP-α-D-MANNOSE-1-PHOSPHATE-GUANYLTRANSFERASE</td>
<td>2.5</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>UTP-α-D-GLUCOSE-1-PHOSPHATE-URIDYLTRANSFERASE</td>
<td>0.31</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>2-3-DIPHOSPHO-D-GLYCERATE-2-PHOSPHO-D-GLYCERATE-PHOSPHOTRANSFERASE</td>
<td>5.0</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>α-D-GLUCOSE-1-6-DIPHOSPHATE-α-D-GLUCOSE-1-PHOSPHATE-PHOSPHOTRANSFERASE</td>
<td>17.2</td>
<td>U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP-DEOXYNUCLEOSIDE-MONOPHOSPHATE-PHOSPHOTRANSFERASE</td>
<td>1.4</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>NUCLEOSIDETRIPHOSPHATE-AMP-PHOSPHOTRANSFERASE</td>
<td>0.82</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-NUCLEOSIDEDIPHOSPHATE-PHOSPHOTRANSFERASE</td>
<td>0.91</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP+CMP=ADP+CDP</td>
<td>1.0</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-D-CMP-PHOSPHOTRANSFERASE</td>
<td>1.49</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-AMP-PHOSPHOTRANSFERASE</td>
<td>2.26</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-5-PHOSPHOMEVALONATE-PHOSPHOTRANSFERASE</td>
<td>1.</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-L-ARGININE-PHOSPHOTRANSFERASE</td>
<td>0.167</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>0.002</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-CREATINE-PHOSPHOTRANSFERASE</td>
<td>7.2e-9</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>2.98e-9</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-L-ASPARTATE-4-PHOSPHOTRANSFERASE</td>
<td>3.5e-4</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-3-PHOSPHO-D-GLYCERATE-1-PHOSPHOTRANSFERASE</td>
<td>3450.0</td>
<td>U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value. Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARBAMATE-KINASE[CARBAMATE]</td>
<td>0.04 U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-ACETATE-PHOSPHOTRANSFERASE</td>
<td>0.008 U</td>
<td>Barman</td>
</tr>
<tr>
<td>PYROPHOSPHATE-L-SERINE-O-PHOSPHOTRANSFERASE</td>
<td></td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>950. U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-PYRUVATE-PHOSPHOTRANSFERASE</td>
<td>6450.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-D-GALACTOSE-1-PHOSPHOTRANSFERASE</td>
<td>26. U</td>
<td>Barman</td>
</tr>
<tr>
<td>ATP-D-HEXOSE-6-PHOSPHOTRANSFERASE</td>
<td>386.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>155.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td>O-PHOSPHO-L-SERINE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td></td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>8.25 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-2-AMINOADIPATE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td></td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>1.32 U</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCINE-OXALOACETATE-AMINOTRANSFERASE</td>
<td>0.0164 U</td>
<td>Barman</td>
</tr>
<tr>
<td>4-AMINOBUTYRATE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td></td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>0.1 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ALANINE-MALONATE-SEMIALDEHYDE-AMINOTRANSFERASE</td>
<td></td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>0.2 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ORNITHINE-2-OXOACID-AMINOTRANSFERASE</td>
<td>71. U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value. Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-HISTIDINOLPHOSPHATE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td>2500.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ALANINE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td>2.2 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ASPARTATE-2-OXOGLUTARATE-AMINOTRANSFERASE</td>
<td>0.165 U</td>
<td>Barman</td>
</tr>
<tr>
<td>GUANOSINE-ORTHOPHOSPHATE-RIBOSYLTRANSFERASE</td>
<td>0.019 U</td>
<td>Barman</td>
</tr>
<tr>
<td>NUCLEOSIDE-PURINE[PYRIMIDINE]DEOXYRIBOSYLTRANSFERASE</td>
<td>15.1 U</td>
<td>Barman</td>
</tr>
<tr>
<td>CELLOBIOSE-ORTHOPHOSPHATE-GLUCOSYLTRANSFERASE</td>
<td>0.23 U</td>
<td>Barman</td>
</tr>
<tr>
<td>UDP-GLUCOSE-FRUCTOSE-6P-GLUCOSYL-TRANSFERASE</td>
<td>3250. U</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>53. U</td>
<td>Barman</td>
</tr>
<tr>
<td>UDP-GLUCOSE-D-FRUCTOSE-2-GLUCOSYL-TRANSFERASE</td>
<td>1.6 U</td>
<td>Barman</td>
</tr>
<tr>
<td>MALTOSE-ORTHOPHOSPHATE-GLUCOSYLTRANSFERASE</td>
<td>0.23 U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value, Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACETYL-COA-ACETYL-COA-ACETYLTRANSFERASE</td>
<td>2.0e-5 U</td>
<td>Barman</td>
</tr>
<tr>
<td>ACETYL-COA-ORTHOPHOSPHATE-ACETYLTRANSFERASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0135 U</td>
<td>Barman</td>
</tr>
<tr>
<td>ACETYL-COA-CARNITINE-O-ACETYLTRANSFERASE</td>
<td>1.67 U</td>
<td>Barman</td>
</tr>
<tr>
<td>ACETYL-COA-CHOLINE-O-ACETYLTRANSFERASE</td>
<td>5100.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td>SEDOHEPTULOSE-7-PHOSPHATE-D-GLYCERALDEHYDE-3-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-PHOSPHATE-DIHYROXYACETONE-TRANSFERASE</td>
<td>1.05 U</td>
<td>Barman</td>
</tr>
<tr>
<td>TRANSKETOLASE[ERYTHROSE]</td>
<td>66.7 U</td>
<td>Barman</td>
</tr>
<tr>
<td>TRANSKETOLASE[FRUCTOSE]</td>
<td>0.84 U</td>
<td>Barman</td>
</tr>
<tr>
<td>SEDOHEPTULOSE-7-PHOSPHATE-D-GLYCERALDEHYDE-3-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-PHOSPHATE-GLYCOL-ALDEHYDETRANSFERASE</td>
<td>0.95 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ARGININE-GLYCINE-AMIDINOTRANSFERASE</td>
<td>1.07 U</td>
<td>Barman</td>
</tr>
<tr>
<td>CARBAMOYLPHOSPHATE-OXAMATE-CARBAMOYLTRANSFERASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.623 U</td>
<td>Barman</td>
</tr>
<tr>
<td>CARBAMOYLPHOSPHATE-L-ORNITHINE-CARBAMOYLTRANSFERASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100000.0 U</td>
<td>Barman</td>
</tr>
<tr>
<td>METHYLMALONYL-COA-PYRUVATE-CARBOXYLTRANSFERASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.526 U</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-FORMYL-L-GLUTAMATE-TETRAHYDROFOLATE-5-FORMYLTRANSFERASE</td>
<td>13.0</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>N-FORMIMINO-L-GLUTAMATE-TETRAHYDROFOLATE-5-FORMIMINOTRANSFERASE</td>
<td>1.5</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>N-FORMIMINOGLYCINE-TETRAHYDROFOLATE-5-FORMIMINOTRANSFERASE</td>
<td>0.32</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-SERINE-TETRAHYDROFOLATE-5-10-HYDROXYMETHYLTRANSFERASE</td>
<td>10.2</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>REDUCED-NADP-NAD-OXIDOREDUCTASE</td>
<td>1.43</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>5-10-METHYLENETETRAHYDROFOLATE-NADP-OXIDOREDUCTASE</td>
<td>7.55</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>5-6-7-8-TETRAHYDROFOLATE-NADP-OXIDOREDUCTASE</td>
<td>1.79e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-4-DIAMINOPENTANOATE-NAD[P]-OXIDOREDUCTASE[DEAMINATING]</td>
<td>1.04e-14</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCINE-NAD-OXIDOREDUCTASE</td>
<td>2.72e-11</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-LEUCINE-DEHYDROGENASE</td>
<td>1.11e-13</td>
<td>M</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
Table 7.6, continued

<table>
<thead>
<tr>
<th>Name</th>
<th>Value. Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-GLUTAMATE-NADP-OXIDOREDUCTASE[DEAMINATING]</td>
<td>4.5e-14 M²</td>
<td>Barman</td>
</tr>
<tr>
<td>L-GLUTAMATE-NAD-OXIDOREDUCTASE[DEAMINATING]</td>
<td>4.5e-14 M²</td>
<td>Barman</td>
</tr>
<tr>
<td>L-ALANINE-NAD-OXIDOREDUCTASE[DEAMINATING]</td>
<td>6.98e-14 M</td>
<td>Barman</td>
</tr>
<tr>
<td>BUTYRYL-COA-DEHYDROGENASE[PYOCYAMINE]</td>
<td>0.22 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-4,5-DIHYDRO-OROTATE-OXYGEN-OXIDOREDUCTASE</td>
<td>4.34e-10 M</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYOXYLATE-NADP-OXIDOREDUCTASE[ACYLATING-COA]</td>
<td>7.25e-6 U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GLYCERALDEHYDE-3-PHOSPHATE-NAD-OXIDOREDUCTASE[PHOSPHORYLATING]</td>
<td>0.5 M-1</td>
<td>Barman</td>
</tr>
<tr>
<td>ALDEHYDE-NAD-OXIDOREDUCTASE[ACYLATING-COA]</td>
<td>1.2e-4 U</td>
<td>Barman</td>
</tr>
<tr>
<td>L-LACTATE-OXALOACETATE-OXIDOREDUCTASE</td>
<td>1.8 U</td>
<td>Barman</td>
</tr>
<tr>
<td>MANNITOL-DEHYDROGENASE[NADP]</td>
<td>5.0e-10 M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-THREONATE-NAD-OXIDOREDUCTASE</td>
<td>3.7e-11 M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-DEOXY-D-GLUCONATE-NAD-OXIDOREDUCTASE</td>
<td>3.4e-12 M</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-ALDOHEXOSE-NADP-1-OXIDOREDUCTASE</td>
<td>1.0e-7</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>INDOLELACTATE-NAD-OXIDOREDUCTASE</td>
<td>6.3e-6</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-CARNITINE-NAD-OXIDOREDUCTASE</td>
<td>1.3e-11</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>D-3-PHOSPHOGLYCERATE-NAD-OXIDOREDUCTASE</td>
<td>1.0e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCOLLATE-NADP-OXIDOREDUCTASE</td>
<td>3.0e-18</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCEROL-DEHYDROGENASE[NADP]</td>
<td>1.67e-6</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GLUCONATE-NAD[P]-OXIDOREDUCTASE</td>
<td>3.5e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>MANNITOL-NAD-OXIDOREDUCTASE</td>
<td>3.6e-9</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>5.3e-9</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>4-HYDROXYBUTYRATE-NAD-OXIDOREDUCTASE</td>
<td>2.6e-7</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>TARTRONATE-SEMIALDEHYDE-REDUCTASE</td>
<td>5.1e-14</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>3-HYDROXYPROPIONATE-NAD-OXIDOREDUCTASE</td>
<td>9.0e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>RIBITOL-NAD-OXIDOREDUCTASE</td>
<td>4.5e-11</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>0.00717</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GLUCOSE-6-PHOSPHATE-NADP-OXIDOREDUCTASE</td>
<td>6.0e-7</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GALACTOSE-NAD-OXIDOREDUCTASE</td>
<td>570.0</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>GLUCOSE-DEHYDROGENASE</td>
<td>3.1e-7</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>3.0e-6</td>
<td>M</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value. Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-MALATE-NADP-OXIDOREDUCTASE[DECARBOXYLATING]</td>
<td>0.051 M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-MALATE-NAD-OXIDOREDUCTASE</td>
<td>6.4e-13 M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-3-HYDROXYACYL-COA-NAD-OXIDOREDUCTASE</td>
<td>6.3e-11 M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>2.17e-10 M</td>
<td>Barman</td>
</tr>
<tr>
<td>3-HYDROXYISOBUTYRATE-NAD-OXIDOREDUCTASE</td>
<td>3.0e-11 M</td>
<td>Barman</td>
</tr>
<tr>
<td>D-3-HYDROXYBUTYRATE-NAD-OXIDOREDUCTASE</td>
<td>0.13 U</td>
<td>Barman</td>
</tr>
<tr>
<td>D-GLYCERATE-NAD-OXIDOREDUCTASE</td>
<td>6.1e-14 M</td>
<td>Barman</td>
</tr>
<tr>
<td></td>
<td>3.55e-13 M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-LACTATE-NAD-OXIDOREDUCTASE</td>
<td>2.76e-6 U</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCOLLATE-NAD-OXIDOREDUCTASE</td>
<td>1.65e-15 M</td>
<td>Barman</td>
</tr>
<tr>
<td>SHIKIMATE-NADP-OXIDOREDUCTASE</td>
<td>0.176 U</td>
<td>Barman</td>
</tr>
<tr>
<td>QUINATE-NAD-OXIDOREDUCTASE</td>
<td>0.00461 M</td>
<td>Barman</td>
</tr>
<tr>
<td>D-MANNITOL-1-PHOSPHATE-NAD-OXIDOREDUCTASE</td>
<td>4.9e-10 M</td>
<td>Barman</td>
</tr>
<tr>
<td>IDITOL-DEHYDROGENASE[SORBITOL]</td>
<td>1.14e-9 M</td>
<td>Barman</td>
</tr>
<tr>
<td>IDITOL-DEHYDROGENASE[XYLITOL]</td>
<td>5.4e-11 M</td>
<td>Barman</td>
</tr>
<tr>
<td>XYLITOL-NADP-OXIDOREDUCTASE[L-XYLULOSE-FORMING]</td>
<td>2.97e-11 M</td>
<td>Barman</td>
</tr>
</tbody>
</table>

continued
### Table 7.6, continued

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-GLYCEROL-3-PHOSPHATE-NAD-OXIREDUCTASE</td>
<td>1.0e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>GLYCEROL-NAD-OXIREDUCTASE</td>
<td>5.1e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>2-3-BUTANEDIOI-NAD-OXIDOREDUCTASE</td>
<td>2.74e-10</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>L-HOMOSERINE-NAD-OXYDOREDUCTASE</td>
<td>0.0011</td>
<td>U</td>
<td>Barman</td>
</tr>
<tr>
<td>ALCOHOL-NAD-OXIDOREDUCTASE</td>
<td>8.0e-12</td>
<td>M</td>
<td>Barman</td>
</tr>
<tr>
<td>PYRUVATE-DEHYDROGENASE</td>
<td>-8.0</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>A-KETOGLUTARATE-DEHYDROGENASE</td>
<td>-8.0</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>ISOCITRATE-DEHYDROGENASE</td>
<td>-5.0</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>6PHOSPHOFRUCTOKINASE</td>
<td>-3.4</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td></td>
<td>-34.9</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>SUCCINATE-DEHYDROGENASE</td>
<td>0.0</td>
<td>kJ/mol</td>
<td>ESTIMATED</td>
</tr>
<tr>
<td>PYRUVATE-DECARBOXYLASE</td>
<td>-12.761001</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>OXALATE-DECARBOXYLASE</td>
<td>-50.461</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>OXALOACETATE-DECARBOXYLASE</td>
<td>-25.161</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>ASPARAGINASE</td>
<td>-12.1</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>CIS-ACONITATE-HYDRATASE</td>
<td>8.4</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>SUCCINATE-COA-LIGASE[GDP]</td>
<td>-0.7</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>D-GLUCOSE-6-PHOSPHATE-KETOL-ISOMERASE</td>
<td>0.4</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-GLYCERALDEHYDE-3-PHOSPHATE-KETOL-ISOMERASE</td>
<td>1.83</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>L-ASPARTATE-AMMONIA-LYASE</td>
<td>18.0</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>2-PHOSPHO-D-GLYCERATE-HYDROLYASE</td>
<td>0.44</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>CITRATE-OXALOACETATE-LYASE[COA-ACETYLATING]</td>
<td>-7.7</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>FRUCTOSE-1,6-DIPHOSPHATE-D-GLYCERALDEHYDE-3-PHOSPHATE-LYASE</td>
<td>5.73</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>L-GLUTAMINE-AMIDOHYDROLASE</td>
<td>-11.7</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>2-3-DIPHOSPHO-D-GLYCERATE-2-PHOSPHO-D-GLYCERATE-PHOSPHOTRANSFERASE</td>
<td>-1.06</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>ATP-CREATINE-PHOSPHOTRANSFERASE</td>
<td>4.1</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>ATP-3-PHOSPHO-D-GLYCERATE-1-PHOSPHOTRANSFERASE</td>
<td>-4.5</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>ATP-PYRUVATE-PHOSPHOTRANSFERASE</td>
<td>-7.5</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td></td>
<td>-14.1</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>ATP-D-HEXOSE-6-PHOSPHOTRANSFERASE</td>
<td>-4.0</td>
<td>kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td></td>
<td>-31.9</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>REDUCED-NADP-NAD-OXIDOREDUCTASE</td>
<td>-2.9</td>
<td>kJ/mol</td>
<td>Hinz</td>
</tr>
</tbody>
</table>

continued
Table 7.6, continued

<table>
<thead>
<tr>
<th>Name</th>
<th>Value. Units</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-6-7-8-TETRAHYDROFOLATE-NADP-OXIDOREDUCTASE</td>
<td>24.1 kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>L-GLUTAMATE-NAD-OXIDOREDUCTASE[DEAMINATING]</td>
<td>32.2 kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>L-ALANINE-NAD-OXIDOREDUCTASE[DEAMINATING]</td>
<td>38.5 kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>D-GLYCERALDEHYDE-3-PHOSPHATE-NAD-OXIDOREDUCTASE[PHOSPHORYLATING]</td>
<td>1.5 kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>L-MALATE-NAD-OXIDOREDUCTASE</td>
<td>7.1 kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td>L-LACTATE-NAD-OXIDOREDUCTASE</td>
<td>29.7 kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td>ALCOHOL-NAD-OXIDOREDUCTASE</td>
<td>6.0 kcal/mol</td>
<td>Lehninger</td>
</tr>
<tr>
<td></td>
<td>25.9 kJ/mol</td>
<td>Hinz</td>
</tr>
<tr>
<td></td>
<td>23.8 kJ/mol</td>
<td>Hinz</td>
</tr>
</tbody>
</table>

Table 7.6 completed
CHAPTER 8

COLLISION-LIMIT BASED ESTIMATION

OF MAXIMUM RATES

FOR ENZYMATIC REACTIONS
8.1. INTRODUCTION

8.1.1. PREVIOUS WORK

It is frequently mentioned in the literature that the maximum rate of a single-reactant single-product irreversible enzymatic reaction is equal to the rate of collisions between the enzyme and the substrate [Hammes and Schimmel, 1970, Fersht, 1977, Hiromi, 1979]:

$$r_{\text{max}} = k e_{\text{total}} [A]$$  \hspace{1cm} (8.1)

where $e_{\text{total}}$ is the concentration of the enzyme, $[A]$ is the concentration of the substrate, and $k$ is the bimolecular collision parameter. Maximum values for $k$ are reported as roughly $10^9 \text{ M}^{-1} \text{s}^{-1}$.

In another treatment [Albery and Knowles, 1976] this maximum is used to evaluate the efficiency of the enzyme, defined as the ratio of the observed rate to the maximum collision rate.

In all of the above previous work, the maximum rate is simply viewed as equal to the rate of encounter of the enzyme with a substrate; no additional constraints, stemming from a more detailed picture of the reaction mechanism or from the reversibility of the reaction, are employed.

8.1.2. THE METHOD PRESENTED HERE

In this work the estimation of maximum rates is extended to multi-substrate, multi-product reactions, reversible or irreversible; a general methodology that allows the estimation of maximum rates for any enzymatic reaction is presented.
This technique assumes a typical fast mechanism for the reaction, with certain constraints imposed on the kinetic parameters. The constraints employed require that kinetic parameters be consistent with the equilibrium constant, and that the rate of each bimolecular step be smaller than the collision rate of the species participating in the step. The kinetic parameters are then determined so that the constraints are satisfied and the rate is optimized.

The method can also be used to estimate a bound on the concentration of a reactant or a product, provided that the reaction rate is known.

In the next sections of this chapter, a more precise statement of the problem is posed, along with necessary assumptions. The problem is then recast in terms of dimensionless quantities, and solve it for several reaction schemes. Reasonable (as opposed to the literature's unrealistically high) collision parameters are then estimated, along with ranges for the dimensionless quantities involved. Simple numerical examples are finally discussed in order to demonstrate the principles involved and present other potential applications and future directions. Extensive application of the method is also involved in Chapter 10, where a case study on the production of lysine is discussed.

8.1.2. SIGNIFICANCE

The significance of this method lies in that it allows the estimation of bounds for parameters related to reaction kinetics, such as rates and concentrations, for which exact data are very often not available. It is equally applicable to enzymatic steps of intracellular biochemical pathways as well as extracellular enzymatic transformations.
As we show in Chapter 10, the method can be used for the identification of kinetic bottlenecks in biochemical pathways. Along with the pathway synthesis method (Chapter 9), the method allows the construction of bypasses to overcome the kinetic bottlenecks.

Although it is implied that the enzyme catalyzing the reaction is a known, existing enzyme, the method can also be used when an enzyme is sought to carry out a certain transformation. If the method determines that, under given conditions, the transformation cannot take place with the desired rate, then it can be a priori concluded that no suitable enzyme exists.
8.2. PROBLEM AND ASSUMPTIONS

For a given enzymatic reaction, it is assumed that the following are known:

- The order of substrate binding
- The order of product release from the enzyme
- The equilibrium constant $K_e$
- The enzyme concentration $e_{total}$
- The substrate and product concentrations

It is further assumed that the conditions of the reaction correspond to an aqueous solution without any macroscopic concentration gradients, and that the enzyme works alone, instead of being part of a multi-enzyme complex. Under these assumptions, the objective is to find the maximum rate of the enzymatic reaction under steady-state conditions.

The reaction attains its maximum rate when its main limitation is the physicochemical collision of the species. The biochemical literature usually refers to this as "diffusion limitation", but the term "collision limitation" will be used in this work, because "diffusion" implies macroscopic concentration gradients which are not considered here.

The steps of dissociation, intramolecular rearrangement, and rearrangement within an enzyme-substrate complex can take place instantaneously. This is justified since the characteristic rate constants can be as high as $10^{10}$ s$^{-1}$ for dissociation [Fersht, 1977], $10^{12}$ s$^{-1}$ for intramolecular rearrangement [Fersht, 1977], but only
$10^6 \text{s}^{-1}$ for a bimolecular collision. The last result can be obtained from Equation (8.1) for $[A]=1\text{mM}$ and $k=10^9 \text{M}^{-1} \text{s}^{-1}$.

An additional assumption in the method is that the species that bind to the enzyme come from the bulk of the solution and not from another site on the enzyme. Thus, $\text{H}^+$, $\text{OH}^-$, and $\text{H}_2\text{O}$ should not participate in the reaction mechanism, as they may be supplied by other sites of the enzyme or a metabolite. This is not a restrictive assumption, but rather a guideline in the use of the methodology: Whenever such species do occur, they are assumed to bind instantaneously, and the corresponding steps must be excluded from the mechanism.
8.3. TREATMENT OF A TWO-REACTANT TWO-PRODUCT ORDERED MECHANISM

To demonstrate the mechanics of the method we will apply it, step by step, for a particular reaction mechanism.

8.3.1. DERIVATION OF THE RATE EQUATION

An ordered mechanism with two reactants A and B and two products P and Q has the form indicated in Figure 8.1. The rate of the reaction at steady state, \( r \), can be determined from the system:

\[
\begin{align*}
    r &= k_1 [E] [A] - k_{-1} [EA] \\
    r &= k_2 [EA] [B] - k_{-2} [EAB] \\
    r &= k_3 [EAB] - k_{-3} [EQ] [P] \\
    r &= k_4 [EQ] - k_{-4} [E] [Q] \\
    e_{\text{total}} &= [E] + [EA] + [EAB] + [EQ]
\end{align*}
\]  

(8.2)  
(8.3)  
(8.4)  
(8.5)  
(8.6)

Since the concentrations of multi-species complexes will not be considered again, it is convenient to drop the brackets from concentrations. As a function of A, B, P, Q, \( e_{\text{total}} \), and the kinetic parameters, the rate is expressed as:

\[
r = e_{\text{total}} (k_1 k_2 k_3 k_{\text{AB}} - k_{-1} k_{-2} k_{-3} k_{\text{PQ}}) D^{-1}
\]

(8.7)

where
\[ D = k_{-4}k_{-3}k_{2}BPQ + k_{-4}k_{-3}k_{-1}PQ + k_{-4}k_{-3}k_{2}PQ + k_{-4}k_{2}k_{3}BQ \\
+ k_{-4}k_{-1}k_{3}Q + k_{-4}k_{-2}k_{-1}Q + k_{-3}k_{1}k_{2}ABP + k_{-3}k_{2}k_{1}AP + k_{-3}k_{-2}k_{-1}P \\
+ k_{1}k_{2}k_{4}AB + k_{1}k_{2}k_{3}AB + k_{2}k_{3}k_{4}B + k_{1}k_{3}k_{4}A + k_{2}k_{1}k_{4} \\
+ k_{-1}k_{3}k_{4} + k_{-2}k_{-1}k_{4} \] (8. 8)
Figure 8.1:

The general ordered mechanism
for a two-reactant two-product enzymatic reaction.
8.3.2. INTRODUCTION OF CONSTRAINTS

There are two kinds of constraints that can be imposed on the kinetic parameters. The first is that the ratio of the forward rate constants to the backward rate constants must be equal to the equilibrium constant:

\[ K_e = \left( k_1 k_2 k_3 k_4 \right) \left( k_{-1} k_{-2} k_{-3} k_{-4} \right)^{-1} \]  \hspace{1cm} (8.9)

The second limitation applies only to (forward or backward) bimolecular steps. Since a collision is required for the two reacting species to form a complex, the rate of each bimolecular step cannot exceed the rate of collision of the two species in the aqueous solution. In effect:

\[ k_i \leq b_i \quad \text{for} \quad i=1,2,3,4 \]  \hspace{1cm} (8.10)

where \( b_i \) is the collision-determined upper bound for the rate constant \( k_i \). Under the constraints (8.9) and (8.10), the objective is to find values for the kinetic parameters, \( k_i \) (for \( i=\pm1, \pm2, \pm3, \) and \( \pm4 \)), so that the rate, \( r \), in Equations (8.7) and (8.8) is maximized.

Note that the kinetic parameters, \( k_i \), are not necessarily intrinsic reaction rate constants, since they may be limited by a physical process such as the collision of the species. Despite this partly phenomenological character of the kinetic parameters the equilibrium relation (8.9) must still be satisfied because the collision rate constants do not depend on the concentrations.
8.3.3. TRANSFORMATION OF PARAMETERS

The mathematical solution of the problem is simplified through the introduction of the parameters:

\[ m_i = k_i / k_i \quad \text{for} \quad i = 1, 2, 3, 4 \]  \hspace{1cm} (8.11)

and substitution of the kinetic parameters \( k_{-1}, k_{-2}, k_3, \) and \( k_4 \) in the rate expression as follows:

\[ k_{-1} = k_1 / m_1 \]  \hspace{1cm} (8.12)

\[ k_{-2} = k_2 / m_2 \]  \hspace{1cm} (8.13)

\[ k_3 = k_{-3} \cdot m_3 \]  \hspace{1cm} (8.14)

\[ k_4 = k_{-4} \cdot m_4 \]  \hspace{1cm} (8.15)

The equilibrium constraint is satisfied by the substitution:

\[ m_3 = K_e / (m_1 m_2 m_4) \]  \hspace{1cm} (8.16)

8.3.4. DERIVATIVES

The parameters \( k_1, k_2, k_{-3}, k_{-4}, m_1, m_2, \) and \( m_4 \) are all non-negative, while \( k_1, k_2, k_{-3}, \) and \( k_{-4} \) are also upper-bounded by collision limits. Working with this set of seven independent variables, the derivatives of the reaction rate are determined as follows:
\[ \frac{\partial r}{\partial k_1} = G \left( k_{-4}^2 k_4^2 k_{-3}^2 k_3^2 k_{-1}^2 m_2^2 m_4^2 K_e \right) (m_1 m_2 m_4 BPQ + m_1 m_4 PQ + m_4 K_e B + K_e BQ) \] 

(8.17)

\[ \frac{\partial r}{\partial k_2} = G \left( k_{-4}^2 k_4^2 k_{-3}^2 k_3^2 k_{-1}^2 m_1^2 m_4^2 K_e \right) (m_1 m_2 m_4 PQ + K_e Q + m_1 m_4 K_e A + m_4 K_e) \] 

(8.18)

\[ \frac{\partial r}{\partial k_3} = G \left( k_{-4}^2 k_4^2 k_{-3}^2 k_3^2 m_4^2 K_e \right) (Q + m_1 m_2 m_4 AB + m_1 m_4 A + m_4) \] 

(8.19)

\[ \frac{\partial r}{\partial k_4} = G \left( k_{-4}^2 k_4^2 k_{-3}^2 k_3^2 m_2^2 m_4^2 K_e \right) (m_1 m_2 m_4 AB + m_4 m_1 m_4 AP + m_4 P + K_e AB) \] 

(8.20)

\[ \frac{\partial r}{\partial m_1} = G \left( k_{-4} k_{-3} k_{-1} k_2 m_4 K_e \right) \left( -k_{-4} k_{-3} k_{-1} k_2 k_3 m_2 m_4 BPQ \right. \]

\[ -k_{-4} k_{-3} k_2 m_4 PQ + k_{-4} k_{-3} k_4 K_e Q -k_{-3} k_1 k_1 k_2 m_2 m_4 ABP \]

\[ -k_{-4} k_{-3} k_1 k_2 m_4 AP + k_{-4} k_{-3} k_1 m_4 K_e -k_{-4} k_1 k_1 k_2 m_2 m_4 AB \]

\[ -k_{-4} k_{-3} k_1 m_2 k_2 m_4 A) \] 

(8.21)

\[ \frac{\partial r}{\partial m_2} = -G \left( k_{-4} k_{-3} k_{-1} m_1^2 k_2 m_4^2 K_e \right) (k_{-4} k_{-3} m_1 k_2 BPQ \]

\[ +k_{-4} k_{-3} k_1 PQ + k_{-3} k_1 m_1 k_2 AB + k_{-4} k_1 m_1 k_2 m_4 AB) \] 

(8.22)
\[
\frac{\partial r}{\partial m_4} = G (k_{-4} k_{-3} k_1 m_1 k_2 K_e) (k_{-4} k_{-3} m_1 k_2 K_e BQ + k_{-4} k_{-3} k_1 m_1 k_2 K_e AB - k_{-4} k_{-3} m_1 k_2 m_2 m_4 AB - k_{-4} k_{-3} k_1 m_1 k_2 m_4^2 A - k_{-4} k_{-3} k_1 k_2 m_4^2) 
\]

where:
\[
G = e^{\text{total} (AB-PQK_e^{-1})} D^{-2}
\]

8.3.5. **OPTIMIZATION**

Since a positive driving force is required for the reaction to take place, G must be positive. Consequently, Equations (8.17) to (8.20) yield:

\[
\frac{\partial r}{\partial k_i} > 0, \quad \text{for } i=1, 2, -3, -4.
\]

Thus, to yield the maximum possible rate, the four bimolecular step parameters \( k_1, k_2, k_{-3}, \) and \( k_{-4} \) must be at their collision limits. Since

\[
\frac{\partial r}{\partial m_2} < 0
\]

and \( m_2 \geq 0 \), in order to maximize the rate \( r \), \( m_2 \) must be equal to 0. This directly implies that \( k_{-2} \to \infty \) since it has already been concluded that \( k_2 \) is non-zero.

Considering that the mechanism steps -2 and 3 compete for the same intermediate, the fact that \( k_{-2} \to \infty \) implies that the concentration of that intermediate is zero, and the reaction can only take place if \( k_3 \to \infty \). The only remaining independent variables are
m₁ and m₄, which do not assume extreme values. They are obtained from the solution of
the system of equations:

\[
\frac{\partial r}{\partial m_1} = 0
\]  
(8.27)

\[
\frac{\partial r}{\partial m_4} = 0
\]  
(8.28)

with k₁, k₂, k₃, and k₄ at their collision limits and m₂ = 0, as indicated above. The
system cannot be solved analytically.

8.3.6. RESULT

Returning to the original set of independent variables, the results can be
summarized as follows. To maximize the rate of the reaction, as given in Equations
(8.7) and (8.8):

• Set k₂ → ∞ and k₃ → ∞

• Set the parameters k₁, k₂, k₃, and k₄ to their collision-determined upper
  bounds

• Obtain k₁ and k₄ from the system:

\[
k_{-1} k^2_{4} k_{-3} k_{2} BK Q + k_{-1} k_{-4} k_{3} k_{e} Q + k_{-1} k_{-4} k_{3} k_{1} k_{2} ABK e = k^2_{4} k^2_{1} k_{2} A + k_{-1} k^2_{4} k_{1} k_{2}
\]  
(8.29)

\[
k_{4} k_{-4} k_{3} k_{2} PQ + k_{4} k_{-3} k_{1} k_{2} AP + k^2_{4} k^2_{1} k_{2} A = k^2_{-1} k_{-4} k_{3} k_{e} Q + k^2_{4} k_{1} k_{4} k_{3} k_{e}
\]  
(8.30)
Under the assumptions stated, substitution of all these values for the kinetic parameters into the rate equation will yield the maximum rate.
8.4. NONDIMENSIONALIZATION

The algebraic expressions and the optimization procedure can be simplified if appropriate dimensionless parameters are introduced.

In this section, the ordered reaction mechanism, for any number of reactants and products, is described in terms of a set of dimensionless parameters. Then the example from the previous section is recast in a dimensionless form, and final results for other mechanisms are listed.

8.4.1. GENERAL ORDERED ENZYMATIC REACTION

For a biochemical reaction with \( n \) reactants and \( m \) products:

\[
\sum_{i=1}^{n} S_i \rightarrow \sum_{i=n+1}^{n+m} S_i
\]

the general ordered mechanism is:

\[
S_i + E_i \rightleftharpoons E_{i+1}, \quad i=1, \ldots, n
\]

\[
E_i \rightleftharpoons E_{i+1} + S_i, \quad i=n+1, \ldots, n+m
\]

For more compact notation, the symbol \( E_{n+m+1} \) is allowed as an alternative token for \( E_1 \).

Let \( t_i \) be the characteristic time of step \( i \), where \( i=\pm 1, \ldots, \pm n \):
\[ t_1 = k_1^{-1} S_i^{-1}, \quad t_i = k_i^{-1}, \quad \text{for } i=1, \ldots, n \quad (8.34) \]
\[ t_i = k_i^{-1}, \quad t_i = k_i^{-1} S_i^{-1}, \quad \text{for } i=n+1, \ldots, n+m \quad (8.35) \]

### 8.4.2. DIMENSIONLESS PARAMETERS

The dimensionless parameters that will be used are:

\[ u_i = \frac{t_i}{t_1}, \quad i=2, \ldots, n \quad \text{or} \quad i=-(n+m), \ldots, -(n+1) \quad (8.36) \]

\[ h_i = \frac{t_i}{t_1}, \quad i=1, \ldots, n \quad \text{or} \quad i=-(n+m), \ldots, -(n+1) \quad (8.37) \]

\[ f = \frac{\prod_{i=n+1}^{n+m} S_i^{-1}}{\prod_{i=1}^{n} S_i} \quad (8.38) \]

\[ e_i = \frac{E_i}{n+m-1} = \frac{E_i}{E_{\text{total}}} \quad (8.39) \]

\[ r_{\text{nd}} = \frac{r \cdot t_1}{e_{\text{total}} (1-f)} \quad (8.40) \]

### 8.4.2.1. Physical Significance of the Parameters

Using \( t_1 \) as the global time scale of reference, the physical interpretation of the dimensionless parameters is as follows:
• \( u_i \) is simply the scaled characteristic time of step \( i \). It can also be viewed as the "activity" of \( S_i \) scaled by the activity of \( S_1 \). Two compounds can have the same activity \( u_i \), even with different concentrations, if their rate constants for reactions to enzyme species are inversely proportional to their concentrations.

• \( h_i \) is the equilibrium constant of step \( i \), with respect to enzyme species alone. At equilibrium, \( h_i \) is equal to \( e_{i+1}/e_i \).

• \( f \) is the mass action ratio divided by the equilibrium constant; the dimensionless driving force for the reaction is \( 1-f \).

• \( e_i \) is the concentration of enzyme species \( E_i \) as a fraction of total enzyme.

• \( r_{nd} \) is the dimensionless rate per unit driving force, per mole of total enzyme, per unit time \( t_1 \). It is essentially the inverse of the dimensionless "resistance" that the mechanism puts up against the driving force.

### 8.4.3. DIMENSIONLESS TREATMENT OF THE EXAMPLE

In terms of dimensionless parameters the two-reactant two-product system (Equations 8.2 to 8.6) can be written as:

\[
\begin{align*}
    r_{nd} &= e_1^{-1} - e_2^{-1}/h_1 \\
    r_{nd} &= u_2^{-1} (e_2^{-1} - e_3^{-1}/h_2) \\
    r_{nd} &= u_3^{-1} (e_3^{-1}/h_3 - e_4^{-1})
\end{align*}
\] (8.41) (8.42) (8.43)
\[ r_{nd} = u^{-1}_4 (e^{4}_1 / h^{4}_1 - e^{1}_1) \] 
(8.44)

\[ e^{1}_1 + e^{2}_1 + e^{3}_1 + e^{4}_1 = 1 \] 
(8.45)

\[ f = h^{4}_3 / h^{4}_1 h^{2}_2 \] 
(8.46)

The maximum rate for the reaction, in terms of dimensionless parameters, can be expressed as:

\[ r_{nd} = (u^{-1}_3 h^{4}_1 f h^{-1}_4 + u^{-1}_3 f h^{-1}_4 + u^{-1}_2 h^{4}_1 h^{-1}_1 + u^{-1}_4 h^{-1}_1 f + u^{-1}_3 f + h^{-1}_1 + u^{-1}_4 f + u^{-1}_2 + u^{-1}_4 h^{-1}_4 + h^{-1}_4 + 1)^{-1} \] 
(8.47)

where \( u_2, u_3, \) and \( u_4 \) are collision limited and \( h_1 \) and \( h_4 \) are determined from the following system of equations:

\[ u^{-1}_3 h^{2}_1 f + u^{-1}_4 h^{2}_1 h^{-1}_4 f + h^{2}_1 h^{-1}_4 f = h^{2}_4 u^{-1}_2 + h^{-1}_4 u_2 \] 
(8.48)

\[ u^{-1}_3 h^{2}_1 f + u^{-1}_3 h^{-1}_1 f = u^{-1}_2 h^{2}_4 - u^{-1}_4 h^{2}_1 h^{-1}_4 h^{2}_1 h^{-1}_4 \] 
(8.49)

### 8.4.4. RESULTS FOR OTHER ORDERED MECHANISMS

Similarly to the above example, the optimization results for all ordered reaction schemes that involve three or fewer reactants and three or fewer products have been obtained, and are listed in Tables 8.1 to 8.9. The results have been, whenever possible, simplified algebraically.
Table 8.1:

Result of dimensionless maximization of the rate for an ordered mechanism with 1 reactant and 1 product.

<table>
<thead>
<tr>
<th>Reaction: $S_1 \rightarrow S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate:</td>
</tr>
</tbody>
</table>
| \[
<table>
<thead>
<tr>
<th>\frac{1}{r_{nd}} = u_{2}^{-f+1}</th>
</tr>
</thead>
</table>
Table 8.2:
Result of dimensionless maximization of the rate for an ordered mechanism with 1 reactant and 2 products.

<table>
<thead>
<tr>
<th>Reaction: ( S_1 \rightarrow S_2 + S_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate: ( \frac{1}{r_{nd}} = \frac{u_{-2} f + u_{-3} f + 2 h_{-3} u_{-3} u_{-3} + 2 h_{-3} + 1}{h_{-3}} )</td>
</tr>
<tr>
<td>With ( h_{-3} ) from the equation: ( h_{-3} = \left( \frac{u_{-2} f}{u_{-3} + 1} \right)^{1/2} )</td>
</tr>
</tbody>
</table>
Table 8.3:

Result of dimensionless maximization of the rate for an ordered mechanism
with 1 reactant and 3 products.

Reaction: \( S_1 \rightarrow S_2 + S_3 + S_4 \)

Rate:
\[
\frac{1}{r_{nd}} = 2 \frac{u_{-2}f^2}{h_{-3}h_{-4}} + \frac{u_{-2}f}{h_{-3}} + u_{-2}f + 2 \frac{u_{-3}f}{h_{-4}} + u_{-3}f + u_{-4}f + h_{-3}u_{-3}^{+1}
\]

With \( h_{-3} \) and \( h_{-4} \) from the equations:
\[
h_{-4}u_{-2}f + u_{-2}f = h_{-3}^2h_{-4}u_{-3} + h_{-3}h_{-4}^2u_{-4} + h_{-3}^2h_{-4}^2
\]
\[
u_{-2}f + u_{-3}h_{-3}f = h_{-3}^2h_{-4}u_{-4} + h_{-3}^2h_{-4}^2 + h_{-3}h_{-4}^2u_{-4} + h_{-3}^2h_{-4}^2
\]
Table 8.4:
Result of dimensionless maximization of the rate for an ordered mechanism
with 2 reactants and 1 product.

<table>
<thead>
<tr>
<th>Reaction: $S_1 + S_2 \rightarrow S_3$</th>
</tr>
</thead>
</table>
| Rate: \[
\frac{1}{r_{nd}} = u_3 f + \frac{2u_2}{h_1} + u_2 + 1
\]
| With $h_1$ from the equation: \[
h_1 = \left( \frac{u_2}{u_3 f + f} \right)^{1/2}
\] |
Table 8.5:
Result of dimensionless maximization of the rate for an ordered mechanism
with 2 reactants and 2 products.

Reaction: $S_1 + S_2 \rightarrow S_3 + S_4$

Rate:
$$\frac{1}{r_{nd}} = \frac{u_{-3} h_1 f}{h_{-4}} + 2 \frac{u_{-3} f}{h_{-4}} + u_{-3} f + u_{-4} f + \frac{h_{-4} u_2}{h_1} + 2 \frac{u_2}{h_1} + u_2 + 1$$

With $h_1$ and $h_{-4}$ from the equations:
$$u_{-3} h_1^2 f + h_{-4} u_{-4} h_1 f + h_{-4} h_1^2 f = h_{-4} u_2 + h_{-4} u_2$$
$$u_{-3} h_1^2 f + u_{-3} f = h_{-4} u_2 + h_{-4} u_{-4} h_1 + h_{-4} h_1$$
Table 8.6:
Result of dimensionless maximization of the rate for an ordered mechanism with 2 reactants and 3 products.

<table>
<thead>
<tr>
<th>Reaction: ( S_1 + S_2 \rightarrow S_3 + S_4 + S_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
\frac{1}{r_{nd}} &= 2 \frac{u_{-3} h_1 f}{h_{-5} h_{-4}} + 2 \frac{u_{-4} h_1 f}{h_{-5}} + u_{-5} h_1 f + h_1 f + 2 \frac{u_{-3} f}{h_{-5} h_{-4}} + \frac{u_{-4} f}{h_{-5}} \\
&+ u_{-3} f + 2 \frac{u_{-4} f}{h_{-5}} + u_{-4} f + u_{-5} f + \frac{h_{-5}}{h_1} \frac{h_{-4} u_2}{h_1} + u_2 + h_{-4} u_{-4} + 1
\end{align*}
\] |
| With \( h_1, h_{-4}, \) and \( h_{-5} \) from the equations: |
| \[
\begin{align*}
u_{-3} h_1^2 f + h_{-4} u_{-4} h_1^2 f + h_{-5} u_{-5} h_{-4} h_1^2 f + h_{-5} h_{-4} h_1^2 f &= 0 \\
&= h_{-5}^2 h_{-4}^2 u_2 + h_{-5}^2 h_{-4} u_2 + h_{-5}^2 h_{-4} u_2 \\
u_{-3} h_1^2 f + h_{-5} u_{-3} h_1 f + u_{-3} h_1 f &= 0 \\
&= h_{-5}^2 h_{-4}^2 u_2 + h_{-5}^2 h_{-4} u_2 + h_{-5}^2 h_{-4} u_2 + h_{-5}^2 h_{-4} h_1^2 f \\
h_{-4} u_{-4} h_1^2 f + h_{-4} u_{-4} h_1^2 f + h_{-5} u_{-3} h_1 f &= 0 \\
&= h_{-5}^2 h_{-4}^2 u_2 + h_{-5}^2 h_{-4} u_2 + h_{-5}^2 h_{-4} h_1^2 f + h_{-5}^2 h_{-4} h_1^2 f \\
&= h_{-5}^2 h_{-4}^2 u_2 + h_{-5}^2 h_{-4} h_1^2 f + h_{-5}^2 h_{-4} h_1^2 f + h_{-5}^2 h_{-4} h_1^2 f
\end{align*}
\] |
Table 8.7:

Result of dimensionless maximization of the rate for an ordered mechanism with 3 reactants and 1 product.

<table>
<thead>
<tr>
<th>Reaction: $S_1 + S_2 + S_3 \rightarrow S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate: $\frac{1}{r_{nd}} = h_2 u_2 f + u_{-4} f + 2 \frac{u_3}{h_1 h_2} + \frac{u_3}{h_2} + u_{3} + 2 \frac{u_2}{h_1} + u_2 + 1$</td>
</tr>
</tbody>
</table>

With $h_1$ and $h_2$ from the equations:

$h_1 h_2 u_2 f + u_{-4} h_1 h_2 f + h_1 h_2 f = h_1 u_3 + u_3$

$u_{-4} h_1 h_2 f + h_1 h_2 f - h_1 h_2 u_2 f = h_2 u_2 - h_1 u_3$
Table 8.8:

Result of dimensionless maximization of the rate for an ordered mechanism:

with 3 reactants and 2 products.

<table>
<thead>
<tr>
<th>Reaction: ( S_1 + S_2 + S_3 \rightarrow S_4 + S_5 )</th>
</tr>
</thead>
</table>

**Rate:**

\[
\frac{1}{r_{nd}} = h_2 u_2 f + \frac{u_{-4} f}{h_{-5}} + u_{-4} f + u_{-5} f + 2 \frac{h_{-5} u_3}{h_1 h_2} + 2 \frac{u_3}{h_1 h_2} + \frac{h_{-5} u_2}{h_1} + 2 \frac{u_2}{h_1} + u_2 + h_{-5} u_{-5} + h_{-5} + 1
\]

With \( h_1, h_2, \) and \( h_{-5} \) from the equations:

\[
u_{-4} h_1^2 h_2 f + h_{-5} u_{-5} h_1^2 h_2 f + h_{-5} h_1^2 h_2 f - h_{-5} h_1 h_2^2 u_2 f =
\]

\[= h_{-5} h_2 u_2 + h_{-5} h_2 u_2 - h_{-5} h_1 u_3\]

\[h_{-5} h_1 h_2 u_2 f + u_{-4} h_1^2 h_2 f + h_{-5} u_{-5} h_1 h_2 f + h_{-5} h_1 h_2 f =
\]

\[= h_{-5} h_1 u_3 + h_{-5} u_3 + h_{-5} u_3\]

\[u_{-4} h_1^2 h_2 f + u_{-4} h_1^2 h_2 f + u_{-4} h_1 h_2 f =
\]

\[= h_{-5} u_3 + h_{-5} h_2 u_2 + h_{-5} u_{-5} h_1 h_2 + h_{-5} h_1 h_2\]
Table 8.9:

Result of dimensionless maximization of the rate for an ordered mechanism with 3 reactants and 3 products.

Reaction: \( S_1 + S_2 + S_3 \rightarrow S_4 + S_5 + S_6 \)

Rate:

\[
\frac{1}{r_{nd}} = h_2 u_2 f + \frac{u_{-4} f}{h_{-6} h_{-5}} + \frac{u_{-4} f}{h_{-5}} + u_{-4} f + \frac{u_{-5} f}{h_{-6}} + u_{-5} f + u_{-6} f + 2 \frac{h_{-6} h_{-5} u_3}{h_1 h_2} + 2 \frac{h_{-6} u_3}{h_1 h_2} + 2 \frac{u_3}{h_1 h_2} + 2 \frac{u_3}{h_1} + 2 \frac{h_{-6} h_{-5} u_2}{h_1} + 2 \frac{h_{-6} u_2}{h_1} + 2 \frac{u_2}{h_1} + u_2 + h_{-5} + h_{-6} u_{-5} + h_{-6} u_{-6} + h_{-6} + 1
\]

(continued on the next page)
With \( h_1, h_2, h_5, \) and \( h_6 \) from the equations:

\[
\begin{align*}
&u_{-6} h_1^2 f + h_{-5} u_{-4} h_1^2 f + h_{-5} h_{-4} h_1^2 f + h_{-5} h_{-5} h_1^2 f - \\
&h_{-5} h_1 h_2^2 u_2 f = h_{-5} h_1 h_2^2 u_2 + h_{-5} h_1 h_2^2 u_2 + h_{-5} h_1 h_2^2 u_2 - h_{-5} h_1 u_3
\end{align*}
\]

\[
\begin{align*}
&h_{-6} h_{-5} h_1 h_2^2 u_2 f + u_{-4} h_{-5} h_1 h_2^2 u_2 f + h_{-5} u_{-4} h_{-5} h_1 h_2^2 f + h_{-6} u_{-5} h_{-5} h_1 h_2^2 f + \\
&h_{-6} h_{-5} h_1 h_2^2 f = h_{-6} h_{-5} h_1 u_3 + h_{-6} h_{-5} h_1 u_3 + h_{-6} h_{-5} h_1 u_3 + h_{-6} h_{-5} h_1 u_3
\end{align*}
\]

\[
\begin{align*}
&u_{-4} h_1 h_2 f + u_{-4} h_1 h_2 f + h_{-6} u_{-4} h_1 h_2 f + u_{-4} h_1 h_2 f = h_{-6} h_{-5} u_3 + \\
&h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2
\end{align*}
\]

\[
\begin{align*}
&h_{-5} h_{-4} h_1 h_2^2 f + h_{-5} h_{-4} h_1 h_2^2 f + h_{-5} h_{-5} h_1 h_2^2 f - h_{-4} h_{-5} h_1 h_2^2 f = h_{-6} h_{-5} u_3 + \\
&h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2 + h_{-6} h_{-5} h_1 h_2
\end{align*}
\]
8.5. EXAMPLE OF ASYMPTOTIC BEHAVIOR

The dimensionless quantities allow easier examination of limiting cases that facilitate the understanding of the maximum-rate concepts. These cases are particularly useful in light of the complexity of the general analytic expressions.

8.5.1. EFFECT OF $f$ IN A SIMPLIFIED SYSTEM

As an example, the effect of $f$ (the mass action ratio divided by the equilibrium constant) on the maximum rate of a bioreaction with one reactant and three products will be examined.

To reduce the number of independent parameters, the analysis is restricted, by assuming that $u_i=1$ for $i=-2,-3,-4$. This is the case when the products have approximately equal molecular weights and the concentration of each product is roughly 60% of the concentration of the reactant.

This restriction simplifies the equations for the maximum rate (Table 8.3) to:

$$r_{nd} = \left( fh^{-1}h^{-1} + fh^{-1} + fh^{-1} + 3f + 2h h + h + 2h + 1 \right)^{-1}$$  \hspace{1cm} (8. 50)

with $h_3$ and $h_4$ being the solution of the following system, simplified through some algebraic manipulation:

$$f = 2h^2h_3$$  \hspace{1cm} (8. 51)

$$2h_4(1+h_4) = h_3(1+2h_4)$$  \hspace{1cm} (8. 52)
8.5.2. REACTION VERY FAR FROM EQUILIBRIUM

To find the asymptotic behavior when \( f << 1 \), Equations (8.51) and (8.52) are solved for \( h_3 \) and \( h_4 \), and substituted in Equation (8.50) to yield \( r_{nd} \).

8.5.2.1. First-Order Approximation

As a first approximation, only terms of order \( f^{1/3} \) are retained:

\[
\begin{align*}
h_3 & = (2f)^{1/3} \\
h_4 & = (f/4)^{1/3} \\
r_{nd}^{-1} & = 1 + 3 (2f)^{1/3}
\end{align*}
\] (8.53) (8.54) (8.55)

This result does not offer sufficient accuracy, because, even for quite small values of \( f \), terms of order \( f^{1/3} \) can be comparable to 1.

8.5.2.2. Second-Order Approximation

To ensure the applicability of the approximation, it is wise to obtain the next term in the asymptotic expansion by retaining terms of order \( f^{2/3} \) as well:

\[
\begin{align*}
h_3 & = (2f)^{1/3} - \frac{1}{3} (2f)^{2/3} \\
h_4 & = (f/4)^{1/3} + \frac{1}{12} (2f)^{2/3} \\
r_{nd}^{-1} & = 1 + 3 (2f)^{1/3} + \frac{5}{2} (2f)^{2/3}
\end{align*}
\] (8.56) (8.57) (8.58)
8.5.2.3. Region of Influence of the Reversibility of the Reaction

Application of Equations (8.56), (8.57), and (8.58) shows that even for a reaction removed 100-fold from equilibrium ($f=0.01$, i.e., the mass action ratio is 100 times smaller than the equilibrium constant) the maximum rate is less than half the rate expected for an irreversible reaction ($f=0$).

Hence, the fact that the reaction is not fully irreversible exerts a significant effect on the maximum rate.

8.5.3. REACTION NEAR EQUILIBRIUM

At the other end of the spectrum, for $f \rightarrow 1$ Equations (8.50), (8.51), and (8.52) after similar solution and substitution yield:

$$h_{-3} = 1 - \frac{1 - f}{2 + 2^{1/2}}$$  \hspace{1cm} (8.59)

$$h_{-4} = \frac{1}{2^{1/2}} - \frac{1 - f}{4}$$  \hspace{1cm} (8.60)

$$r_{nd} = \left( 2^{1/2} 3 + (3 + 2^{-1/2}) (1 + f) \right)^{-1} = \left( 7.95 + 3.71 f \right)^{-1}$$  \hspace{1cm} (8.61)

For $f \rightarrow 1$, $r_{nd}$ varies approximately linearly with $f$, and no severe effect of $f$ on the rate is observed. It should be noted that $r_{nd}$ is already expressed as rate per unit driving force $(1-f)$, so the effect of $f$ examined here is in addition to the effect expected solely from the reduction of the driving force. Figure 8.2 shows the numerical solution
of the example, for any value of $f$, along with the two asymptotic cases that were examined.
Figure 8.2:

General numerical solution and analytic asymptotic solutions for one-reactant three-product enzymatic reactions, with \( u_1 = 1 \), \( u_2 = 1 \), and \( u_3 = 1 \); (a) asymptotic solution for \( f \to 0 \), retaining terms of order \( f^{1/3} \); (b) asymptotic solution for \( f \to 0 \), retaining terms of order \( f^{2/3} \); (c) asymptotic solution for \( f \to 1 \); (d) General solution for any value of \( f \) between 0 and 1.
8.6. ESTIMATION OF COLLISION PARAMETERS

Collision-determined upper bounds for the kinetic parameters, along with other ranges useful in the methodology, are estimated in this section. The collision limit \( k \) for a reaction between a substrate \( S \) (of small molecular weight) and an enzyme \( E \):

\[
E + S \rightarrow^{k} ES
\]

is given by [Hiromi, 1979]:

\[
k = \rho_S \rho_E k_0
\]

where \( \rho_S \) and \( \rho_E \) are steric factors, smaller than one, and \( k_0 \) the encounter constant.

We will estimate each of these three parameters below.

8.6.1. STERIC FACTORS

8.6.1.1. Substrate

The steric factor \( \rho_S \) ranges from as small as roughly 0.01 (for large molecules) to as large as comparable to 1 (for small molecules).

The participation of water, \( H^+ \), or \( OH^- \) in reactions will be neglected, because, as discussed in Section 2, these molecules do not necessarily originate from the bulk of the solution. Other molecules occurring in biochemical reactions are generally complicated enough to have unreactive orientations, leading to \( \rho_S \) significantly smaller than 1.

A conservatively high value will be nevertheless assumed for \( \rho_S \). Specifically, it is assumed that \( \rho_S = 1/3 \), signifying that one third of the possible orientations of \( S \) are suitable for binding.
8.6.1.2. Enzyme

Since the active site of an enzyme is the area surrounding only a small number of bonds, it is modelled here as a circle of diameter 8 Å, or $8 \times 10^{-10}$ m. The point of interest here is essentially the inaccuracy allowed in the collisions. The value assumed above implies that, if the substrate misses the exact binding spot by more than 8 Å, the collision will be ineffective.

In this sense, estimation of $\rho_E$ as the area of the active site divided by the total area of the enzyme [Hiromi, 1979] yields

\[ \rho_E = \left( \frac{2 \times 10^{-10}}{r_E} \right)^2 \]

(8.64)

where $r_E$ is the enzyme radius.

8.6.2. APPROXIMATE RATE OF ENCOUNTER

For uncharged molecules, the rate of encounter $k_0$ is given by [Gutfreund, 1972, Hiromi, 1979]:

\[ k_0 = 4 \pi N (D_E + D_S) (r_E + r_S), \quad \text{in} \quad M^{-1}s^{-1} \]

(8.65)

where $N$ is Avogadro's number ($6 \times 10^{26}$ kmol$^{-1}$), $D_E$ and $D_S$ are the diffusion coefficients of the enzyme and the substrate in m$^2$/s, and $r_E$ and $r_S$ are molecular radii of the enzyme and the substrate in m.
8.6.2.1. Simplifications Due to the Relative Sizes of the Species

Since E is a big molecule and S a small molecule, $D_S$ is much larger than $D_E$, and $r_E$ is much larger than $r_S$. This allows the elimination of the effect $D_E$ and $r_S$, through the introduction of a constant correction constant factor equal to 1.25:

\[
(D_E + D_S) (r_E + r_S) \approx 1.25 D_S r_E
\]  
(8.66)

Substitution of Equation (8.66) in Equation (8.65) then yields:

\[
k_0 \approx 5 \pi N D_S r_E
\]  
(8.67)

8.6.2.2. Substrate Diffusion Coefficient

For the diffusion coefficient of the substrate, data from the CRC Handbook of Chemistry and Physics [Weast and Astle, 1985] yield the approximation:

\[
D_S = 7.3 \times 10^{-9} M_S^{-0.45}, \quad \text{at } 25^\circ\text{C}
\]  
(8.68)

where $M_S$ is the molecular weight of S. The elevation of $D_S$ with temperature is roughly 2% to 3% per degree.

8.6.2.3. Enzyme Radius

The radius of the enzyme molecule is related to the enzyme diffusivity through the Stokes-Einstein relation:

\[
D_E = \frac{RT}{6\pi N r_E \eta}
\]  
(8.69)
For viscosity \( \eta \) equal to that of the solvent (water), and using \( D_E \) data from the literature [Hlroml, 1979], Equation (8.69) yields:

\[
\tau_E = 4.5 \times 10^{-11} M_E^{0.38}
\]  

(8.70)

where \( M_E \) is the molecular weight of E.

8.6.2.4. Resulting Rate of Encounter

Through substitution of Equations (8.68) and (8.70) into Equation (8.67), \( k_0 \) can be expressed as an empirical function of the molecular weights \( M_E \) and \( M_S \):

\[
k_0 = 3 \times 10^9 M_S^{-0.45} M_E^{0.38}, \text{ in } M^{-1} s^{-1}
\]  

(8.71)

8.6.3. FINAL COLLISION PARAMETER

Substitution of Equations (8.64) and (8.71) into Equation (8.63), under the stated assumption that \( \rho_S = 1/3 \), yields an expression of \( k \) as an empirical function of the molecular weights \( M_E \) and \( M_S \):

\[
k = 2 \times 10^{10} M_S^{-0.45} M_E^{-0.38}, \text{ in } M^{-1} s^{-1}
\]  

(8.72)

This relation (8.72) is valid if one assumes that there are no electrostatic effects, the viscosity of the solution is equal to that of water, and the temperature is 25\textdegree C.
8.6.4. RANGES FOR OTHER PARAMETERS

Rough ranges and typical values for some of the parameters discussed, based on the above result, are shown in Table 8.10. The correlation between $k_1$ and $k_1$, due to the fact that the molecular weight $M_E$ is the same, was taken into account for the estimation of the range of $u_1$ in Table 8.10. The numbers are only meant to be used in the investigation of realistic parameter values. The bounds are not absolute and the "typical values" are just order-of-magnitude estimates and not statistical averages of observed values.
Table 8.10:
Ranges for parameters pertinent in the maximum-rate methodology.

<table>
<thead>
<tr>
<th>QUANTITY</th>
<th>MINIMUM</th>
<th>MAXIMUM</th>
<th>TYPICAL VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$</td>
<td>$5 \times 10^{-6}$ M</td>
<td>$5 \times 10^{-3}$ M</td>
<td>$10^{-4}$ M</td>
</tr>
<tr>
<td>$M_S$</td>
<td>20</td>
<td>800</td>
<td>200</td>
</tr>
<tr>
<td>$M_E$</td>
<td>$10^4$</td>
<td>$10^6$</td>
<td>$10^5$</td>
</tr>
<tr>
<td>Collision $k_i$</td>
<td>$6 \times 10^6$ M$^{-1}$s$^{-1}$</td>
<td>$2 \times 10^8$ M$^{-1}$s$^{-1}$</td>
<td>$2 \times 10^7$ M$^{-1}$s$^{-1}$</td>
</tr>
<tr>
<td>$t_i=1/k_iS_i$</td>
<td>$3 \times 10^{-2}$ s</td>
<td>$10^{-6}$ s</td>
<td>$4 \times 10^{-4}$ s</td>
</tr>
<tr>
<td>$u_i=t_i/t_1$</td>
<td>$2 \times 10^{-4}$</td>
<td>$6 \times 10^3$</td>
<td>1</td>
</tr>
<tr>
<td>$f$</td>
<td>$10^{-12}$</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>$e_{total}$</td>
<td>$2 \times 10^{-8}$ M</td>
<td>$10^{-4}$ M</td>
<td>$10^{-6}$ M</td>
</tr>
<tr>
<td>$r/e_{total}$</td>
<td>$2$ s$^{-1}$</td>
<td>1000 s$^{-1}$</td>
<td>$100$ s$^{-1}$</td>
</tr>
<tr>
<td>$r_{nd}$</td>
<td>$10^2$</td>
<td>$10^9$</td>
<td>$3 \times 10^5$</td>
</tr>
</tbody>
</table>
8.7. NUMERICAL EXAMPLES

In this section, simple examples of the application of this methodology are presented. More substantial applications, involving the identification and elimination of kinetic bottlenecks, will be discussed in Chapter 10, in the context of a case study on the synthesis of lysine.

8.7.1. ISOLATED REACTIONS

A schematic of simple applications of the maximum-rate estimation method is shown in Figure 8.3. We perform estimations for two reactions from the serine pathway:

- In the reaction *Phosphoserine transaminase*, we assume values for the concentrations of the reactants and products, and estimate the enzyme turnover.

- In the other reaction, *Phosphoglycerate dehydrogenase*, we assume a value for the enzyme turnover and for all except one of the concentrations. By solving the same maximum-rate equations (with respect to the concentration rather than the rate), we estimate a bound for the missing concentration.
phosphoserine transaminase

\[
\begin{align*}
\text{a-keto-glutarate: } & 1.0 \text{ mM} \\
\text{P-serine: } & 2.0 \text{ mM} \\
\text{P-OH-pyruvate: } & 2.5 \text{ mM} \\
\text{glutamate: } & 1.0 \text{ mM}
\end{align*}
\]

maximum rate: 1000 mol substr./mol enz. s

PG-dehydrogenase

maximum rate no less than: 1000

\[
\begin{align*}
\text{NAD+: } & 2.0 \text{ mM} \\
\text{P-OH-pyruvate: } & 1.5 \text{ mM} \\
\text{NADH: } & 0.4 \text{ mM}
\end{align*}
\]

3-Phosphoglycerate no less than 2.0 mM

Figure 8.3:

Simple maximum-rate estimations for reactions from the serine pathway
8.7.2. A SHORT P. THWAY

We will estimate the maximum rate of the reactions in a section (shown in Figure 8.4) of the glycolytic pathway.

8.7.2.1. Parameters for Small Molecules

Estimates for the necessary parameters of the metabolites involved are shown in Table 8.11.

The concentrations for the pathway intermediates in Table 8.11 were taken to be equal to their concentrations in human erythrocytes, reported in the literature [Lehninger 1975].

8.7.2.2. Parameters for Enzymes

The molecular weights of enzymes are often not known, and the molecular weight of each enzyme in the pathway is assumed here to be equal to 40,000. This is probably an underestimate for most enzymes and will result in higher values for the maximum rates. This follows a general strategy: When only a rough range for a parameter is known, the parameter is assumed to have the value that leads to a conservative overestimate of the maximum rate.

The equilibrium constant of each step (Table 8.12) can be estimated from Gibbs Free-Energy values from the literature [Lehninger 1975].
Figure 8.4:
A section of the glycolytic pathway.
Table 8.11:

Estimation of important parameters for intermediates of the glycolytic pathway.

<table>
<thead>
<tr>
<th>Intermediate</th>
<th>Mol. Weight</th>
<th>Collision $k_l$</th>
<th>Concentration $\left(\frac{k_i S_i}{I_j}\right)^{-1} = t_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP</td>
<td>503</td>
<td>2.2</td>
<td>1850 $\mu$M</td>
</tr>
<tr>
<td>ADP</td>
<td>424</td>
<td>2.3</td>
<td>138 $\mu$M</td>
</tr>
<tr>
<td>Glucose</td>
<td>180</td>
<td>3.5</td>
<td>5000 $\mu$M</td>
</tr>
<tr>
<td>Glc6P</td>
<td>258</td>
<td>2.9</td>
<td>83 $\mu$M</td>
</tr>
<tr>
<td>Fru6P</td>
<td>258</td>
<td>2.9</td>
<td>14 $\mu$M</td>
</tr>
<tr>
<td>FruDP</td>
<td>336</td>
<td>2.6</td>
<td>31 $\mu$M</td>
</tr>
</tbody>
</table>
8.7.2.3. Results

Table 8.12 shows the maximum-rate calculations for the analyzed steps of the pathway. To obtain values for the maximum rate itself, values for enzyme concentrations are required. Since these are not known, maximum specific activities for the enzymes \( \frac{r}{e_{\text{total}}} \) were calculated, in units of moles of substrate per mole of enzyme per second. Some other interesting results are worth mentioning:

(a) If the order of the products or reactants are reversed, the maximum rate estimate change by less than 15%. Considering the other uncertainties in the estimation, the binding order is not expected to have a significant effect on the maximum rate.

(b) If the reaction catalyzed by \textit{Hexokinase} is assumed to be irreversible, the constraint relating the kinetic parameters to the equilibrium constant is removed, and a simplified algebraic expression is obtained for the dimensionless maximum rate:

\[
  r_{nd} = (1+u_2)^{-1}
\]  

(8.73)

Application of Equation (8.73) yields \( \frac{r}{e_{\text{total}}} = 3.2 \times 10^4 \). This result does not deviate significantly from the result obtained without the irreversibility assumption \( \frac{r}{e_{\text{total}}} = 2.9 \times 10^4 \). In effect it is quite reasonable to assume that the reaction is irreversible.

Using the results of the calculations, whenever enzyme concentrations are given, the maximum rate for each step and the overall maximum rate for the pathway can be calculated. If, on the other hand, the overall pathway rate is given, the minimum
required concentration for each enzyme can be calculated, and possible bottlenecks, where high enzyme concentrations are required, can be predicted.
Table 8.12:
Estimation of maximum-rate parameters for some steps of the glycolytic pathway.

<table>
<thead>
<tr>
<th>Step</th>
<th>$K_e$</th>
<th>$f$</th>
<th>$r_{nd}$</th>
<th>$r/e_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hexokinase</td>
<td>$8.6 \times 10^2$</td>
<td>$1.4 \times 10^{-6}$</td>
<td>0.17</td>
<td>$2.9 \times 10^4$</td>
</tr>
<tr>
<td>Glucose-6P-Isomerase</td>
<td>$5.0 \times 10^{-1}$</td>
<td>$3.4 \times 10^{-1}$</td>
<td>0.65</td>
<td>$5.3 \times 10^2$</td>
</tr>
<tr>
<td>Phosphofructokinase</td>
<td>$3.1 \times 10^2$</td>
<td>$5.3 \times 10^{-4}$</td>
<td>0.89</td>
<td>$3.6 \times 10^2$</td>
</tr>
</tbody>
</table>
8.8. EXTENSIONS

8.8.1. EXTREMA FOR OTHER PARAMETERS

Although the problem was stated as maximization of the rate, the same procedures and formulae also answer the following optimization questions:

(a) What is the minimum substrate concentration $S_j$ that can achieve a desired or observed rate $r$, for given concentrations of the remaining substrates and products $S_k$, $k \neq l$, and given enzyme concentration $e_{total}$?

(b) What is the minimum enzyme concentration $e_{total}$ that can achieve a given rate $r$, for given substrate and product concentrations $S_l$?

(c) What is the maximum concentration for a product $S_p$ that can accommodate a given rate $r$, for given concentrations of the remaining products and substrates $S_k$, $k \neq p$, and given enzyme concentration $e_{total}$? This upper bound for the concentration will normally be quite high, and thus useless, but when the reaction is close to equilibrium reasonable concentrations will be obtained.

(d) What is the minimum equilibrium constant $K_e$ that allows a given rate $r$, for given substrate and product concentrations $S_k$, and given enzyme concentration $e_{total}$? In this case, an obvious minimum is the mass action ratio of the reaction, but this methodology will yield a tighter (i.e., larger) minimum.
Case (b) was already mentioned in the example. Cases (a) and (c) combined can yield both upper and lower bounds for the concentration of an intermediate in a metabolic pathway, when the rate of the pathway and concentrations of all the other metabolites are given.

8.8.2. EXPLOITATION OF ADDITIONAL CONSTRAINTS

Further narrowing of the bounds yielded by the methodology (i.e., increase of the lower bounds and decrease of the upper bounds) can be achieved through the exploitation of partial information on the rate of the reaction.

Knowledge on specific activity, Michaelis constants, or reaction rate data under some particular conditions, introduces additional constraints in the maximization. With more constraints, the maximum rate can be reduced further and other parameter bounds accordingly tightened.

8.8.2.1. Constraints on the Maximum Enzyme Turnover

As an illustration, consider the example two-reactant two-product ordered mechanism, which was examined in detail in Sections 8.2 and 8.3. The maximum enzyme turnover, $V_{\text{max}}$, of the enzyme can be derived from the general rate equations (8.7) and (8.8), by setting $P=0$, $Q=0$, $A \rightarrow \infty$, and $B \rightarrow \infty$:

$$
V_{\text{max}} = \frac{r}{e_{\text{total}}} = \frac{k_3 k_4}{k_3 + k_4}
$$

(8.74)
Any equality or inequality constraint on $V_{\text{max}}$ yields, through this equation, an additional constraint involving the kinetic parameters. The constraint can be used in the maximization to yield tighter bounds.

### 8.8.2.2. Constraints on the Michaelis Constant

To incorporate information on the Michaelis constant of reactant B, assume that $P = 0$, $Q = 0$, and $A \rightarrow \infty$, to obtain from the rate equations (8.6) and (8.7):

$$
\frac{r}{e_{\text{total}}} = \frac{k_1 k_2 k_3 k_4 B}{(k_3 + k_4) k_1 k_2 B + (k_{-2} + k_3) k_1 k_4}
$$

(8.75)

From Equation (8.75) it follows that $V_{\text{max}}$ is given by Equation (8.74), and the Michaelis constant for B, $K_m$, by the equation:

$$
K_m = \frac{(k_{-2} + k_3)k_4}{(k_3 + k_4)k_2}
$$

(8.76)

Thus, as with $V_{\text{max}}$, it is possible to express constraints on $K_m$ as constraints among the kinetic parameters, and use them in the optimization.

### 8.8.2.3. Other Constraints

Suppose, that the actual rate is known under some particular concentrations of metabolites, and a maximum rate estimate is desired for some new concentrations. The rate datum is translated into a constraint among kinetic parameters through direct substitution of the related concentrations and the rate datum in Equations (8.7) and (8.8). Then the constrained maximization can be performed using the new concentrations.
The introduction of any additional constraint, especially one stemming from a rate datum, further complicates analytical maximization, and may necessitate numerical application of the whole method. It is nevertheless useful, when the data are available, and bounds significantly tighter than yielded by the original method are sought.
8.9. DISCUSSION

The results show that the rate of an enzymatic reaction reaches a maximum when kinetic parameters for bimolecular steps approach their collision limits, while kinetic parameters for unimolecular steps take either extreme values (such as $+\infty$) or intermediate optimal values.

The rate constraints derived by this methodology are always stronger than the collision limits previously known. On the other hand, they are expected to be weaker than constraints derived from actual kinetic data.

8.9.1. SIMPLIFICATION OF THE RESULTS

The complicated algebraic results produced by the methodology can be simplified by employing dimensionless parameters, and examining physically meaningful limiting cases.

One interesting limiting case, mentioned in the numerical example (Section 8.7.2.3), is the assumption of irreversibility for a reaction. This assumption leads to drastically simplified analytical results, like Equation (8.73), for any ordered reaction mechanism.

8.9.2. HANDLING RANDOM-ORDER MECHANISMS

The restriction of the results to ordered mechanisms is not severe in practical applications.

For some reaction classes the order is known, as is the case with dehydrogenases, for which the coenzyme binds before the other substrate [Dixon and Webb, 1979].
Otherwise, the mathematical and numerical analysis of random-order mechanisms can be carried out similarly to ordered mechanisms, except that the algebraic expressions would be much more complicated.

A simpler solution is to assume the random-order maximum to be equal to the maximum among all possible ordered mechanisms. Since the binding of each substrate changes the conformation of the enzyme, even formally random-order mechanisms should have a preferred order that accounts for most of the activity of the enzyme. For the purposes of a rough maximum-rate analysis, the preferred mechanism can be assumed as fully representative of the order of binding, i.e., accounting for all of the observable reaction rate.

The examples we have examined certainly indicate that the binding-order does not significantly affect the maximum rate. However, it is important to examine all ordered mechanisms when additional constraints stemming from $V_{\text{max}}$, Michaelis constants, or other data are exploited in the maximization, because the additional data may be very sensitive to binding order.

8.9.3. RANGES FOR OTHER PARAMETERS

As discussed previously (Section 8.8.1 and numerically in the example of Section 8.7.1), the outlined maximum-rate methodology is not just a method for estimating a rate; it can also be used in the reverse direction.

Given a value or design specification on the rate of reaction, the method can be used to estimate a range for another parameter affecting the rate of the biochemical reaction. This includes parameters like metabolite concentrations or the equilibrium constant of the reaction.
8.9.4. APPLICATIONS

Applying the methodology, biochemists can check whether a proposed pathway or reaction mechanism is consistent with the observed rates, while biochemical engineers can check whether a desired process performance is feasible.

The methodology can be useful, for example, in the identification of kinetic bottlenecks (i.e., steps whose maximum rate cannot keep up with the rest of the pathway), or rate-limiting steps of biochemical pathways. An assumption that a step is rate-limiting provides concentration ranges for intermediates in the pathway. If, under these concentrations, the maximum rate estimate is below the observed rate, the candidate rate-limiting step must be rejected.

8.9.5. MULTI-ENZYME COMPLEXES

When the enzyme catalyzing the reaction is part of a multi-enzyme complex, the effective concentrations of reactants and products may differ drastically from their bulk concentrations in the cell. The substrate and intermediate molecules may be transferred, within the complex, from one enzyme to the other, without visiting the bulk of the solution.

Thus, the methodology cannot be applied to an enzyme belonging to a multi-enzyme complex. However, it can be applied to the complex as a whole, for which the assumptions of the methodology hold.

8.9.6. ACTIVE CONCENTRATIONS OF CURRENCY METABOLITES

Another problem where this methodology is pertinent to is the determination of the active concentration of a "currency" metabolite that participates in several
bioreactions. Under some conditions a significant fraction of the total metabolite may be bound to enzymes, leaving only the remainder as reactive concentration.

It would be, thus, particularly interesting to employ the maximum rate methodology to determine the minimum complexed (inactivated) metabolite concentration for given \( r, e_{\text{total}} \), and \( S_i \) for a large set of reactions that involve the metabolite. Then one could estimate the remaining active concentration and compare it to the total concentration.

8.9.7. METABOLIC EFFICIENCY AND EVOLUTION OF ENZYMES

It would be very useful to examine the metabolic efficiency of enzymes, that is, their achieved rate divided by the predicted maximum. Based on their efficiency, enzymes could be classified as almost perfected when their efficiency is comparable to 1 (e.g., larger than 0.01 or 0.05), or imperfect when the efficiency is much smaller than 1.

The level of the efficiency of an enzyme is expected to relate to two important factors:

(a) The intrinsic difficulty of the reaction. Some reactions are so hard that the internal chemical barriers always overshadow physical collision factors.

(b) The metabolic importance of the bioreaction. There is stronger evolutionary pressure on enzymes that are more critical for the overall efficiency of a microorganism. Those enzymes are more likely to be almost perfected.
8.9.8. CONCLUSIONS

The novel methodology that was presented allows the estimation of a maximum rate for an enzymatic reaction, provided that the equilibrium constant of the reaction, the concentration of the enzyme, and the concentrations of the reactants and products are all known.

The method can estimate an extremum for another parameter if the actual rate of the reaction is known. The methodology is based solely on physicochemical considerations and can be applied in the absence of kinetic data, or exploit partial kinetic data to produce tighter bounds.

The assumption of ordered mechanisms can be overcome by assuming the fastest binding order as predominant, or through numerical optimization on the random-order mechanism.

The main difficulties in the application of the methodology stem from the complexity of the algebraic results. This can be partly alleviated through the investigation of limiting cases, such as irreversible reactions, near-equilibrium reactions, and initial rates.

Investigation of potential biological regularities involving enzyme efficiency may provide useful insights on the intrinsic reaction difficulty and the evolutionary perfection of enzymes.
8.10. SUMMARY

In this chapter, a general methodology that allows the estimation of maximum rates for enzymatic reactions was described.

Starting with a typical fast mechanism for an enzymatic reaction, the rate is expressed in terms of the kinetic parameters. The kinetic parameters are require to obey certain constraints, which dictate that the ratio of forward and backward rate constants be equal the equilibrium constant, and that the rate of any bimolecular-reaction steps be less than the rate of collision of the reactant species. Other information on the reaction rate can also be expressed in terms of constraints. By maximizing the rate with respect to the kinetic parameters, subject to all applicable constraints, an upper bound is obtained for the rate of the reaction, based on first principles.

If the reaction rate is actually known, the methodology can alternatively estimate an extremum for the concentration of the enzyme, a substrate, or a product. A nondimensionalization which facilitates the algebraic manipulation was outlined, and asymptotic behavior with respect to proximity to equilibrium was examined.

Ranges were estimated for the collision parameters, which are necessary for the application of the method in practice. Potential extensions of the methodology were discussed. Applications of the method include the identification and bypassing of kinetic bottlenecks (Chapter 10).
CHAPTER 9

SYNTHESIS OF BIOCHEMICAL PATHWAYS

SATISFYING STOICHIOMETRIC SPECIFICATIONS
9.1. INTRODUCTION

9.1.1. VARIANTS OF THE SYNTHESIS PROBLEM

In general, the synthesis of biochemical pathways involves the construction of pathways (i.e., sets of enzyme-catalyzed bioreactions) that meet certain specifications. There are three, essentially equivalent variants in the formulation of the synthesis problem:

- In the more synthetic variant we start with no pathway at all and aim to synthesize whole pathways that meet a set of specifications.

- In pathway improvement, we start with a complete existing pathway and want to improve it by eliminating a section of the pathway that contains a bottleneck.

- Finally, we may start with a partial pathway which we want to complete.

We will discuss, in the paragraphs below, these variants of the problem, and we will argue that they are essentially equivalent.

9.1.1.1. Base Problem: Synthesis of Whole Pathways

In the very initial steps of the design of a biochemical process, we are faced with a synthesis problem. We want to produce certain target bioproducts, under partial constraints on the available substrates (reactants), allowed by-products, desired yield, productivity, etc.
Given such specifications, we would like to systematically identify all pathways that can achieve the transformation of the available substrates to the desired products, subject to the other imposed constraints. In this formulation of the synthesis problem, it should be noted that:

- A pathway must include all necessary reactions to convert initial substrates that the microorganism consumes to final products that will be recovered from the fermentation broth. It is not sufficient to give a pathway leading from the intermediary metabolism to the product.

- The pathways sought are not restricted to those commonly considered as the pathways the microorganisms use to produce the product for their own needs. In fact, we are particularly interested in non-obvious alternatives including pathways that are not present in any single microorganism.

- The complexity and density of the intermediary metabolism of the cell drastically increase the number of possibilities, even if the final section of a pathway, leading from the intermediary metabolism to the final product, can be uniquely identified.

9.1.1.2. Modifications to Existing Pathways

9.1.1.2.1. Bottlenecks in Pathways

Consider a given pathway, in which a bottleneck has been identified through the use of analytical methodologies. The bottleneck could be:
• A thermodynamic infeasibility caused by an unfavorable equilibrium constant (estimated by the group-contribution method of Chapter 7).

• A kinetic infeasibility caused by a slow enzyme, determined through either actual enzyme kinetics or maximum rates (Chapter 8).

• A regulatory infeasibility because of regulation which is, because of experimental difficulties, hard to modify in practice. The regulation might be at any level:
  ◊ Regulation of gene expression
  ◊ Post-transcriptional modification
  ◊ Allosteric control
  ◊ Inhibition by substrate or product analogues

9.1.1.2.2. Removal of Bottlenecks

Depending on the nature of a bottleneck and the experimental context, one of two courses of action can be followed:

(a) One can replace the enzyme with an enzyme from a different source

(b) The structure of the pathway can be altered by replacing a whole bioreaction or even a set of bioreactions (i.e., a segment of the pathway around the bottleneck)
In case (b) we are clearly faced with another synthesis problem. Since we would like to keep the well-functioning segments of the pathway intact, we only want to synthesize a bypass to the pathway's bottleneck.

9.1.1.2.3. Equivalence to Base Problem

This synthesis problem is exactly the same as the problem of synthesizing whole pathways.

The only difference is that in this case the designated reactants and products while not be true substrates and final products of fermentations, but rather intermediates of an existing biochemical pathway.

Another way of looking at it is that we are synthesizing pathways that achieve the same transformation as the initial pathway but exclude the bottleneck-bioreaction: from the synthesized pathways we can then select only those pathways that share a large part of their structure with the original pathway.

9.1.1.3. Completion of Partial Pathways

The third variant of the synthesis problem arises if we know what pathway to use for part of the desired transformation, and we want to augment the known partial pathway so that the whole transformation is achieved. This commonly arises in two contexts:

- The known pathway consumes (and produces) metabolites which were perceived as currency metabolites at some stage, but we now want to include in the pathway appropriate bioreactions that will
eliminate these currency metabolites from the overall, net stoichiometry of the pathway.

- The known pathway starts from the intermediary metabolism and proceeds to the final product. We want to bridge the gap between the available (specified) substrates and the appropriate metabolic intermediates, which serve as net reactants of the known pathway.

This variant of the problem is only on the surface different from the previous formulations. It is reducible to the problem of bypassing bottlenecks (discussed in Section 9.1.1.2, above) if we imagine that there was a bottleneck-reaction, with the correct stoichiometry that completed the pathway, and we now want to eliminate that reaction while retaining the rest of initial pathway intact.

9.1.2. NATURE OF THE SPECIFICATIONS AND CONSTRAINTS

In Sections 3.2.2 and 4.4.2 (and especially Section 4.4.2.1) we discussed specifications that may be imposed on the pathways to be synthesized. These specifications have many different forms, and refer to different attributes of the pathway or the process. Here we will focus only on those specifications which can be used, directly or indirectly, by the algorithmic synthesis methodology that is the subject of this chapter.
9.1.2.1. Stoichiometric Specifications

A whole class of specifications can be formulated by classifying each building block (each metabolite and each bioreaction) from the database according to the role it is required or allowed to play in the pathways to be synthesized.

The fact that stoichiometric specifications are linear is a key to the success of the synthesis algorithm. We will examine each of four roles (reactants, products, intermediates, and bioreactions) of building blocks below, describing the specifications stemming from it.

9.1.2.1.1. Reactants

Each metabolite from the database can be classified according to whether it is allowed to be a net reactant or substrate (we use the terms substrate and reactant interchangeably) of the pathway:

(1) Required reactants (or desired reactants) must be consumed by the pathway

(2) Allowed reactants may or may not be consumed by the pathway

(3) Excluded reactants (or prohibited reactants) must not be consumed by the pathway

Clearly, only specifications (1) and (3) are true constraints. Specification (2), on the other hand, is not a true constraint because it does not restrict the set of acceptable pathways. However, the default characterization of this role for each
metabolite is specification (3): The bulk of metabolites in the database are not allowed to be overall reactants of the pathway.

Thus, there is a very large number of constraints which must be taken into account by the synthesis algorithm.

9.1.2.1.2. Products

Each metabolite can be similarly classified according to whether it is allowed to be a net product of the pathway:

1. **Required products** (or desired products) must be produced by the pathway

2. **Allowed products** may or may not be produced by the pathway

3. **Excluded products** (or prohibited products) must not be produced by the pathway

These specifications are analogous to the specifications on reactants (Section 9.1.2.1.1, above). As with reactants, specification (2) is not a true constraint. The default characterization is specification (3), introducing more constraints to be taken into account.

9.1.2.1.3. Intermediates

Each metabolite can be classified according to whether it is allowed to be an intermediate of the pathway:
(1) **Required intermediates** (or desired intermediates) *must* participate in the pathway

(2) **Allowed intermediates** may or may not participate in the pathway

(3) **Excluded intermediates** (or prohibited intermediates) *must not* participate in the pathway

As with reactants and products (Sections 9.1.2.1.1 and 9.1.2.1.2), specifications (1) and (3) are true constraints, while specification (2) is not restrictive. Contrary to reactants and products, the default characterization of this role for each metabolite is specification (2): The bulk of metabolites are, in fact, *allowed* to be internal intermediates of the pathway.

Thus, this role will not, in general, introduce many constraints.

9.1.2.1.4. **Bioreactions**

In a manner similar to the classification of intermediates (Section 9.1.2.1.3, above) each bioreaction from the database can be classified according to whether it is allowed to participate in the pathway:

(1) **Required bioreactions** (or desired bioreactions) *must* participate in the pathway

(2) **Allowed bioreactions** may or may not participate in the pathway

(3) **Excluded bioreactions** (or prohibited bioreactions) *must not* participate in the pathway
These specifications are analogous to the specifications on intermediates (Section 9.1.2.1.3). As with intermediates, specification (2), which not a true restrictive constraint, and it is the default characterization.

Since bioreactions are in general reversible, each reaction can receive a designation for each of its two possible directions. The two designations are not completely independent. For example, it is not meaningful to specify a reaction as required in both the forward and reverse direction; a reaction that is required in one direction must be excluded in the other direction. Thus, there are in total 5 possible designations (out of a total of 3x3=9 simple-minded combinations) for a reversible reaction:

- Allowed in both directions; this is the only specification introducing no restrictions
- Required in the forward direction (and excluded in the backward direction)
- Required in the backward direction (and excluded in the forward direction)
- Allowed in the forward direction (and excluded in the backward direction)
- Allowed in the backward direction (and excluded in the forward direction)

9.1.2.2. Thermodynamic Specifications

One would presumably like any synthesized pathways to contain only thermodynamically feasible bioreactions.
To this end, one can introduce some kind of constraint on the equilibrium constants (or the Gibbs Energies) of the bioreactions that can be used; one possible constraint is a simple upper bound on the Gibbs Energy (or a lower bound on the equilibrium constant). Note that the same kind of constraint ought to apply to both the forward and reverse (backward) direction of the reaction, as the nominal direction of a reaction in the database is often arbitrary. Reversible reactions can be used in pathways in either direction.

The default we use for thermodynamic feasibility is not as clearly defined as the defaults of stoichiometric specifications. We require that the maximum of all known values of the equilibrium constant of the reaction be larger than the minimum possible mass-action ratio, under physiologically acceptable concentrations of the reactants and products. From various literature sources [Ingraham et al., 1983, Lehninger, 1982], it appears that concentrations under physiological conditions are generally in the range between $5 \times 10^{-6}$ M and $5 \times 10^{-3}$ M.

9.1.2.3. **Handling Specifications**

Thermodynamic specifications can be handled right from the start by applying the supplied criterion to determine what reactions are feasible in the forward and backward directions. The infeasible reactions are eliminated from consideration. For reversible reactions some additional processing is necessary, as we will discuss in Section 9.3.5.

The main body of the algorithm that will be presented in Section 9.2 is devoted to the satisfaction of stoichiometric requirements.
9.1.3. PREVIOUS WORK

There has been only one previous effort to devise a methodology for the synthesis of biochemical pathways [Seressiolis and Bailey, 1988]. There are the following problems with that approach:

- The approach is computational rather than algorithmic. It involves a computer program which synthesizes pathways, but does not supply a clearly defined algorithm. From the vague description of the methodology, it appears that many modifications had to be made to the program to make it produce reasonable pathways. This may be the reason that the program was not even decoded into an algorithm at the end.

- The formulation of the problem is very restrictive. It involves the specification of only one required reactant and one required product and no further refinement in terms of allowed reactants and by-products.

- While all pathways constructed must involve the required reactant and the required product, there is no other declarative description of the set of pathways that are synthesized. They are only vaguely defined in terms of various procedural issues arising in the program.

Given these deficiencies of the previous work, we proceeded to devise a new methodology for the synthesis of biochemical reaction pathways, overcoming the restrictions of previous efforts.
In the next sections we will further formalize the synthesis problem, give all the
details of the algorithm, and discuss the algorithm's efficient use through simple
examples.

More interesting applications are provided in the next chapter, which discusses
at length an example, applying the synthesis algorithm along with other methodologies
from this work. The case study (Chapter 10) deals with the production of lysine from
glucose.
9.2. PROBLEM FORMALIZATION

Let \( R_i \) \((i=1, \ldots, m)\) be the bioreactions available in the database, and \( S_k \) \((k=1, \ldots, n)\) the metabolites which the reactions interconvert. Let \( c_{ki} \) \((i=1, \ldots, m\) and \( k=1, \ldots, n)\) be the stoichiometric coefficient of the \(i\)-th reaction with respect to the \(k\)-th metabolite (molecule).

9.2.1. STOICHIOMETRIES OF A PATHWAY

A biochemical pathway \( P \) is partly defined by the coefficients \( a_i \) of the bioreactions as used by the pathway, which we will call the reaction stoichiometry description, and partly by the structural arrangement of the bioreactions, which we will call connectivity or structural description.

The main issue in the connectivity description is the designation of main reactants and main products for each bioreaction (Section 4.4.1.2). We will not use this kind of description in our synthesis methodology.

Let \( b_k \) be determined as:

\[
b_k = \sum_{i,k} (c_{ki} a_i)
\]  \( (9.1) \)

or in vector notation

\[
B = CA
\]  \( (9.2) \)

The vector \( B \) defines the molecular stoichiometry of the pathway. It describes the net stoichiometry, i.e., the net reactants and products of the pathway. Let the vector \( T \) describe the intermediate stoichiometry of the pathway, with its elements \( t_k \) defined as:
\[ t_k = \left( \sum_{l,k} (c_{kl} a_l)^2 \right)^{1/2} \]  

(9.3)

We are mainly interested in the sign of \( t_k \), which is positive or zero so other definitions of \( t_k \) are also possible and some simpler one can be adopted. In general, \( t_k \) can be defined as any norm of the vector of Equation (9.3) and not necessarily the Euclidian norm that we used.

### 9.2.2. FORMALIZATION OF STOICHIOMETRIC SPECIFICATIONS

Constraints on the stoichiometry (discussed in Section 9.1.2) can now translated into formal requirements on the three stoichiometries we defined. In the formalization below, we will not take into account the redundancies and interactions which arise from the fact that \( b_k \neq 0 \) yields \( t_k > 0 \).

#### 9.2.2.1. Constraints on the Molecular Stoichiometry

Here, we cover the specifications of Sections 9.1.2.1.1 and 9.1.2.1.2. In the general case \( b_k \) is not restricted at all. The restrictions that may arise from the specifications are:

- \( S_k \) must be used as a substrate (reactant): \( b_k < 0 \)
- \( S_k \) must be produced as a product: \( b_k > 0 \)
- \( S_k \) may not be used as a substrate: \( b_k \geq 0 \)
- \( S_k \) may not be used as a product: \( b_k \leq 0 \)
- \( S_k \) must be neither a substrate nor a product: \( b_k = 0 \)
9.2.2.2. Constraints on the Intermediate Stoichiometry

Here, we cover the specifications of Section 9.1.2.1.3. In the general case, i.e., when no restrictions are present, \( t_k \geq 0 \). The restrictions that may arise from the specifications are:

- \( S_k \) must occur in the pathway (even if only as an intermediate): \( t_k > 0 \)

- \( S_k \) must not occur at all in the pathway: \( t_k = 0 \)

9.2.2.3. Constraints on the Reaction Stoichiometry

Here, we cover the specifications of Section 9.1.2.1.4. In the general case there is no restriction on the value of \( a_k \). The restrictions that may arise from the specifications are:

- \( R_k \) must occur in the pathway, in its forward direction: \( a_k > 0 \)

- \( R_k \) must occur in the pathway, in its backward direction: \( a_k < 0 \)

- \( R_k \) must not occur in its forward direction: \( a_k \leq 0 \)

- \( R_k \) must not occur in its backward direction: \( a_k \geq 0 \)

- \( R_k \) must not occur at all in the pathway: \( a_k = 0 \)
9.3. ALGORITHM FOR THE SYNTHESIS OF BIOCHEMICAL REACTION PATHWAYS

We will present here an algorithm which, given a set of stoichiometric requirements and a database of biochemical reactions (i.e., reaction stoichiometries) synthesizes all biochemical pathways that satisfy the stoichiometric requirements and certain additional constraints.

9.3.1. INTRODUCTION OF ADDITIONAL CONSTRAINTS

9.3.1.1. Pathway Properties and Relations

In this section, we define some properties of pathways (and relations among pathways), which we then use to introduce some additional constraints into the pathway synthesis problem. The properties and relations will also be useful in the detailed description and proof of the algorithm.

The length of a pathway is simply the number of reactions which participate in the pathway (with coefficients different from zero).

A combination of a set of constituent pathways is a pathway whose stoichiometries are linear combinations of the stoichiometries of the constituent pathways, with positive combination coefficients.

Let P and Q be pathways derived from the same reaction database. We will call P a subpathway of Q if, and only if, every reaction that has a non-zero coefficient in the stoichiometry of Q has a coefficient with the same sign in the stoichiometry of P. We will equivalently call Q a superpathway of P.
A pathway $P$ is a strict subpathway of $Q$ if, and only if, $P$ is a subpathway of $Q$ and the length of $P$ is smaller than the length of $Q$.

If $Q$ is a superpathway of $P$, then there is another pathway, $W$, such that the pathway combination of $W$ and $P$ yields $Q$, while $W$ is a subpathway of $Q$. We say, then, that $Q$ can be partitioned into the subpathways $P$ and $W$; we call $W$ the complement of $P$ with respect to $Q$.

Let $Q$ be a pathway satisfying a set of requirements (which are not restricted to the requirements discussed in this section). We will call $Q$ a minimal pathway, with respect to the requirements, if there is no strict subpathway of $Q$ satisfying the requirements. In effect, over the space of reactions that $Q$ contains, $Q$ is the pathway with the smallest length that will satisfy the requirements.

### 9.3.1.2. Additional Constraints

While the algorithm can synthesize pathways of arbitrary length, we will restrict the algorithm to considering only pathways of size less than some arbitrary limit. This limit becomes an additional part of the formal description of the problem. Thus, we have introduced an extra constraint (apart from the specifications mentioned in Section 9.2):

- The pathway must have size less than or equal to $w$: At most $w$ of its reaction-stoichiometry coefficients are allowed to be different from zero.

The motivation for this restriction is to maintain the computational complexity of the algorithm manageable.
In practice, it is not difficult to place an upper bound on the size allowed for a pathway. More importantly, this upper bound is independent of the size of the database: Having information about more and more reactions increases the size of the search space and possibly the number of solutions, but there is no reason to expect each solution to be longer.

There is also a genetic engineering consideration here: If the pathway consists of too many steps, it becomes impractical to implement the pathway, through genetic engineering techniques, into a microorganism.

While we are interested in all pathways satisfying the specifications (and there is, in general, an infinite number of them), we would like to have the pathways described in terms of a set of minimal pathways. In other words, if pathway $P$ satisfies the constraints posed (in their strict-inequality or loose-inequality form, as will be discussed later) but is not minimal, we do not want $P$ to be directly produced by the algorithm. Instead, we want its constituent pathways that also satisfy the constraints to be produced directly; $P$ will be known indirectly, because combinations of acceptable pathways are also acceptable.

This point will be clarified further in the description of the metabolite-processing and pathway-marking phases of the algorithm. At this stage it is sufficient to note that we do not want non-minimal pathways to be directly produced by the algorithm.

9.3.2. OVERVIEW OF THE ALGORITHM

The algorithm is based on the iterative satisfaction of requirements, and the transformation of the initial set of reactions (which can be thought of as one-step
pathways) into a final set of pathways which satisfy all requirements. The algorithm consists of three phases:

- A preliminary reaction-processing phase, in which:
  - The reversibility of reactions is introduced through the creation of reverse reactions.
  - Excluded reactions are removed.

- An intermediate metabolite-processing phase. For each metabolite, constraints pertaining to the metabolite's use (as a reactant, product, or intermediate) as allowed or excluded are satisfied through the iterative transformation of a current set of pathways:
  - Pairs of pathways, with one pathway producing the metabolite and the other consuming it, may be linearly combined to eliminate the metabolite from the molecular stoichiometry.
  - Pathways may be removed from the current set if they violate the constraint at hand.

- A final pathway-marking phase, in which:
  - The final set of pathways is marked to indicate which of the constraints pertaining to required reactants, products, intermediates, or reactions are satisfied by the pathway.
  - The existence of a solution is established.
  - Composite pathways that are correct answers to the synthesis problem can be formed.
9.3.3. REACTION-PROCESSING PHASE

This is a preliminary phase of the synthesis algorithm, whose main aims are:

(1) To account for the reversibility of biochemical reactions

(2) To satisfy some of the constraints on the reaction stoichiometry, namely those prescribing excluded reactions

9.3.3.1. Reaction Reversibility

In order to account for the reversibility of reactions, we create new reactions, which are the inverses of those reactions from the original set that are in principle (i.e., thermodynamically) reversible. In effect, we decompose reversible reactions into their forward and backward portions.

With the new, expanded, reaction set, we require that the reaction-stoichiometry coefficients be non-negative. We transform the constraints placed on each of the original reactions into a constraint on that reaction and a constraint on its inverse (both of which belong to the new reaction set), as shown in Table 9.1.

From Table 9.1 it is evident that, in the new reaction set, there are only three possible prescriptions on reaction stoichiometry (as opposed to six for the initial set shown in Section 9.2.2.3):

- $R_k$ may occur in the pathway (which is the general case, indicating the absence of restrictions): $t_k \geq 0$
- $R_k$ must occur in the pathway: $t_k > 0$
- $R_k$ must not occur in the pathway: $t_k = 0$
Table 9.1:
Transformation of the stoichiometric constraint placed on a reversible reaction, into stoichiometric constraints on the forward and backward portions into which the reaction is decomposed

<table>
<thead>
<tr>
<th>Initial Constraint</th>
<th>Transformed Constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_k \geq 0 )</td>
<td>( a_k \geq 0 )</td>
</tr>
<tr>
<td>( a_k = 0 )</td>
<td>( a_k = 0 )</td>
</tr>
<tr>
<td>( a_k \geq 0 )</td>
<td>( a_k \geq 0 )</td>
</tr>
<tr>
<td>( a_k &gt; 0 )</td>
<td>( a_k &gt; 0 )</td>
</tr>
<tr>
<td>( a_k \leq 0 )</td>
<td>( a_k = 0 )</td>
</tr>
<tr>
<td>( a_k &gt; 0 )</td>
<td>( a_k = 0 )</td>
</tr>
</tbody>
</table>
9.3.3.2. Excluded Reactions

From the constraints of the previous section, the last category (dictating that certain reactions be excluded from the constructed pathways) can be satisfied right from the start. We simply eliminate such reactions from consideration by removing them from the set of available reactions.

In the computer implementation, it is sufficient to eliminate the corresponding backward-pointers, which point from metabolites to reactions in which they participate. These pointers are, as will be described below, part of the state of the design. Once a reaction is not accessible from any metabolite, it will never be used in the construction of pathways.

The constraints referring to required reactions (corresponding to a strict inequality for the corresponding reaction stoichiometry coefficient) cannot be processed at this preliminary stage. We will see that their satisfaction is achieved in the last, pathway-marking phase.

9.3.4. METABOLITE-PROCESSING PHASE

This is the main phase of the algorithm and will involve construction of pathways to satisfy constraints on molecular stoichiometry.

9.3.4.1. State of the Design

At each stage in the synthesis of algorithm, the problem state (or the state of the design [Mostow, 1983, Mostow, 1984]) is characterized by a set of composite reactions (i.e., partial pathways) from which we can construct, as linear combinations, pathways
to satisfy the remaining requirements. Each partial solution, then, consists of the following elements:

- The set of constraints (on the stoichiometry) we have yet to satisfy. As each constraint corresponds to a particular metabolite, this is the same as the set of metabolites that still remain to be processed.

- The set of active, incomplete pathways constructed so far. These are pathways that satisfy the already-processed constraints. The initial set of reactions is the starting set of (one-step) pathways, when no constraints have been processed yet. The starting set is transformed in subsequent steps of the design.

- Back-pointers which show, for each remaining metabolite, the pathways in which it participates. These data-structures must be initially created by passing over each of the initial one-step pathways and updating the metabolites participating in the pathway's molecular stoichiometry.

- Other bookkeeping information, detailing (for each of the pathways constructed as partial solutions) the evolution from the initial problem to the current state, and interrelationships among partial solutions.

### 9.3.4.2. Pathway Expansion

At each pathway-expansion step, one requirement from the set of remaining requirements is chosen as a goal, and the system finds a modification of the set of pathways such that all surviving pathways satisfy the requirement. This involves the
construction of new pathways, as linear combinations of existing ones, as well as deletion of pathways from the working set.

More specifically, for the metabolite being processed and using the backward-pointers readily available in each metabolite, we assemble two subsets of the current pathway set:

- The list of pathways that produce the metabolite.
- The list of pathways that consume the metabolite.

9.3.4.2.1. Pathways Constructed and Deleted

The pathways that may, at this step of the algorithm, be constructed as linear combinations will be combinations of exactly one pathway from the first list (above) and exactly one pathway from the second list. The pathways that may be deleted from the current set will be all pathways from either (or both lists), depending on the exact nature of the constraint.

We will delineate the procedure for each kind of constraint below:

• Molecular Stoichiometry Constraints
  ◊ In the general case $b_k$ is not restricted at all and $S_k$ may be a reactant or product of the pathways. No processing is necessary and the set of pathways is carried intact to the next iteration.
  ◊ $S_k$ must be used as a substrate (reactant): $b_k < 0$. All combination pathways are constructed. The producing pathways are deleted. This constraint will receive additional attention in the pathway-marking phase.
◊ $S_k$ must be produced as a product: $b_k > 0$. All combination pathways are constructed. The consuming pathways are deleted. This constraint will receive additional attention in the pathway-marking phase.

◊ $S_k$ may not be used as a substrate: $b_k \geq 0$. All combination pathways are constructed. The consuming pathways are deleted.

◊ $S_k$ may not be used as a product: $b_k \leq 0$. All combination pathways are constructed. The producing pathways are deleted.

◊ $S_k$ must be neither a substrate nor a product: $b_k = 0$. All combination pathways are constructed. All producing and consuming pathways are then deleted.

• INTERMEDIATE STOICHIOMETRY

◊ In the general case $t_k \geq 0$, and the metabolite may participate freely in the pathway. No processing is necessary and the set of pathways is carried intact to the next iteration.

◊ $S_k$ must occur in the pathway (even if only as an intermediate), i.e., $t_k > 0$. No processing is necessary and the set of pathways is carried intact to the next iteration. This constraint will receive additional attention in the pathway-marking phase.

◊ $S_k$ must not occur at all in the pathway, i.e., $t_k = 0$. No combinations of pathways are constructed. All producing and consuming pathways are deleted.
9.3.4.2.2. Requirements that Remain

After the processing of the constraint, we have a new set of pathways, which almost satisfy the constraint. The qualification stems from the fact that, for strict-inequality constraints, i.e., $b_k > 0$ (required product), $b_k < 0$ (required reactant), and $t_k > 0$ (required intermediate), only the corresponding equality constraints are guaranteed to be satisfied, i.e., the constraints $b_k \geq 0$ (allowed product but excluded reactant), $b_k \leq 0$ (allowed reactant but excluded product), and $t_k \geq 0$ (allowed intermediate) respectively.

The last phase of the algorithm (pathway-processing, described in Section 9.3.5) serves exactly to enforce the satisfaction of the strict inequality constraints.

9.3.4.2.3. Redundancy Checks

Because of the requirement that pathways directly produced by the algorithm be minimal, new pathways are not directly inserted in the current set. We must first perform two checks:

- If the pathway is redundant, i.e., If the pathway (or a linear multiple of it, with positive coefficients) is already present in the current set, then it is not re-inserted.

- If the pathway is not minimal with respect to loose-inequality constraints discussed above, then the pathway is discarded.
9.3.4.3. Updating the State of the Design

There is significant bookkeeping information that must be updated with each removal or addition of a pathway from the current set, outlined in the following paragraphs.

9.3.4.3.1. Deletion of an Existing Pathway

Deletion of a pathway from the active pathway set entails:

◊ Removal of pointers which point to the pathway from metabolites (which participate in the molecular stoichiometry of the pathway).

◊ Construction of a record explaining what constraint (metabolite) dictated the abolition of this pathway.

◊ Removal of the pathway from the set of active pathways (but only after it has been used in the construction of combinations, when that is appropriate).

9.3.4.3.1. Addition of a New Pathway

Introduction of a pathway into the set of active pathways entails:

◊ Computation of the molecular and reaction stoichiometries for the pathway

◊ Check of redundancy and minimality.
Construction of data-structures detailing the constituent pathways and the constraint (metabolite) which required the creation of this combination.

Installation of the backward-pointers in the metabolites that appear in the molecular stoichiometry of the new pathway, in order to make the pathway accessible and usable in the construction of later combinations.

Inclusion of the new pathway in the set of active pathways.

9.3.5. PATHWAY-MARKING PHASE

At the end of the metabolite-processing phase, we have a set of pathways satisfying all of our requirements, except for the strict-inequality constraints, for which only the corresponding loose inequalities are satisfied. Because of the linear nature of the requirements, all linear combinations of pathways also satisfy the constraints, in the same loose form.

9.3.5.1. Satisfaction of Strict-Inequality Constraints

Some pathways from the final set also satisfy the set of original strict-inequality constraints, or a subset of that set. Combinations of pathways (with non-negative coefficients, as defined in Section 9.3.1.1) satisfy the union of the constraints satisfied by their constituent pathways.

Thus, if we mark each pathway with the strict-inequality constraints it satisfies, we can obtain the final answer to the synthesis problem:
• The pathways satisfying the original stoichiometric constraints of the synthesis problem are all those linear combinations (with non-negative coefficients) of pathways from the final set which have at least one constituent pathway satisfying each of the strict-inequality constraints (i.e., the constraints referring to required reactants, products, intermediates, or reactions). In other words, we must include in the combination:

◊ at least one pathway consuming each required reactant,

◊ at least one pathway producing each required product, and

◊ at least one pathway containing each required intermediate.

◊ at least one pathway in which each required reaction participates.

Naturally, a single constituent pathway may possess many of the strict-inequality constraints and can serve to satisfy many of the above requirements. Thus, if a pathway satisfies all of the strict inequalities, then that pathway itself is acceptable as one solution to the overall synthesis problem; any combination of that pathway with other pathways from the final set is also acceptable.

If there is a strict-inequality requirement which is not satisfied by any of the pathways in the final set, then there is no solution to the original synthesis problem. If each strict-inequality requirement is satisfied by at least one pathway in the final set, then a feasible pathway exists.
9.3.5.2. **Neutral Pathways**

One may wonder whether there are, in the final set, pathways which do not satisfy any of the strict-inequality requirements, and whether there is any reason to construct or keep such pathways.

There are indeed such pathways, which we will call *neutral*, generated by the algorithm. Since these pathways do not contribute to the satisfaction of any strict-inequality requirements, it is not *necessary* to use them in combinations constructed from the final set, but they may be freely included in such combinations as they neither prevent any requirements from being satisfied nor introduce additional requirements.

The neutral pathways must be constructed, because:

- They are required for the mathematical completeness of the solution. We want a description of *all* pathways satisfying our set of requirements; these include pathways that are formed as combinations, with neutral pathways as some of their constituent pathways.

- In terms of the mechanics of the algorithm, there is no advantage to be gained by omitting neutral pathways. In other words, it is not possible to decide *a priori* whether certain pathways in the current set will only lead to combinations that are neutral pathways. Thus, the efficiency of the algorithm is independent of whether the neutral pathways are kept or not.

It can be easily shown that composite pathways which contain neutral pathways are not minimal.
9.3.6. PROPERTIES OF THE ALGORITHM

In this section, we will first discuss some mathematical properties of the algorithm, showing that the algorithm possesses:

- Correctness; it generates only feasible pathways, i.e., only pathway satisfying the stoichiometric requirements imposed.

- Completeness; it generates (a description of) all pathways satisfying the requirements.

- Complexity polynomial in the size of the database that describes the stoichiometries of the reactions

We will then discuss issues related to the computation performance of the algorithm and the ways in which the algorithm takes advantage of the structure of the problem.

9.3.6.1. Mathematical Properties

9.3.6.1.1. Correctness Theorem

If a linear combination of pathways from the final set (produced by the synthesis algorithm) contains at least one constituent pathway satisfying each of the strict inequality requirements (referring to required reactants, products, intermediates, or reactions), then the combination pathway satisfies all of the initial stoichiometric requirements.
9.3.6.1.2. Correctness Proof

The proof is based on three facts:

- Whenever, in the synthesis algorithm, a new pathway is created, its molecular and reaction stoichiometries are consistent, and consistent with those of its constituent pathways. This is obvious, because all the stoichiometries of a pathway are linearly combinable.

- In the metabolite-processing phase of the algorithm, after the processing of any particular metabolite, the current set of active pathways satisfies the stoichiometric constraints imposed on that metabolite, at least in their loose inequality form.

- As outlined in the pathway-marking phase of the algorithm, a linear combination of pathways from the final set satisfies the union of the strict-inequality requirements satisfied by its constituent pathways.

The last claim is true because those pathways that do not satisfy the strict inequalities satisfy the corresponding loose inequalities. For example, for a required reactant there will be some pathways will have the corresponding stoichiometric coefficient negative, while others will have it equal to zero. In a combination of such pathways, if there is one non-zero (in this case negative) coefficient then the linear combination of the coefficients is negative as well.
9.3.6.1.3. Completeness Theorem

The synthesis algorithm creates a final set of pathways such that: Any pathway satisfying the original stoichiometry constraints is a linear combination of pathways from the final set.

9.3.6.1.4. Completeness Proof

The reaction-processing and pathway-marking phases play no significant role in eliminating pathways. Thus, we focus on the metabolite-processing phase.

At the beginning of the phase, the algorithm has an initial set of (one-step) pathways. Since that set contains all the feasible reactions (unless they have been designated as excluded) any feasible pathway is (by definition) a linear combination of pathways from that set.

We simply need to prove that

If before processing a particular metabolite there exists a pathway that: (a) Satisfies the constraints on the metabolite; and (b) can be constructed as a combination pathway from the current set of active pathways

then, after processing the metabolite, the pathway can still be constructed from the (changed) active set.

This holds because of the way we handle each requirement. Consider, as an example, a metabolite whose constraint is that it may not occur at all in the stoichiometry of the pathway (excluded reactant and excluded product), and all the partial pathways in the initial active set that consume or produce it.
• If a composite pathway does not involve any of the partial pathways that consume or produce the metabolite, then it can be constructed after the processing exactly the same way it was constructed before, since its constituent pathways remain unaffected by the processing of the metabolite.

• If a pathway involves constituent partial pathways that consume or produce the metabolite, then for each of these partial pathways let: \( x_i \) be the coefficient of the metabolite in the partial pathway; and \( y_i \) the (non-negative) coefficient with which the partial pathway participates in the composite pathway at hand. In order for the constraint on the metabolite to be satisfied, its coefficient in the composite pathway must be zero:

\[
\sum_i x_i y_i = 0 \quad \text{(9.4)}
\]

Let \( Y \) be the total consumption and total production of the metabolite:

\[
Y = \sum_i x_i y_i = \sum_{i, (x_i > 0)} x_i y_i + \sum_{i, (x_i < 0)} x_i y_i \quad \text{(9.5)}
\]

Then:

\[
\sum_i x_i y_i = \sum_i \sum_{j, (x_j < 0)} (x_i y_i f_j + x_j y_j f_i) \quad \text{(9.6)}
\]

where:

\[
f_i = x_i y_i / Y \quad \text{(9.7)}
\]

Note that:
\[ x_i y_j f_{ij} + x_j y_i f_{ji} = 0 \]  \hspace{1cm} (9.8)

Once the parameters \( f_i \) are chosen, using Equation (9.7) with the metabolite being processed in mind, Equation (9.6) holds for all other metabolites as well — although Equation (9.8) does not.

Thus, the transformation in Equation (9.6) denotes that the composite pathway can be written as a sum of pairs of constituent partial pathways (with \( f_i \) the coefficients used in the formation of the pairs), such that for each pair the metabolite has zero total coefficient, as Equation (9.8) states. But these pairs are exactly the combinations created by the algorithm. Hence, a composite pathway that satisfied the constraint before the metabolite was processed can still be constructed after the metabolite is processed, using the combination-pair partial pathways created in the processing.

9.3.6.1.5. Polynomial Complexity Theorem

If we restrict the size of the synthesized pathways to a reaction length \( k \), independent of the size of the database, then the outlined algorithm has time-complexity polynomial in the size of the database.

9.3.6.1.6. Polynomial Complexity Proof

The number of pathways of size at most \( k \) that can be formed by \( n \) reactions is at most the number of subsets if the reaction set with cardinality at most \( k \). Thus, the number of pathways of size at most \( k \) is equal to:

\[ \sum_{i=1}^{k} \binom{n}{i} \]  \hspace{1cm} (9.9)

which is a polynomial in \( n \). Since the construction of each pathway of size \( i \) takes at
most i pathway-combination steps, each of which has a cost (in terms of forming the various stoichiometries of the pathway, checking redundancy, checking minimality, and all other bookkeeping) of order polynomial in i.

Thus, the complexity of synthesizing, from a database of size n, all pathways of size at most k is polynomial in n.

9.3.6.1.7. Corollary on the Complexity with respect to Pathway Size

From the proof above, it follows that with respect to the size of the pathways, k, the complexity of the algorithm is at worst exponential, or, more precisely, \( O(p(n)^q(k)) \) where \( p \) and \( q \) are polynomials.

If the size of the pathways is not restricted, the complexity becomes at worst \( O(n^n) \), or, more precisely \( O(p(n)^q(n)) \) where \( p \) and \( q \) are polynomials.

9.3.6.2. Computational Properties

The algorithm was implemented as part of the facillites for pathway design described in Appendix A. The performance of the implementation of the algorithm greatly varies with the exact formulation of the problem, because the number of solutions (final set of pathways) depends on subtle points of the problem formulation such as choosing a key metabolite and designating it as an allowed reactant and allowed product. This issue will be discussed in the context of specific examples (in Section 9.4.1.2.) for which time and space requirements will be given.

We will provide here some rough estimates of time requirements and space usage, for Symbolics 3640 and 3650 machines.
• The requirements for setting up the initial data structures (which are essentially a local copy of the database) are proportional (linear) to the size of the database. Rough requirements per database object (i.e., per metabolite or bioreaction) are:

  ◊ Time elapsed: 0.05 s
  ◊ List words used: 70
  ◊ Structure words used: 70

• The requirements for the main body of the algorithm appear proportional to the number of solutions for those cases in which results were obtained. Thus, per synthesized pathway:

  ◊ Time elapsed: 0.15 s
  ◊ List words used: 200
  ◊ Structure words used: 100

Thus, the requirements for the average solved problem are as follows:

• Setting up the initial data structures for 220 reactions and 400 metabolites requires 35s, 40k list words, and 40k structure words

• Construction of 5000 pathways requires 8 minutes, 1M list words, and 500k structure words.
9.4. DISCUSSION AND SIMPLE EXAMPLES

9.4.1. DEPENDENCE OF THE COMPUTATIONAL EFFICIENCY ON THE FORMULATION OF THE PROBLEM

While we believe that the algorithm is reasonably efficient, i.e., it will run in a short time (of the order of a few minutes) for carefully-formulated problems, we cannot overemphasize the importance of extreme care and caution in the formulation of a problem.

In this section we give some direct guidelines about proper formulation of synthesis problems, and discuss the computational requirements of some examples to further clarify the way in which a good formulation can be constructed in an iterative fashion.

9.4.1.1. Common Currency Metabolites

If one does not designate common energy or oxidation-reduction currency metabolites (such as NAD or ATP) to be allowed reactants and allowed products, the algorithm is very likely to run forever.

For example, consider the currency metabolites involved in the management of Gibbs Energy (PPI, Pi, ATP, ADP, and AMP) and three classes of subpathways:

- Many subpathways acting as futile cycles converting ATP to ADP and Pi (or, in general, higher Gibbs-Energy metabolites to lower Gibbs-Energy metabolites). Let k be the number of these subpathways.
• m subpathways which achieve the main function (the transformation specified by the particular synthesis problem) while gaining energy (in terms of ATP and ADP).

• Similarly, n subpathways achieving the function at the expense of energy (i.e., consumption of ATP).

If the energy-currency metabolites are excluded from appearing in the net stoichiometry, then we must combine subpathways from the three classes to form stoichiometrically acceptable pathways. The number of combinations that are created may be of order higher than \( O(nmk) \), depending on the exact stoichiometries with respect to the currency metabolites.

If each of \( n, m, k \) is of the order of 100, the total number of pathways may be of the order of \( 10^6 \) or larger. Regardless of the efficiency of the implementation this number of pathways will lead to impossible time and space requirements.

The examples in the next paragraphs will show that other metabolites, which are not considered normally currency-metabolites, may cause similar effects and must be designated as allowed reactants and allowed products. Table 9.1 shows certain abbreviations of the names of metabolites; these abbreviations will be used in the depiction of pathways through the rest of this chapter.
Table 9.2: Abbreviations of the names of metabolites that will be used in the pathways for serine (Section 9.4.1.1) and alanine (Section 9.4.1.2)

<table>
<thead>
<tr>
<th>ABBREVIATION</th>
<th>METABOLITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glc</td>
<td>Glucose</td>
</tr>
<tr>
<td>Glc6P</td>
<td>Glucose-6-phosphate</td>
</tr>
<tr>
<td>Fru6P</td>
<td>Fructose-6-phosphate</td>
</tr>
<tr>
<td>FruDP</td>
<td>Fructose-1,6-diphosphate</td>
</tr>
<tr>
<td>GAP</td>
<td>Glyceraldehyde-3-phosphate</td>
</tr>
<tr>
<td>DHAP</td>
<td>Dihydroxyacetone-phosphate</td>
</tr>
<tr>
<td>3PG</td>
<td>3-phosphoglycerate</td>
</tr>
<tr>
<td>2PG</td>
<td>2-phosphoglycerate</td>
</tr>
<tr>
<td>PEP</td>
<td>Phosphoenolpyruvate</td>
</tr>
<tr>
<td>Pyr</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>AcCoA (or Acetyl-CoA)</td>
<td>Acetyl-Coenzyme-A</td>
</tr>
<tr>
<td>OxAc</td>
<td>Oxaloacetate</td>
</tr>
<tr>
<td>Fum</td>
<td>Fumarate</td>
</tr>
<tr>
<td>Mal</td>
<td>Malate</td>
</tr>
<tr>
<td>αkG</td>
<td>α-ketoglutarate</td>
</tr>
<tr>
<td>Glt or Glu</td>
<td>Glutamate</td>
</tr>
<tr>
<td>Asp</td>
<td>Aspartate</td>
</tr>
<tr>
<td>3P-Ser or P-Ser</td>
<td>3-Phospho-serine</td>
</tr>
<tr>
<td>3P-OH-Pyr</td>
<td>3-Phosphohydroxypyruvate</td>
</tr>
</tbody>
</table>
9.4.1.1. Synthesis of Serine

We consider here the simple problem of synthesis of serine from glucose. There is a significant fan-out factor at the two ends of the pathway as:

- There are 6 reactions in which glucose participates:
  - Glucose dehydrogenase
  - Aldohexose dehydrogenase
  - Maltose phosphorylase
  - Celllobiose phosphorylase
  - Trehalose-phosphorylase
  - Hexokinase

- There are 4 reactions in which serine participates:
  - Serine hydroxy-methyl-transferase
  - Methyl-serine hydroxy-methyl-transferase [hydroxy-methyl-serine]
  - Pyrophosphate serine phospho-transferase
  - Phosphoserine phosphatase

9.4.1.1.1. Initial Formulation

A first formulation of the problem would involve the following stoichiometric specifications:

- Required reactant: Glucose
- Required product: Serine
- Allowed reactant: NH₃
- Allowed product: CO₂
• Allowed reactants and allowed products:
  ◊ Oxidation-reduction currency metabolites:
    NAD, NADH, NADP, NADPH, FAD, FADH₂
  ◊ Direct energy-currency metabolites: ATP, ADP, AMP
  ◊ Indirect energy-currency metabolites:
    GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CDP, CMP
  ◊ Other: Coenzyme-A (CoA), Phosphate (Pi), Pyrophosphate (PPI)

The problem appears at first to be well formulated: We are looking for pathways that can synthesize serine from glucose and ammonia, with carbon dioxide as a possible by-product.

Testing this formulation, we find that it cannot run within practical time and appears to get stuck in the processing of constraints on Acetyl-CoA (which can be considered a currency-metabolite but was not treated the same way as the other currency metabolites in the present formulation).

9.4.1.1.2. Inclusion of Acetyl-CoA

In the absence of any deep insights in the structure of the solution space, we modify the problem specification by introducing as an allowed reactant and allowed product that metabolite at which the program stalled — in this case Acetyl-CoA.

However, the program still does not produce results, and it now stalls in the processing of malate.
9.4.1.1.3. Inclusion of Malate

This leads us to introduce malate as an allowed reactant and allowed product, to arrive at the formulation:

- Required reactant: Glucose
- Required product: Serine
- Allowed reactant: NH₃
- Allowed product: CO₂

- Allowed reactants and allowed products: Acetyl-CoA, malate, NAD, NADH, NADP, NADPH, FAD, FADH₂, ATP, ADP, AMP, GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CD P, CMP, CoA, Pi, PPi

For this formulation the program runs to completion. The computational requirements (beyond the requirements for initialization which are constant, as discussed in Section 9.3.8) are:

- Time elapsed: 6 minutes
- List words: 750k
- Structure words: 450k

With respect to the number of constructed pathways:

- A total of 4356 pathways were constructed
- 2830 of the pathways involve neither serine nor glucose.

They are, in effect, neutral pathways (Section 9.3.5.2) interconverting allowed reactants and allowed products.
• The remaining 1526 pathways involve either serine or glucose. The longest of these pathways involves 26 reactions.

Figures 9.1 through 9.6 depict some of the pathways derived from this particular formulation of the problem.

Figure 9.1 depicts what one would consider as the normal pathway for the synthesis of serine, with glucose as the main reactant. Glucose follows glycolysis to 3-phosphoglycerate which is converted to serine through 3-phospho-hydroxy-pyruvate and 3-phosphoserine. The glutamate required for the nitrogen of serine is produced by glutamate-dehydrogenase.

Figure 9.2 shows a pathway in which malate, rather than glucose, is the main reactant. Malate is converted to pyruvate and carbon dioxide by malate-dehydrogenase[decarboxylating]. Pyruvate follows glycolysis upstream to 3-phosphoglycerate which is converted to serine by the same sequence of three bioreactions as in the pathway of Figure 9.1.

Figure 9.3 depicts a pathway identical to the pathway of Figure 9.2, but with one additional enzyme, p-ruvate-dehydrogenase, which recovers the carbon-dioxide produced by malate-dehydrogenase[decarboxylating], by converting Acetyl-CoA to pyruvate.
Figure 9.1:
Synthesis of serine from glucose,
with recovery of glutamate by *Glutamate-dehydrogenase*
Figure 9.2:

Synthesis of serine from malate,

with recovery of glutamate by Glutamate-dehydrogenase
Figure 9.3:

Synthesis of serine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate by \textit{Glutamate-dehydrogenase}
The pathways of Figures 9.4, 9.5, and 9.6 are similar to the pathways of Figures 9.1, 9.2, and 9.3, with one important difference. The recovery of glutamate is not done through glutamate-dehydrogenase, but rather through a set of four bioreactions, which involve as intermediates fumarate, aspartate, oxaloacetate, and malate.

This particular option for converting α-keto-glutarate and ammonia into glutamate is merely one of a number of possibilities. Another similar possibility is the well-known route through glutamine, using the bioreactions glutamine synthetase and glutamine synthase.
Figure 9.4:

Synthesis of serine from glucose,
with recovery of glutamate through a loop
involving oxaloacetate, aspartate, fumarate, and malate
Figure 9.5:

Synthesis of serine from malate,
with recovery of glutamate through a loop
involving oxaloacetate, aspartate, fumarate, and malate
Figure 9.6:

Synthesis of serine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate.
9.4.1.1.4. Inclusion of Oxaloacetate

In the last refinement of the problem, the number of pathways produced and the computational resources required would have been reduced further, if we had taken the bolder step to include not only malate but also oxaloacetate (which was the second most difficult metabolite) as allowed reactants and allowed products. Such a step would lead to the following statistics:

- Computational requirements:
  ◇ Time elapsed: 1.5 minutes
  ◇ List words: 180k
  ◇ Structure words: 120k

- Number of pathways:
  ◇ Total: 1293
  ◇ Using serine or glucose: 246

Figures 9.7, 9.8, and 9.9 show some of the pathways that are produced in this formulation but not in the previous formulations.

The pathways in the three figures correspond to the previous pathways of Figures 9.4, 9.5, and 9.6 which recovered glutamate through a loop that involved oxaloacetate, aspartate, fumarate, and malate. The difference between the new pathways and the old ones is that, since now both malate and oxaloacetate are allowed reactants and allowed products, one of the reactions of the glutamate-recovering loop, namely malate-dehydrogenase, has not been included.

The net result is that the recovery of glutamate in Figures 9.7, 9.8, and 9.9 involves consumption of malate and production of oxaloacetate.
Figure 9.7:

Synthesis of serine from glucose,

with recovery of glutamate accompanied by

consumption of Malate and production of oxaloacetate.
Figure 9.8:

Synthesis of serine from malate, with recovery of glutamate accompanied by consumption of Malate and production of oxaloacetate.
Figure 9.9:

Synthesis of serine from malate and Acetyl-CoA, with recovery of glutamate accompanied by consumption of Malate and production of oxaloacetate.
9.4.1.2. Synthesis of Alanine

We will discuss here the synthesis of alanine from glucose. With respect to the fan-out factors, there are in our database 6 bioreactions involving glucose and 4 bioreactions involving alanine. The latter are:

- *Alanine dehydrogenase*
- *Methyl-serine hydroxy-methyl-transferase*
- *Alanine aminotransferase*
- *β-Alanine aminotransferase*

9.4.1.2.1. Initial Formulation

- Required reactant: Glucose
- Required product: Alanine
- Allowed reactant: NH$_3$
- Allowed product: CO$_2$
- Allowed reactants and allowed products: NAD, NADH, NADP, NADPH, FAD, FADH$_2$, ATP, ADP, AMP, GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CDP, CMP, CoA, Pi, PPI

This formulation is, as with serine (Section 9.4.1.1.1), too tight to generate a manageable number of pathways.
9.4.1.2.2. Inclusion of Malate and Acetyl-CoA

We must add, as before (Sections 9.4.1.1.2 and 9.4.1.1.3), malate and Acetyl-CoA to the list of allowed reactants and allowed products, to reach the formulation:

- Required reactant: Glucose
- Required product: Alanine
- Allowed reactant: NH₃
- Allowed product: CO₂
- Allowed reactants and allowed products: Acetyl-CoA, malate, NAD, NADH, NADP, NADPH, FAD, FADH₂, ATP, ADP, AMP, GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CDP, CMP, CoA, Pi, PPI

This formulation runs to completion. The statistics are:

- Computational requirements:
  ◊ Time elapsed: 5 minutes
  ◊ List words: 650k
  ◊ Structure words: 460k

- Number of pathways:
  ◊ Total: 4065
  ◊ Using alanine or glucose: 1446

Figures 9.10, through 9.18 show nine of the pathways synthesized for the production of alanine for this particular formulation of the problem.
Figure 9.10 depicts what one would consider as the normal pathway for the synthesis of alanine, with glucose as the main reactant. Glucose follows glycolysis to pyruvate which is converted to alanine by *Alanine aminotransferase*. The glutamate required by this reaction (for the amino group of alanine) is produced by *glutamate-dehydrogenase*.

Figure 9.11 shows a pathway in which malate, rather than glucose, is the main reactant. Malate is converted to pyruvate and carbon dioxide by *malate-dehydrogenase*[decarboxylating]. Pyruvate is converted to alanine by the same sequence of bioreactions as in the pathway of Figure 9.10.

Figure 9.12 depicts a pathway identical to the pathway of Figure 9.11, but with one additional enzyme, *pyruvate-dehydrogenase*, which recovers the carbon-dioxide produced by *malate-dehydrogenase*[decarboxylating], by converting Acetyl-CoA to pyruvate.
Figure 9.10:

Synthesis of alanine from glucose, with recovery of glutamate by Glutamate-dehydrogenase.
Figure 9.11:

Synthesis of alanine from malate, with recovery of glutamate by Glutamate-dehydrogenase.
Figure 9.12:

Synthesis of alanine from malate and Acetyl-CoA,
without production of carbon dioxide,
and with recovery of glutamate by Glutamate-dehydrogenase.
The pathways of Figures 9.13, 9.14, and 9.15 are similar to the pathways of Figures 9.10, 9.11, and 9.12, with one important difference. The recovery of glutamate is not done through glutamate-dehydrogenase, but rather through a set of four bioreactions, which involve as intermediates fumarate, aspartate, oxaloacetate, and malate. Thus, these pathway resemble the pathways for the synthesis of serine (Figures 9.4, 9.5, and 9.6).

As we mentioned for serine, the conversion of α-keto-glutarate and ammonia into glutamate can take place through a number of other routes, such as the bioreactions glutamine synthetase and glutamine synthase.
Figure 9.13:

Synthesis of alanine from glucose,
with recovery of glutamate through a loop
involving oxaloacetate, aspartate, fumarate, and malate
Figure 9.14:

Synthesis of alanine from malate,
with recovery of glutamate through a loop
involving oxaloacetate, aspartate, fumarate, and malate
Figure 9.15:

Synthesis of alanine from malate and Acetyl-CoA, without production of carbon dioxide, and with recovery of glutamate through a loop involving oxaloacetate, aspartate, fumarate, and malate.
The last family of pathways we will discuss from this problem-formulation is shown in Figures 9.16, 9.17, and 9.18. These pathways are similar to the pathways of Figures 9.10, 9.11, and 9.12, except that alanine is not produced by alanine aminotransferase but rather by alanine dehydrogenase. Since the latter bioreaction uses ammonia directly (rather than glutamate) there is no need for glutamate dehydrogenase or any other scheme for the conversion of α-keto-glutarate to glutamate.

The pathways of Figures 9.16, 9.17, and 9.18 demonstrate this concept, utilizing as reactants glucose (Figure 9.16), malate (Figure 9.17), or a combination of malate and Acetyl-CoA (Figure 9.18).
Figure 9.16:

Synthesis of alanine from glucose, with *Alanine-dehydrogenase*
Figure 9.17:

Synthesis of alanine from malate, with *Alanine-dehydrogenase*
Figure 9.18:
Synthesis of alanine from malate and Acetyl-CoA, without production of carbon dioxide, and with *Alanine-dehydrogenase*
9.4.1.2.3. Decomposition of the Problem

At this point, we have obtained a solution which involves too many pathways, and we look for ways to reduce the number of pathways. We notice that pyruvate is an intermediate in all the pathways that lead to alanine. We can verify this by formulating a synthesis problem with pyruvate as an excluded intermediate; this formulation yields no acceptable pathways.

Thus, we can decompose the original problem into two subproblems that can be investigated separately:

- Synthesis of pyruvate from glucose
- Synthesis of alanine from pyruvate

9.4.1.2.4. Synthesis of Alanine from Pyruvate

We focus here on the second of the two subproblem listed above, which we formulate as follows:

- Required reactant: (none!)
- Required product: Alanine
- Allowed reactant: NH$_3$
- Allowed product: CO$_2$
- Allowed reactants and allowed products: Pyruvate, Malate, Acetyl-CoA, NAD, NADH, NADP, NADPH, FAD, FADH$_2$, ATP, ADP, AMP, GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CDP, CMP, CoA, Pi, PPI
This formulation runs to completion with the following statistics:

- Computational requirements:
  - Time elapsed: 4 minutes
  - List words: 470k
  - Structure words: 210k

- Number of pathways:
  - Total: 2425
  - Using alanine or glucose: 153

We have thus reduced the number of pathways to the point where one could start examining them one by one. The pathways that are derived from this problem formulation are much simpler, since the variety of ways in which pyruvate can be produced are eliminated.

The whole series of pathways for the production of alanine that we have discussed so far (i.e., Figures 9.10 through 9.18), as well as a number of other alternatives, collapse into just three simple pathways, shown in Figures 9.19, 9.20, and 9.21. These pathways represent the three (demonstrated) ways in which alanine can be derived from pyruvate:

- Through Alanine aminotransferase and glutamate dehydrogenase (Figure 9.19).

- Through Alanine aminotransferase and a loop for the recovery of glutamate, with aspartate, fumarate, malate, and oxaloacetate as intermediates (Figure 9.20).

- Through a single bioreaction, Alanine dehydrogenase, which does not require glutamate at all (Figure 9.21).
Figure 9.19:

Synthesis of alanine from pyruvate,

with recovery of glutamate by *Glutamate-dehydrogenase*
Figure 9.20:

Synthesis of alanine from pyruvate, with recovery of glutamate through a loop involving aspartate, fumarate, malate, and oxaloacetate.
Figure 9.21:

Synthesis of alanine from pyruvate through a single bioreaction, *Alanine dehydrogenase*, which does not require glutamate.
9.4.1.2.5. Elimination of Equivalent Bioreactions

A last improvement can be made if we observe that there are, in the database, pairs of bioreactions which happen to achieve the same transformations, except for minor differences in the energy metabolites, which are not important in this particular problem. Four such pairs are:

- *Pyruvate-phosphate-dikinase* and *Phospho-enol-pyruvate kinase* both interconvert pyruvate and phospho-enol-pyruvate (PEP)

- *Phosphoglycerate-phospho-mutase* and *Phosphoglycerate mutase* both interconvert 2-phosphoglycerate and 3-phosphoglycerate.

- *Oxaloacetate-decarboxylase* and *Pyruvate-carboxylase* both interconvert pyruvate and oxaloacetate

- *ATP-citrate-lyase* and *Citrate-synthetase* both interconvert oxaloacetate and citrate

When two bioreactions are interchangeable, we can designate one of them as an excluded reaction, with the understanding that in the end we can substitute one reaction for the other in synthesized pathways. Applying this idea in this case, we formulate the problem:

- Required product: Alanine

- Allowed reactant: NH₃

- Allowed product: CO₂
- Allowed reactants and allowed products: Pyruvate, Acetyl-CoA, NAD, NADH, NADP, NADPH, FAD, FADH₂, ATP, ADP, AMP, GTP, GDP, GMP, TTP, TDP, TMP, UTP, UDP, UMP, CTP, CDP, CMP, CoA, Pi, PPI

- Excluded reactions:
  ◊ Pyruvate-phosphate-dikinase,
  ◊ Phosphoglycerate-phospho-mutase,
  ◊ Oxaloacetate-decarboxylase,
  ◊ ATP-citrate-lyase

This formulation runs to completion with statistics substantially improved over the previous version:

- Computational requirements:
  ◊ Time elapsed: 2.5 minutes
  ◊ List words: 260k
  ◊ Structure words: 130k

- Number of pathways:
  ◊ Total: 1302
  ◊ Using alanine or glucose: 65

If we examine the generated pathways closer (since their number is now small enough), we see that 64 of the synthesized pathways involve the bioreaction Alanine-aminotransferase which is considered the standard route for the synthesis of alanine (Figures 9.19 and 9.20). The remaining pathway uses, instead, the bioreaction Alanine-dehydrogenase, which involves direct use of ammonia rather than glutamate...
None of the pathways involve *Methyl-serine hydroxy-methyltransferase* or *β-Alanine aminotransferase*.

### 9.4.2. TAKING ADVANTAGE OF CHARACTERISTICS OF BIOCHEMICAL NETWORKS

To achieve better performance in the metabolite-processing phase, it is best to choose to satisfy first the constraint that creates the smallest possible number of new pathways to be created, while removing as many pathways as possible.

This strategy leads to very significant progress (i.e., reduction of the cardinality of the active pathway set) at the early stages of the metabolite-processing phase. This is caused by the nature of metabolic reaction networks: Although they have a densely connected intermediary metabolism, they also contain isolated, specialized pathways for products or substrates, connected only loosely (sparsely) with the intermediary metabolism.

- If a specialized compound is produced by such a linear pathway, loosely interacting with the rest of the metabolism, the metabolite-processing phase will initially create a chain leading from the intermediary metabolism to the target compound. Then:

  ◊ If the compound is an excluded product, and there is no pathway to recycle it back to the intermediary metabolism, the whole chain (pathway) will be discarded. In effect, all the reactions in the chain are discarded with very low computational expense, since the whole investigation takes place only once.
◊ If the compound is an allowed or required product then the set of reactions that produces it will be abstracted into a single pathway and will be used as is in the rest of the processing. This saves the computational expense of searching the pathway step by step repeatedly.

◊ If the compound is an excluded product but it can be recycled, then the whole set of reactions that produces and recycles the compound will be similarly abstracted into a single pathway.

• Similarly, if a specialized substrate is consumed by a linear pathway, loosely interacting with the rest of the metabolism, a chain leading from the substrate to the intermediary metabolism will be created. Then:

◊ If the compound is an excluded reactant, and there is no pathway to produce it from the intermediary metabolism, the whole chain (pathway) will be discarded.

◊ If the compound is an allowed or required reactant then the set of reactions that consumes it will be abstracted into a single pathway.

◊ If the compound is an excluded reactant but it can be produced from the intermediary metabolism, then the set of reactions that produces and consumes the compound will be abstracted into a single pathway.
Consider the network of Figure 9.22, and suppose that P, A, and Q are excluded reactants and excluded products, while B is an allowed reactant and R an allowed product. Suppose also that all remaining intermediate metabolites that are outside the intermediary metabolism are excluded reactants and excluded products. Based on the previous discussion, in the early stages of the metabolite-processing phase of the algorithm, the following will take place:

- The whole pathway that synthesizes and degrades P will be constructed, and the individual reactions will be discarded.

- Similarly, pathways will be constructed for the consumption of A and the production of Q, and the individual reactions will be discarded.

- The reactions involved in the metabolism of B and R will be discarded, without constructing any pathway.

Thus, the structured character of metabolic networks allows us to order the processing of metabolite-constraints to achieve significant gains in computational efficiency through the pruning and abstraction of reaction sets related to special substrates or products.
Figure 9.22: The structured character of biochemical reaction networks is exploited by the synthesis algorithm in early pruning and abstraction.
9.5. SUMMARY

We discussed in this chapter the problem of synthesizing biochemical pathways that satisfy linear stoichiometric constraints.

The algorithm we devised for the solution of the problem is based on the iterative satisfaction of requirements, and the transformation of the initial set of reactions (which can be thought of as one-step pathways) into a final set of pathways, which satisfy all requirements.

The algorithm is correct and complete and has satisfactory computational performance for carefully formulated problems.

In Chapter 10, the utility of the algorithm will be demonstrated in the context of a case study on the production of lysine from glucose.
CHAPTER 10

A CASE STUDY

IN PATHWAY DESIGN
10.1. INTRODUCTION

We are going to perform, in this chapter, a case study on the synthesis and evaluation of biochemical pathways for the production of lysine from glucose and ammonia. The intent is to demonstrate the application of the synthesis algorithm (Chapter 9), the group-contribution method (Chapter 7), and the maximum-rate methodology (Chapter 8) in a real system.

Since the mechanistic and theoretical aspects of the methods have been thoroughly discussed in their respective chapters, we will focus here on what one can achieve with the methods. As we use the methods, we will avoid citing the particular method used or indicate the precise problem formulation.

We should emphasize, however, that all the analysis in this chapter is based solely on the synthetic and analytic methodologies described in previous chapters and literature data (on reaction stoichiometries, decompositions of molecules into groups, Gibbs Energies, and equilibrium constants) stored in our database.

10.1.1. LYSINE

The selection of lysine for this case study is based on the fact that there is on-going experimental and theoretical work by other researchers on the production of lysine from glucose and ammonia.

It should be emphasized right from the start that the analysis we perform in the Chapter is not exhaustive; our aim was merely to demonstrated the concerted application and utility of our methods in a real system and not to arrive to definitive answers on the synthesis of lysine. However, we did try to find the most interesting results that our
methods can provide. Thus, it is possible (but not likely) that better results might be obtained with more detailed use of the methods.

Another important point that must be put forth here is that lysine possesses no particular qualities (positive or negative) with respect to our methodologies. Thus the utility of our methods in examining the synthesis of other products is not expected to vary radically from their performance in the case of lysine.

10.1.2. FRAMEWORK OF THE CASE STUDY

The basic procedure we will follow in this case study is as follows:

- We synthesize a pathway as close as possible to the pathway believed to prevail

- We identify bottlenecks in the pathway by performing a maximum-rate analysis for each reaction

- We synthesize pathways that bypass bottlenecks

- We synthesize other pathways to explore alternatives that omit key enzymes

- We try to identify fundamental constraints on the structure and yield of the pathways

Note that this is a procedure suggested from the point of view of the goals of the analysis. The exact application of the methods may take place following a number of different structures.

For example, one can generate all pathways producing the desired product from the substrates a priori, carry out all the maximum rate calculations for all pathways,
and then perform all the tasks by appropriate search through this (potentially very big) set of pathways.

Alternatively, one can match the application of the methods to the framework, step by step. We believe, however, that there is no practical way to carry out the analysis without the methods developed in this work.
10.2. A BASIC PATHWAY FOR THE PRODUCTION OF LYSINE

10.2.1. BIOCHEMICAL NETWORK

Table 10.1 shows the abbreviations that we will use for metabolic intermediates throughout this chapter. The core of the bioreaction network with which we will work is shown in Figure 10.1. It includes:

- Glycolysis

- Lactate dehydrogenase, converting pyruvate to lactate (a common anaerobic fate for pyruvate)

- The usual citric acid cycle (or tricarboxylic acid cycle, which will be referred to as TCA), with the exception of the bioreaction α-ketoglutarate dehydrogenase which will be assumed to be absent or non-functional

- The glyoxylate shunt to complement TCA and make up for the absence of α-ketoglutarate dehydrogenase

- The bacterial pathway that leads from oxaloacetate to aspartate and on to lysine.

- Glutamate dehydrogenase and glutamine synthetase for the synthesis of glutamate and glutamine

Figure 10.1 was constructed to conform to the bioreaction network used in the analysis of experiments of lysine production. Note that the figure is substantially simplified, as:
• Many side-reactants and side-products are not shown.

• Many reactions are lumped together. In particular, the arrow drawn from aspartate-semialdehyde (ASA) to lysine represents 6 individual bioreactions.
Table 10.1:
Abbreviations for metabolic intermediates used in biological reaction networks

<table>
<thead>
<tr>
<th>ABBREVIATION</th>
<th>METABOLITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glc</td>
<td>Glucose</td>
</tr>
<tr>
<td>Glc6P</td>
<td>Glucose-6-phosphate</td>
</tr>
<tr>
<td>Fru6P</td>
<td>Fructose-6-phosphate</td>
</tr>
<tr>
<td>GAP</td>
<td>Glyceraldehyde-3-phosphate</td>
</tr>
<tr>
<td>3PG</td>
<td>3-phosphoglycerate</td>
</tr>
<tr>
<td>2PG</td>
<td>2-phosphoglycerate</td>
</tr>
<tr>
<td>PEP</td>
<td>Phosphoenolpyruvate</td>
</tr>
<tr>
<td>Pyr</td>
<td>Pyruvate</td>
</tr>
<tr>
<td>AcCoA (or Acetyl-CoA)</td>
<td>Acetyl-Coenzyme-A</td>
</tr>
<tr>
<td>Cit</td>
<td>Citrate</td>
</tr>
<tr>
<td>i-Cit</td>
<td>Isocitrate</td>
</tr>
</tbody>
</table>

(continued on the next page)
<table>
<thead>
<tr>
<th>ABBREVIATION</th>
<th>METABOLITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OxAc</td>
<td>Oxaloacetate</td>
</tr>
<tr>
<td>Glyox</td>
<td>Glyoxylate</td>
</tr>
<tr>
<td>Fum</td>
<td>Fumarate</td>
</tr>
<tr>
<td>Mal</td>
<td>Malate</td>
</tr>
<tr>
<td>Suc</td>
<td>Succinate</td>
</tr>
<tr>
<td>SucCoA</td>
<td>Succinyl-Coenzyme-A</td>
</tr>
<tr>
<td>αkG</td>
<td>α-ketoglutarate</td>
</tr>
<tr>
<td>Glt or Glu</td>
<td>Glutamate</td>
</tr>
<tr>
<td>Gln</td>
<td>Glutamine</td>
</tr>
<tr>
<td>Asp</td>
<td>Aspartate</td>
</tr>
<tr>
<td>ASA</td>
<td>Aspartate-semialdehyde</td>
</tr>
<tr>
<td>Lys</td>
<td>Lysine</td>
</tr>
<tr>
<td>Gly</td>
<td>Glycine</td>
</tr>
<tr>
<td>Ala</td>
<td>Alanine</td>
</tr>
</tbody>
</table>
Figure 10.1: The basic bioreaction network for the synthesis of lysine
10.2.2. CONSTRUCTION OF BASIC PATHWAY

One of the basic pathways believed to function in bacteria (such as *Brevibacterium Flavum*) for the conversion of glucose to lysine is shown in Figure 10.2. The pathway uses the glyoxylate shunt to bypass *α*-ketoglutarate dehydrogenase, which has not been included in the network.

The pathway of Figure 10.2 is actually only a partial one, because the pathway leading from aspartate to lysine requires succinyl-CoA and glutamate (at the same time producing succinate and a-ketoglutarate).

We can complete the pathway by looking for the smallest possible pathway which contains all the reactions already marked and additional reactions to balance the stoichiometries for succinyl-CoA, glutamate, succinate, and a-ketoglutarate. The stoichiometries of other metabolites (such as ATP or NAD) are also unbalanced, but we will consider them to be allowed reactants and allowed products.

The completed pathway is shown in Figure 10.3. The reactions added are succinate kinase and glutamate dehydrogenase, which, as expected, comprise the simplest alternative.
Figure 10.2:

Partial pathway of the normal route of lysine production
Figure 10.3:
Completion of the basic pathway for the synthesis of lysine.
10.3. MAXIMUM-RATE ANALYSIS

In this section, we will calculate the minimum enzyme requirement, which is a parameter related to the maximum rate of each bioreaction. By comparing the rates of the bioreactions in the basic pathway of Figure 10.3 we will uncover kinetic bottlenecks.

To evaluate the significance of the bottlenecks, we will estimate a rate based on experimental data and compare it to the ideal maximum rate.

10.3.1. PARAMETER VALUES

For the estimation of the maximum rate of a bioreaction, we need the equilibrium constant of the reaction, the concentrations of the reactants and products of the reaction, and the concentrations of enzymes:

- We would like to be as conservative as possible in choosing a value for the equilibrium constant, because it has a very strong effect on the maximum rate. We take the equilibrium constant of a bioreaction to be equal to the maximum value among:
  - Any available data from the literature (residing in the database)
  - The value estimated by the group-contribution method

- The concentrations of all metabolites are assumed to be in the default range we normally use for physiologically acceptable conditions. Thus, the concentrations of the products of each
bioreaction are set to $5 \times 10^{-6}$ and the concentrations of reactants are set to $5 \times 10^{-3}$.

- The concentration of the enzyme is not assumed to have any particular value. Since the maximum rate is proportional to the concentration of the enzyme, we can estimate the quantity $r/E$, i.e., maximum rate over enzyme concentration, leaving the enzyme concentration unspecified.

10.3.2. BASE OF REFERENCE FOR CONCENTRATIONS AND RATES

In the maximum-rate methodology the volume reference for all concentrations and rates is the volume of intracellular material. In effect, the quantity $E$ (in mol/l) refers to intracellular concentration, and $r$ (in mol/s l) refers to rate per unit cell volume.

However, if we are dealing only with the ratio of the two parameters ($r$ and $E$), the volume references or corrections cancel out. Thus we can choose to use culture or reactor volume as the reference. We can even eliminate volume altogether, taking $E$ to denote the absolute amount of enzyme present (in mol) and $r$ to denote the rate of reaction (in mol/s).

Instead of using the ratio $r/E$, we can equivalently estimate the inverse of that ratio, i.e., $E/r$, which denotes the minimum enzyme requirement (per unit rate) for the bioreaction. The actual (i.e., experimental) $E/r$ of a reaction must be higher than our estimate; since actual enzymes are less efficient it takes a higher (than ideally estimated) enzyme concentration to achieve a given rate.
The minimum enzyme requirement, \(E/r\), is a particularly convenient quantity because the minimum enzyme requirement of the whole pathway can be obtained simply by adding together the requirements of all the reactions.

In this context, it is convenient to take \(r\) not as the rate of the bioreaction examined, but rather as the rate of production of the final product. To achieve this transformation of reference-rate, we only need to multiply the initial enzyme requirement of each reaction by the corresponding coefficient of the reaction-stoichiometry of the pathway.

Note that, since a pathway involves many enzymes, the enzyme requirement of the pathway denotes the sum of the concentrations of different enzymes. This is not unreasonable considering that the different enzymes have to coexist and function in the same cell, and compete, in their synthesis, for same limited resources of the cell. Similarly, the pathway as a whole competes for resources for all of its enzymes, because it is functional only when sufficient quantities of all enzymes are present.

Thus, in evaluating a pathway as a whole and comparing it to other pathways, it is useful to lump the concentrations of all the enzymes in the pathway and estimate the minimum enzyme requirement of the pathway.

10.3.3. ESTIMATION OF THE ENZYME REQUIREMENTS OF THE BASIC LYSINE PATHWAY

The minimum enzyme requirement for each bioreaction in the basic pathway of Figure 10.3 is shown in Figure 10.4. For each reaction in Figure 10.4, the number shown is the enzyme requirement of that reaction, in milliseconds.
The total enzyme requirement for the whole pathway of Figure 10.4 is approximately 14 ms.

10.3.3.1. Identification of Kinetic Bottlenecks

About half of the enzyme requirements of the pathway come from the bioreaction *malate dehydrogenase*, which has an enzyme requirement of 6.44 ms.

The next larger contribution, equal to 2.7 ms, comes from *glyceraldehyde-phosphate dehydrogenase*. However, we have very little control over that enzyme since it belongs to glycolysis. Thus, *malate dehydrogenase* remains the main kinetic bottleneck of the pathway.

10.3.3.2. Comparison to Experimental Enzyme Requirements

In order to assess how severe the bottleneck posed by *malate dehydrogenase* is in reality, we will compare the minimum enzyme requirements to requirements based on experimental data.

10.3.3.2.1. Estimation of Experimental Enzyme Requirements

To make a comparison, we need to estimate the concentration, $E$, and the rate, $r$, from experimental data:

- The rate, $r$, of the pathway is equal to the rate of lysine production (in moles per unit time), which, in a batch culture, is easily estimated from successive measurements of the concentration of lysine.
• The enzyme concentration is the product of the following factors:
  ◦ Cell mass present, ln gr, measured experimentally
  ◦ Total protein mass as a fraction of cell mass, approximately equal to 0.5 [Ingraham et al., 1983]
  ◦ Mass fraction of total protein that makes up enzymes of the lysine pathway, assumed to be equal to 0.10.

The last parameter is the most difficult to estimate. Taking into account the number of biochemical reactions that must be active for the survival of the cell, and the extent of the role of proteins in other functions (such as replication, transcription, or translation), the assumed fraction of 0.10 is likely to be on the conservative side. Clearly, the most conservative upper bound is 1.

Based on the above parameters and assumptions, we estimate the enzyme requirement from a set of experimental data (from the lysine fermentation with *Brevibacterium Flavum*; the data were provided by J. Vallino and Greg. Stephanopoulos, Department of Chemical Engineering, Massachusetts Institute of Technology).

For a significant part of the fermentation (30 hours out of a total of 70 hours), the rate of lysine production and the concentration of biomass are constant. For this region, the estimation yields an approximate value of $2 \times 10^3$ ms for the enzyme requirement of the whole pathway.
10.3.3.2.2. Factors causing Disparity between Minimum and Actual Requirements

The maximum rate is about two orders of magnitude away from the experimental rate. There are several factors that must be taken into account in evaluating the deviation of an actual enzymatic reaction rate from the maximum. Some of the most significant factors are:

- Enzyme inefficiencies. In general, enzymes are less efficient (by an order of magnitude or more) than the maximum rate would imply.

- Regulatory effects. If there is any regulation, it may cause an inhibition by a factor of 2 or more, even when the regulation is turned off.

- Tighter concentration regions. In our estimates we assumed concentrations of reactants and products in a fairly wide range, independently for each reaction, and regardless of any physiological requirements on the energy charge and the reduction charge of the cells. In reality, concentrations must be consistent through the whole network and conform to other physiological constraints.

10.3.3.2.3. Interpretation of the Relation between Estimated and Experimental Requirements

If we take into account all of the above factors, the difference between the minimum enzyme requirement and the requirement based on experimental data is not
very high, and there is room for doubt as to whether the pathway is capable of achieving the measured rates.

This suggests that there is good reason to look for bypasses to the kinetic bottleneck caused by malate dehydrogenase. Specifically:

- If the pathway of Figure 10.4 is not the true pathway, a bypass is already functioning and we should set out to uncover it.

- If, on the other hand, the pathway does function then it is constrained by the bottleneck, and any improvement in productivity must involve bypasses.

We will not (and need not) make a commitment to one of the two options. Both of the possible interpretations should be kept in mind as we examine bypasses of the bottleneck and alternative pathways.
Figure 10.4: Calculation of minimum enzyme requirements for the basic pathway for lysine production
10.4. ALTERNATIVE PATHWAYS

We seek now new pathways that eliminate the kinetic bottleneck of malate dehydrogenase.

10.4.1. PATHWAYS INVOLVING CARBOXYLATION OF PYRUVATE

In Figure 10.5 we show a first possibility, which has been already determined (experimentally) to function under certain conditions. This pathway involves the carboxylation of pyruvate, bypassing the whole TCA cycle. This direct conversion of pyruvate to oxaloacetate can be achieved by two distinct bioreactions:

- Pyruvate carboxylase
- Oxaloacetate decarboxylase

The pathway of Figure 10.5 successfully bypasses the kinetic bottlenecks because its minimum enzyme requirement is only 6.4 ms, roughly equal to one half the requirement of the initial pathway.

This pathway also has a more attractive maximum molar yield. Its yield is 100%, i.e., the pathway yields one mole of lysine per mole of glucose, as compared to a molar yield of 67% for the initial pathway of Figure 10.4.
Figure 10.5:
Minimum enzyme requirements for a lysine pathway involving carboxylation of pyruvate
10.4.2. PATHWAYS RETAINING TCA

If the original pathway has some good traits, we might prefer to bypass only the immediate vicinity of the bottleneck and retain much of the structure of the original pathway intact, including the TCA cycle.

10.4.2.1. Pathway through Lactate

A first alternative, shown in Figure 10.6, involves bypassing malate dehydrogenase with a set of just two reactions:

- \textit{Lactate-Malate transhydrogenase} achieves the conversion:
  \[ \text{malate} + \text{pyruvate} \rightarrow \text{oxaloacetate} + \text{lactate} \]

- \textit{Lactate dehydrogenase} achieves the conversion:
  \[ \text{lactate} \rightarrow \text{pyruvate} \]

The combination of the two reactions converts malate to oxaloacetate. Unfortunately, the enzyme requirement of this bypass is approximately the same as that of \textit{malate dehydrogenase}. Specifically, \textit{lactate dehydrogenase} has a requirement of 5.32 ms (compared to 6.44 for \textit{malate dehydrogenase}). Thus, this particular pathway offers little improvement over the original one.

It is interesting to note that this pathway uses \textit{lactate dehydrogenase} in the direction opposite to that originally drawn in Figure 10.1.
Figure 10.6: Pathway converting malate to oxaloacetate, with lactate and pyruvate as intermediates.
10.4.2.2. Pathways deriving Aspartate directly from Fumarate

Two more interesting alternatives are shown in Figures 10.7 and 10.8. They both involve:

- Conversion of malate to fumarate by using Fumarase in the direction opposite to that initially assumed in Figure 10.1
- Conversion of succinate to fumarate by Succinate dehydrogenase as in the original pathway
- Conversion of fumarate into aspartate through Aspartate aminolyase

Since oxaloacetate is used in order to form citrate, half of the aspartate must be recycled back into oxaloacetate to close the TCA loop. The two pathways use different ways to achieve this:

- In the pathway of Figure 10.7 the reaction aspartate glutamate transaminase converts oxaloacetate to aspartate, by operating in the direction reverse to that assumed in the original bioreaction network (Figures 10.1 to 10.4).
- The pathway of Figure 10.8 uses a set of two reactions, Glycine dehydrogenase and Glycine-oxaloacetate aminotransferase, involving interconversion of glycine and glyoxylate.

The pathway of Figure 10.8 is longer, but it is actually the most efficient (kinetically) of all the pathways sharing the TCA structure of the original pathway. Its
minimum enzyme requirement is 8 ms, i.e., almost half of the requirement of the original pathway.

10.4.2.2.1. Persistent Intermediates

In the two pathways discussed above, oxaloacetate is partly bypassed, in that it is needed only for the synthesis of citrate, and not directly for the synthesis of aspartate and lysine. An interesting question is whether we can bypass oxaloacetate altogether and produce aspartate directly from pyruvate or glucose.

With the reactions in our database, this turns out to be impossible. Thus, it appears that oxaloacetate is a key intermediate in the production of aspartate and lysine. The pathways we discuss in this chapter (and other pathways which were constructed but will not be discussed) indicate, in fact, that, the only persistent intermediates, i.e., intermediates that occur in all pathways are:

- The intermediates of glycolysis from glucose to phosphoenolpyruvate, with that section of the pathway fixed

- The intermediates of the pathway from aspartate to lysine, a pathway that is also fixed

- Oxaloacetate, for which no surrounding reaction is fixed, but the intermediate itself is always present (participating in different reactions)
Figure 10.7: The simplest of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate
Figure 10.8: The kinetically most efficient of the pathways bypassing malate dehydrogenase by converting fumarate to aspartate.
10.4.3. ALTERNATIVE ENTRIES OF PYRUVATE INTO TCA

We will investigate here some pathways that bypass not only *malate dehydrogenase* but also *pyruvate carboxylase* and *pyruvate dehydrogenase*, to find other ways in which pyruvate can enter TCA.

10.4.3.1. Carboxylation of PEP or Pyruvate

The first alternative, shown in Figure 10.9, involves the direct carboxylation of phosphoenolpyruvate. This pathway is very similar to the pathway of Figure 10.5, which involved carboxylation of pyruvate.

If we eliminate *PEP carboxylase* from consideration, we find another alternative, even more similar to the pathway of Figure 10.5. It involves conversion of pyruvate to oxaloacetate through two bioreactions:

- *Methyl-malonyl-CoA carboxytransferase:
  \[ \text{pyruvate} + \text{methyl-malonyl-CoA} \rightarrow \text{oxaloacetate} + \text{propionyl-CoA} \]

- *Propionyl-CoA carboxylase:
  \[ \text{propionyl-CoA} + \text{CO}_2 \rightarrow \text{methyl-malonyl-CoA} \]
Figure 10.9:
A pathway that carboxylates PEP, bypassing pyruvate
10.4.3.2. Pathways through Acetate

A whole set of alternative pathways for the entry of pyruvate in the TCA cycle involves the use of acetate as an intermediate. The relevant bioreactions are shown in Figure 10.10.

There are 2 short pathways for the production of acetate from pyruvate, and 3 short pathways for the conversion of acetate to acetyl-CoA or citrate. In this count, we do not include bioreactions that are essentially the same, as discussed in Section 9.4.1.2.5; for example, we consider only one of the two possible bioreactions (ATP-citrate-lyase and Citrate-synthetase) that convert oxaloacetate and acetyl-CoA into citrate.

By forming combinations that include exactly one pathway from each of the two sets, we can construct a total of 6 pathways that convert pyruvate and oxaloacetate into citrate and use acetate as an intermediate. It is exactly this phenomenon, of an intermediate produced and consumed in many ways, that creates the explosion in the number of pathways for poorly formulated synthesis problems.
Figure 10.10:

A set of pathways that bypass pyruvate dehydrogenase by converting pyruvate to acetate and then to Acetyl-CoA or citrate
10.4.4. COMPLEX PATHWAYS

Of all the alternative pathways produced by the synthesis algorithm, we generally discussed, so far, only some of the simplest ones. It should be remembered, however, that the algorithm can find all pathways, and that some very complex pathways will be among them.

We consider here the simple task of conversion of PEP to pyruvate, which can be achieved in one step by pyruvate kinase. The algorithm produced several pathways for this conversion, the most complicated (i.e., longest) of which is depicted in Figure 10.11.

It is interesting to note how the algorithm has managed to find roundabout ways to achieve simple conversions throughout the pathway, forming a nice chain of short coupled loops and a bigger loop surrounding them.
Figure 10.11:
The longest pathway converting PEP to pyruvate
10.5. FUNDAMENTAL CONSTRAINTS

Some of the most interesting results of applying the synthesis algorithm involve not particular pathways found, but rather demonstrations that no pathways exist to meet certain sets of specifications.

We discussed already the fact that there is no pathway that will reach aspartate (and consequently lysine) from glucose without going through oxaloacetate. A second interesting constraint that was uncovered by the algorithm refers to the maximum yield of the pathway:

- The yield can exceed 67% only if carbon dioxide is recovered by some bioreaction.

In effect, if we eliminate reactions that consume carbon dioxide, the yield is restricted to be 67% or less.

A point to keep in mind is that these constraints only hold for the set of reactions present in our database. It is entirely possible that inclusion of additional reactions will change these results.
CHAPTER 11

CONCLUSIONS AND SIGNIFICANCE
11.1. SUMMARY

11.1.1. MOTIVATION AND GOALS

When investigating the production of bioproducts through a biochemical process, one has to select among existing metabolic pathways (which could be used in their initial form or modified) or devise a new pathway for the process. Thus, one needs to systematically construct alternative pathways, evaluate the pathways in order to reveal their limitations, and propose ways to improve them.

This project has as its goal the development of systematic methodologies that:

(a) Analyze given biochemical pathways, determining their potential performance and bottlenecks.

(b) Synthesize alternative pathways, including completely new pathways, which meet process specifications.

11.1.2. WORK PERFORMED

This work has yielded a combination of methods in Artificial Intelligence, Mathematics, and Physical Chemistry, addressing the problem of design of biochemical pathways.

Candidate pathways can be generated by a constrained search algorithm, which is based on the iterative satisfaction of constraints through the transformation of a base-set of pathways. Generated pathways can be screened by heuristic rules and detailed analytical examination. Semiquantitative knowledge can be used within the Q[M]
formalism for Order-of-Magnitude reasoning, which captures engineering commonsense concepts about the relative orders of magnitude of parameters.

Additional methodologies make up for the absence of critical data. They include the estimation of maximum rates for biochemical reactions based on collision limitations, and a group-contribution method for the estimation of equilibrium constants. Using the estimated parameters, the thermodynamic and kinetic feasibility of a pathway can be determined, and the bottlenecks of the pathway can be identified. The bottlenecks can be bypassed by synthesizing appropriate modifications to the pathway.

The goals of the project have thus been met, as the set of methodologies that was developed allows the generation of alternative feasible pathways satisfying specifications, evaluation of pathways, identification of bottlenecks, and construction of bypasses to bottlenecks.

We will reiterate below the individual methods that comprise our pathway-design framework.

11.1.2.1. Pathway Analysis

Pathway analysis consists of examining a given biochemical pathway to produce information about its potential performance, critical parameters, and bottlenecks. Analysis leads to recommendations on how an existing or designed pathway could be improved, and what other information about the pathway should be acquired.

The main obstacle in pathway analysis is the sparseness and qualitativeness of the knowledge available about biochemical systems, requiring methods for the extension of incomplete knowledge and the representation and use of qualitative knowledge.
11.1.2.1.1. Group- Contribution Method

In order to evaluate the feasibility and reversibility of a biochemical pathway, one must know the equilibrium constant of each bioreaction. Unfortunately, for most enzymes, the equilibrium constants are not known.

To overcome this limitation, we developed a group-contribution method for the estimation of Gibbs Energies of Formation of biochemical compounds in aqueous solution. The equilibrium constant of a bioreaction can be calculated from the Gibbs Energies of its substrates and products.

The equilibrium constants are used in the evaluation of the thermodynamic feasibility and reversibility of a biotransformation.

11.1.2.1.2. Estimation of Maximum Rates for Enzymatic Reactions

To evaluate the potential production rate of a pathway and to identify its kinetic bottlenecks, one needs to know the kinetics of each enzyme in the pathway. The estimation of the actual rate of a given pathway is in practice impossible because enzyme kinetics are rarely available, and the properties of an enzyme depend on its source.

To overcome this obstacle, we developed a methodology for the estimation of the maximum allowable rate for any enzymatic reaction. The strategy of the method consists of using a typical fast mechanism for the bioreaction and stating constraints stemming from equilibrium and the fact that bimolecular steps cannot take place faster than the two reacting species collide in the solution. The method uses these constraints on the steps of the mechanism to estimate a maximum rate for the overall bioreaction.
With the results of this method, the kinetic efficiency of a whole pathway can be evaluated. The methodology can be applied to reject proposed mechanisms or pathways as unable to meet rate requirements. The maximum rates of the steps of a pathway also point to kinetic bottlenecks which must be removed if the pathway is to be improved.

### 11.1.2.1.3. Order-of-Magnitude Reasoning

To achieve the use of qualitative and semiquantitative knowledge, a formal system that represents semiquantitative notions is needed. The O[M] Order-of-Magnitude formalism was developed to meet this need.

O[M] reasons with approximate relations among parameters. It is based on Order-of-Magnitude relations that can hold between two quantities, such as the relations "A is much less than B" or "A is slightly larger than B." Inferences within the formalism employ Order-of-Magnitude relations, algebraic constraints, and if-then rules.

O[M] facilitates the acquisition and use of semiquantitative knowledge about biochemical systems and the formalization of important tasks, such as the identification of rate-limiting steps, the analysis of fluxes in a metabolic network, or the qualitative analysis of enzyme inhibition.

### 11.1.2.2. Pathway Synthesis

In the synthesis of biochemical pathways, the goal is to generate metabolic pathways that can achieve a desired transformation. The transformation is described in terms of linear stoichiometric requirements, which prescribe reactants, products, intermediates, and bioreactions as necessary, allowed, and prohibited.
The problem was solved by the development of an algorithm for the construction of all candidate pathways, based on the recursive satisfaction of requirements. At each step, the algorithm transforms a set of partial pathways so that they satisfy one more constraint, until all the requirements are satisfied.

The methodology is useful not only for synthesizing whole metabolic pathways for the production of bioproducts but also for designing modifications to a starting pathway to achieve removal of a bottleneck.
11.2. **REITERATION OF SIGNIFICANT CONTRIBUTIONS**

The central contribution of this work is a framework for the analysis and synthesis of biochemical pathways, as illustrated through the case study on lysine production and the applications of Order-of-Magnitude reasoning. The framework covers the whole spectrum of pathway design, from the generation of alternative pathways to the evaluation of individual pathways, identification of bottlenecks, and creation of bypasses to the bottlenecks.

This work has also contributed a series methods which are part of the framework but have broad applicability beyond their particular role in this work. These distinct methods are:

- A group-contribution method for the estimation of Gibbs Energies of formation or reaction, in aqueous solutions. The method offers an easy way to estimate equilibrium constants of bioreactions using only their stoichiometry and the structure of the reactants and products.

- A method for the estimation of the maximum rate (or the minimum enzyme requirement) of any enzymatic reaction, based on the rates of encounter between the small molecules and the enzyme. The method can be used to evaluate the feasibility of enzymatic processes and the consistency of postulated pathways or enzymatic mechanisms.

- A method for the synthesis of all biochemical pathways satisfying a given set of linear stoichiometric constraints. This method is absolutely necessary when one wants to generate alternative
pathways or test whether there exist pathways to achieve a certain transformation.

- The O[M] formalism for reasoning with orders of magnitude and approximate relations. The system is valuable in reasoning with semiquantitative knowledge and formalizing intuitive Order-of-Magnitude concepts.
11.3. FUTURE DIRECTIONS

Starting from this work, there is a multitude of issues that one might attempt to tackle next. The grand goal would be, of course, the fulfillment of the whole research framework of Chapter 4.

Some more specific problems that are particularly fascinating, as well as challenging, are discussed below.

11.3.1. GENETIC CONSIDERATIONS

The regulation of enzyme synthesis plays an extremely important role in the industrial feasibility of a biochemical pathway. The changes achieved by mutation-selection programs often involve favorable tuning of this regulation.

It is extremely hard to reason on the genetic control, because there is usually almost no knowledge available on its structure and parameters. It is reasonable to expect that as the more substantial body of knowledge on the regulation of enzyme synthesis accumulates, it will be possible to develop methodologies that include genetic considerations in the evaluation of biochemical pathways.

11.3.2. POSTULATING ENZYMES

In the methods we developed for the synthesis of biochemical pathways we only use known enzymes. A logical extension would be to synthesize pathways that include postulated enzymes, not known a priori.

This would require the ability to guess, through analogical reasoning, what enzymes are likely to exist, starting from a large set of existing enzymes. The main
difficulty arising is that there is a large number of enzymes that, at least from the human expert's point of view, could very well exist, i.e., they catalyze reasonable reactions, similar to known enzymatic reactions. From all these enzymes, however, only few are known to exist in practice. Thus, one needs to acquire some additional insights (whose nature is unknown) that allow discrimination between likely and unlikely postulated enzymes.

11.3.3. HOST CELLS

Clearly, a large part of the performance of the process depends not on the biochemical pathway used, but on the host cell within which the pathway functions. Thus, an important next step is the development of techniques that can deal not only with pathways but also with host cells. The techniques should be able to recommend particular hosts, and take into account knowledge on hosts in the selection of pathways.
12

BIBLIOGRAPHY


APPENDIX A

COMPUTER PROGRAMS
A.1. SUMMARY

A software system supporting computer-aided design of biochemical pathways is presented.

A database has been constructed, and contains knowledge about the physicochemical parameters of functional groups, substrates, bioreactions, and enzymes. Both the acquisition and the retrieval of information can be performed through a graphic interface.

The interface allows transparent manipulation of groups to form molecules, molecules to form bioreactions, and bioreactions to form pathways. All the attributes of the entities in the database can also be edited. All operations can be carried out through menus and mouse operations.

The system includes analytical methodologies, such as a group-contribution method for the estimation of equilibrium constants, a method for the estimation of maximum rates for biochemical reactions based on collision limitations, and a formal system for Order-of-Magnitude reasoning. A synthetic methodology for the construction of biochemical pathways satisfying stoichiometric constraints has also been implemented.

The implementation of the supporting methodologies, relies on Symbolics 3640 and 3650 computers and the Symbolics Common LISP language, a superset of Common LISP.

Object-Oriented Programming (OOP) is the predominant programming style for all the applications. Using OOP, all important entities, such as bioreactions, metabolites, pathways, constraints, quantities, relations, and rules, are implemented as
objects. These objects have attributes, whose values can be symbols, lists, numbers, or pointers to other related objects.
A.2. INTRODUCTION

A.2.1. PROGRAMMING STYLE

A.2.1.1. Character of the Problem

It was necessary, in the development of the computer programs in this work, to start solving parts of the problem before finalizing the overall problem formulation and solution strategy. In pathway synthesis, for instance, it was not a priori clear what exact formulation of the synthesis problem was useful. The kind of specifications one ought to impose on the synthesized pathways were only decided upon after trying alternative specifications and inspecting the results.

Hence, in the case of the synthesis problem, it was necessary for the implementation to represent biochemical reaction networks and pathways in a way that corresponds closely to the human perception of these concepts — something clearly not possible with traditional programming paradigms. To accomplish this, we used a programming paradigm that allowed us to implement pathways so flexibly that any operation we could conceptually visualize on bioreactions could be transparently implemented in the programs without a major overhaul of the representation schemes used.

A.2.1.2. LISP

We used the LISP programming language, which offers amazing programming flexibility [Charniak and McDermott, 1985]. LISP is oriented towards the manipulation
of symbols, rather than numbers, and allows very easy construction of complicated and flexible data structures.

In addition, data types in LISP do not have to be predefined; they are instead dynamic and checked at run-time. Storage allocation is also done dynamically, and garbage collection (i.e., reclaiming allocated storage) is automatic and transparent to the programmer. This is especially important to achieve efficient algorithms in pathway synthesis and Order-of-Magnitude reasoning (Section A.5.3).

A.2.1.3. Object-Oriented Programming

Another important part of our paradigm is Object Oriented Programming (OOP). In OOP, chunks of information distinguishable as conceptual entities are represented as objects. Objects belong to classes, and they can be constructed, removed, and mutated using a common protocol of methods that are interpreted differently from one class of objects to another.

Thus, we implemented biochemical reactions, metabolites, and Order-of-Magnitude relations as objects. The combination of LISP procedures (and other data-structures) and OOP yielded a nearly ideal programming paradigm that satisfied all the needs of our problem.

A.2.2. COMPUTING ENVIRONMENT

The programming paradigms we chose, and the complexity and exploratory nature of the problem, also introduced needs on the computing environment.
We needed an environment particularly efficient for rapid, exploratory prototyping and execution, so that programs could be developed initially in small chunks, which could be extended and connected later.

### A.2.2.1. Environment and Facilities

We used Symbolics 3640 and 3650 LISP computers, supporting:

- The Symbolics Common LISP language [Symbolics, 1986], which is a dialect of LISP [Steele, 1984]
- Flavors [Symbolics, 1986], which is a powerful Object-Oriented Programming facility
  - The Genera software environment, including:
    - A powerful Debugger
    - The Zmacs Editor, which supports a special mode for editing LISP programs
    - Dynamic Windows and a Presentations System, allowing high-level, convenient programming of mouse operations
    - An Inspector for examining and modifying data-structures

### A.2.2.2. Advantages of the Environment

Some of the important differences between conventional environments and the environment that we used are the following:
• The environment is uniform, with all the system software written in the same language (LISP). Thus, it is easy to invoke programmatically and use any system facility of the Genera environment.

• The creation and management of multiple processes, running in the foreground or background is easy to accomplish programmatically.

• The window system can be manipulated at a high level through LISP code.

• The usual editing-compiling-linking-loading-running cycle is not time-consuming, because:
  
  ◊ Linking is not needed at all (it is done automatically at run-time).
  
  ◊ Compiling and Loading can be done incrementally and combined into a single step.
  
  ◊ The Editor is constantly present, and one can compile the modified parts of the code from the Editor. The Editor also supports special features for editing LISP code.
  
  ◊ The Debugger allows modification of variables, functions, and data-structures; depending on the nature of the error, one may correct the error and proceed without running the program again from the start.
A.2.2.3. Debugging

In the Symbolics environment an execution-error throws the process in the Debugger, where the error can be investigated by scanning data structures, viewing code, and evaluating chunks of the erroneous code or arbitrary LISP expressions. The whole stack of function calls is available, and one can switch from one frame (function call) to another. One can also perform any other operation in other processes and windows of the environment, returning later to the Debugger which retains its state intact.

The error can be remedied by editing a function. In the process, one can access all the callers of the function or the definitions of other functions, which the editor can automatically locate. Recompiling the function in the editor makes the new (corrected) definition active.

If the consequences of the error are sufficiently local, execution can be continued from the point the error occurred. Otherwise the run must be repeated, but no further compilation, loading, or linking is necessary.

A.2.3. FACILITIES DEVELOPED

The facilities that were implemented cover a series of conceptual methodologies discussed in the main body of the thesis. Specifically, there are computer programs which allow:

- Synthesis of biochemical pathways satisfying stoichiometric specifications (described in Chapter 9; a case study is provided in Chapter 10)
• Fitting data to estimate contributions of groups to the Gibbs Energy of formation of compounds in aqueous solution (described in Chapter 7)

• Use of the contributions of groups to estimate Gibbs Energies and equilibrium constants (described in Chapter 7; involved in the case study of Chapter 10)

• Estimation of the maximum rate of any enzymatic reaction, based on collision limitations (described in Chapter 8; also involved in the case study of Chapter 10)

• Reasoning with orders of magnitude and approximate relations among parameters (described in Chapter 5; applications provided in Chapter 6)

• A menu-driven database for the creation, modification, storage, and retrieval of objects and data.

A.2.4. DATABASE

Among all the facilities that were developed (listed in Section A.2.3 above), the last facility is the only one whose conceptual aspects were not discussed extensively in the main body of the thesis (except for Section 4.3). The reason for this is that the construction of the database did not involve any conceptual breakthroughs.

We will argue below, however, that the database and the way it was implemented did play a significant role in this work.
A.2.4.1. Justification of the Database

The existence of the database was instrumental in the success of the other methodologies and program modules. The database provided, for example, the testing ground for pathway synthesis methodologies. By testing our initial methodologies in a rich database, we were able to refine both the problem formulation and the solution strategy. In the group contribution method, the flexibility of the database allowed us to easily add new data and reevaluate the contributions of groups and the accuracy of the method.

It should be also emphasized that the programming style we have used (LISP and Object-Oriented Programming) allowed radically different actions, methods, and strategies to be applied to the same original representation of the objects and data in the database. A database constructed with a traditional DataBase Management System could not play the same significant role in the conceptual development of methods and their implementation.

A.2.4.2. Character of the Database

The database includes information on functional groups, metabolites (of small molecular weight), biochemical reactions, and biochemical pathways. In the database, and throughout this Appendix, metabolites will also called compounds or molecules, with the understanding that compounds of high molecular weight (such as DNA or enzymes) were not included in the implemented version of the database and will not be discussed here.

The data that the current implementation can deal with include stoichiometries of molecules (in terms of groups) and reactions (in terms of molecules), Gibbs Energies,
equilibrium constants, Enzyme Commission numbers, multiple names, etc. As it is expected that a lot of the information needed for a particular task must be acquired and added to the database specifically to carry out the task at hand, special attention was given to the extensibility of the database.

The database will be discussed further in Section A.3 (with respect to the computer files it entails), Section A.4.6 (with respect to its structure and use), and Section A.5.1 (with respect to the internals of the implementation).
A.3. PROGRAM MODULES

The developed computer-programs comprise of a number of distinct software modules. The modules that refer to biochemical systems are connected through common (or interconvertible) schemes for knowledge representation and a shared interface, so that each module can use methods from others, and the user can freely leap from one module to another.

These are the modules into which all the code and data are organized (Tables A.1 through A.5):

- The DATA MANIPULATION module is used for editing the database. Its files, contained in the directory SLIME:>michael>bio>CHUNKS>.*.*, are described in Table A.1.

- The DATA APPLICATION module allows the estimation of maximum rates of bioreactions and synthesis of biochemical pathways. Its files, contained in the directory SLIME:>michael>bio>MAXRAT>.*.*, are described in Table A.2.

- The DATA STORAGE module provides files for storing the information of the database. Its files, contained in the directory SLIME:>michael>bio>DB>.*.*, are described in Table A.3.

- The DRAWING module supplies facilities for drawing biochemical pathways. Its files, contained in the directory SLIME:>michael>bio>FRAME>.*.*, are described in Table A.4.
• The MAGNITUDE module implements formal Order-of-Magnitude reasoning. Its files, contained in the directory SLIME:michael:OM:.*, are described in Table A.5.

The modules are not completely independent from each other. DATA APPLICATION, for example, assumes that DATA MANIPULATION is present and that DATA STORAGE files have been loaded. The interdependencies of the modules are depicted in Figure A.1.
Table A.1:

Description of files in the DATA MANIPULATION module, for editing the database. The files are contained in the directory SLIME:~michael~bio~CHUNKS~*

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>base-object.lisp</td>
<td>Basic methods and functions applicable to all objects.</td>
</tr>
<tr>
<td>checking-commands.lisp</td>
<td>Functions that perform simple consistency checks on the database.</td>
</tr>
<tr>
<td>commands.lisp</td>
<td>General commands.</td>
</tr>
<tr>
<td>composition.lisp</td>
<td>Functions for manipulating data-structures representing the stoichiometries of objects.</td>
</tr>
<tr>
<td>data-modification-commands.lisp</td>
<td>Functions and commands for modifying the data of the database</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>data-points.lisp</td>
<td>Data-structures and functions for manipulating the data for the group-contribution method</td>
</tr>
<tr>
<td>db.lisp</td>
<td>Information on elements and bonds; this information cannot be modified from the interface.</td>
</tr>
<tr>
<td>frame.lisp</td>
<td>Creation of the frame (and its panes) for the graphic interface of the database.</td>
</tr>
<tr>
<td>gc-use.lisp</td>
<td>Functions for the application of the group-contribution method, i.e., for the estimation of</td>
</tr>
<tr>
<td></td>
<td>equilibrium constants.</td>
</tr>
<tr>
<td>group-contribution.lisp</td>
<td>The core of the group-contribution method performing the regression.</td>
</tr>
<tr>
<td>info-commands.lisp</td>
<td>Commands that give access to information stored in the database.</td>
</tr>
<tr>
<td>init-commands.lisp</td>
<td>Commands and functions executing at the beginning for initializing the database.</td>
</tr>
</tbody>
</table>

continued
Table A.1, continued

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>loadlast.lisp</td>
<td>Presentation translators that achieve mouse sensitivity. Must always be loaded last.</td>
</tr>
<tr>
<td>objects.lisp</td>
<td>Descriptions of the main kinds of objects, i.e., groups, molecules, reactions, and pathways</td>
</tr>
<tr>
<td>okeying-commands.lisp</td>
<td>Commands for indicating which data are acceptable and which should not be saved.</td>
</tr>
<tr>
<td>operations-on-objects.lisp</td>
<td>Methods for the main kinds of objects</td>
</tr>
<tr>
<td>pr-types.lisp</td>
<td>Definition of presentation types.</td>
</tr>
<tr>
<td>read-write-commands</td>
<td>Commands for reading the database from files and saving back into files.</td>
</tr>
<tr>
<td>small-objects.lisp</td>
<td>Data structures for bonds and elements.</td>
</tr>
<tr>
<td>switch-commands.lisp</td>
<td>Commands for changing modes while editing the database (e.g., changing current object).</td>
</tr>
<tr>
<td>synth-objects.lisp</td>
<td>Simple objects specific to the synthesis of pathways</td>
</tr>
<tr>
<td>vars.lisp</td>
<td>Definition of global variables, as well as simple functions and macros.</td>
</tr>
</tbody>
</table>
Table A.2:

Description of files in the DATA APPLICATION module, for estimation of maximum rates of bioreactions and synthesis of biochemical pathways. The files are contained in the directory SLIME:~michael~bio~MAXRAT~*

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>building.lisp</td>
<td>Functions for building data structures necessary for the estimation of maximum rates</td>
</tr>
<tr>
<td>commands.lisp</td>
<td>Commands making available to the interface the high-level functionality of this module</td>
</tr>
<tr>
<td>equation-handling.lisp</td>
<td>Functions for creating and manipulating maximum-rate equations.</td>
</tr>
<tr>
<td>global-methods.lisp</td>
<td>Methods for maximum-rate calculations.</td>
</tr>
<tr>
<td>local-objects.lisp</td>
<td>Facilities for creating copies of objects, so that the database itself is not corrupted.</td>
</tr>
<tr>
<td>reversible-synthesis.lisp</td>
<td>The core of the pathway-synthesis method.</td>
</tr>
<tr>
<td>variables.lisp</td>
<td>Definitions of variables and simple functions</td>
</tr>
</tbody>
</table>
Table A.3:

Description of files in the DATA STORAGE module, for saving the database. The files are contained in the directory SLIME:>michael>bio>DB>*.*

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP.lisp</td>
<td>All data on groups, other than their contributions to the Gibbs Energy.</td>
</tr>
<tr>
<td>group-contributions-base.lisp</td>
<td>The contributions of groups to the Gibbs Energy, estimated by regression.</td>
</tr>
<tr>
<td>MOLECULE.lisp</td>
<td>All data on small biomolecules (metabolites).</td>
</tr>
<tr>
<td>PATHWAY.lisp</td>
<td>All data on biochemical pathways.</td>
</tr>
<tr>
<td>REACTION.lisp</td>
<td>All data on bioreactions.</td>
</tr>
</tbody>
</table>
Table A.4:
Description of files in the DRAWING module,
for drawing pathways. The files are contained
in the directory   SLIME:~michael~blo~FRAME~*:.*:

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>command.lisp</td>
<td>Definitions of commands.</td>
</tr>
<tr>
<td>connectives.lisp</td>
<td>Definition and manipulation of auxiliary objects such as arcs.</td>
</tr>
<tr>
<td>constants.lisp</td>
<td>Definition of constants.</td>
</tr>
<tr>
<td>crecon.lisp</td>
<td>Functions creating and connecting icons.</td>
</tr>
<tr>
<td>flavors.lisp</td>
<td>Basic data-structures behind the icons</td>
</tr>
<tr>
<td>icons.lisp</td>
<td>Graphic features of the icons.</td>
</tr>
<tr>
<td>imageneric.lisp</td>
<td>Functions for the manipulation of images.</td>
</tr>
<tr>
<td>loadlast.lisp</td>
<td>Compilation of methods for more efficient execution.</td>
</tr>
<tr>
<td>ops.lisp</td>
<td>Graphic operations on icons.</td>
</tr>
<tr>
<td>windows.lisp</td>
<td>Definitions of windows and panes.</td>
</tr>
</tbody>
</table>
Table A.5:
Description of files in the MAGNITUDE module,
for Order-of-Magnitude reasoning. The files are contained
in the directory SLIME:=michael:=OM=*.*.*

<table>
<thead>
<tr>
<th>FILE NAME</th>
<th>DESCRIPTION OF FILE CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>infer.lisp</td>
<td>Functions and methods performing the core of the Inferences</td>
</tr>
<tr>
<td>infer-base.lisp</td>
<td>Storage of inference methods that can be memoized, such as tables</td>
</tr>
<tr>
<td></td>
<td>for the transitivity properties of relations</td>
</tr>
<tr>
<td>object-atms.lisp</td>
<td>Objects for assumption-based truth maintenance.</td>
</tr>
<tr>
<td>objects.lisp</td>
<td>Definitions of objects like relations, constraints, etc.</td>
</tr>
<tr>
<td>phys-base.lisp</td>
<td>Information on physical dimensionality of quantities</td>
</tr>
<tr>
<td>rel-base.lisp</td>
<td>Definitions of order-of-magnitude relations</td>
</tr>
<tr>
<td>utilities.lisp</td>
<td>Variables and simple functions</td>
</tr>
</tbody>
</table>
Figure A.1:
Interdependencies among modules.

An arrow from module X to module Y indicates that X depends on Y.
A.4. USING THE PROGRAMS

In this section, we provide instructions on how the programs can be used. The internals of the implementation will be examined in Section A.5.

A.4.1. LOADING THE CODE

The DATA STORAGE module can be loaded from the interface (Section A.4.5) which becomes accessible after loading the DATA MANIPULATION module. In order to load any other module, one simply loads a special file with the name "loader.lisp" residing in the directory of the module, by issuing (in a Lisp Listener of the Symbolics machine) the command:

Load File  <name of the file>

Each loader file contains lisp forms that load the rest of the files in that directory, in the correct order. Note that before loading the desired module, one must load all the modules on which the desired module depends (except for the DATA STORAGE module which can be loaded through the interface). Consequently, to load the DATA APPLICATION module, one issues two commands:

Load File  SLIME::michael>bio>chunks>loader.lisp

Load File  SLIME::michael>bio>maxrat>loader.lisp

A.4.2. PREPARATION OF THE GRAPHIC INTERFACE

Once the programs are loaded (Section A.4.1), the graphic interface for the retrieval and storage of data and the analysis and synthesis of biochemical pathways
(i.e., for modules DATA STORAGE, DATA MANIPULATION, and DATA APPLICATION) is invoked by pressing:

<Select>  <Triangle>

By <Select> and <Triangle> we refer to particular keys of the Symbolics keyboard; the <Select> key is marked as such, while the <Triangle> key is marked with the picture of a triangle. The invocation can take place from any window or process and regardless of concurrent execution of other forms in the same process, because <Select> corresponds to an asynchronous character.

With a delay of a few seconds, a frame consisting of an assortment of panes appears, and a pop-up menu offers to initialize everything. After one clicks the left mouse-button in the "Yes" box of the pop-up menu, the program reads and updates the entire database. Through the rest of this document, whenever we refer to mouse-clicking but do not specify which mouse-button is clicked, we will always mean the left mouse-button.

The database is now in a clean state, and the graphic-interface frame is as depicted in Figure A.2.
Figure A.2:

Initial form of the interface frame
A.4.3. WINDOW PANES

In the configuration of Figure A.2, there are in total 7 panes:

- The Logo pane, which is the small, square pane bearing the sign "BioPath", has no particular function.

- The Mode Indicator pane, immediately to the right of the Logo pane, allows the selection of the mode of operation, which corresponds to the type of objects that can be edited (as will be discussed in Section A.4.4, below).

- The Commander pane, immediately below the Logo and Mode Indicator panes, allows the mouse-driven invocation of commands (described in Section A.4.5). Its purpose is to make the use of the commands easier and faster. The same commands can alternatively be input, through the keyboard, into the interaction pane.

- The Interaction pane, at the bottom left corner of the frame (i.e., below the Commander pane), receives commands (described in Section A.4.5) and LISP forms, and displays the values returned by their execution.

- The List Chooser pane, at the bottom right section of the frame, displays (and makes mouse-sensitive) all available objects that can be used in the composition (stoichiometry) of the object being edited. For example, if a reaction is being edited, all molecules are displayed in the List Chooser.

- The Object Values pane, at the top right section, provides for the display and modification of various attributes of objects. These
attributes include the acceptable and preferred names of the object, Enzyme Commission numbers, Gibbs Energies, and equilibrium constants.

- The Display pane, at the top left corner, displays the name of the object being edited, and its current compositions (stoichiometries).

A.4.4. MODES OF OPERATION

The mode of operation determines what kind of objects are allowed to be edited. It can be selected by simply clicking on the desired mode in the Mode Indicator pane. However, only the modes "Group", "Molecule", "Reaction", and "Pathway" are currently supported.

When a mode is selected, the List Chooser which is initially empty (Figure A.2) displays all the objects that can be used in the composition of objects of the current type. Thus, in Group mode all elements and bonds are displayed, in Molecule mode all groups are displayed, in Reaction mode all molecules are displayed (Figure A.3), and in Pathway mode all reactions are displayed (Figure A.4).

The scrolling bar on the left side of the pane can be used to scroll through the long lists displayed.
A.4.5. COMMANDS

One can invoke a command either by typing the command's name into the Interaction pane or by clicking on the command in the Commander pane. The commands, which are all present independently of the mode of operation, are listed and explained in Table A.6.

Some of the commands require additional arguments to be supplied. These can be symbols (words) to be typed, or objects from the database. In the latter case, they are normally objects displayed in the List Chooser pane. One can select an object from the List Chooser as an argument to a command by simply clicking on it.
Table A.6: Commands for the manipulation of the database

<table>
<thead>
<tr>
<th>COMMAND NAME</th>
<th>EXPLANATION OF THE COMMAND'S ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add Set Of Species</td>
<td>Prompts for the name of an object, and adds the stoichiometry of that object to the stoichiometry of the current object being edited.</td>
</tr>
<tr>
<td>Add To Composition</td>
<td>Prompts for an object and adds that object to the stoichiometry (composition) of the current object.</td>
</tr>
<tr>
<td>Backward Pointers</td>
<td>Updates the backward pointers, i.e., the entries in the &quot;participates-in&quot; slots, of all objects.</td>
</tr>
<tr>
<td>Calculate Rate</td>
<td>Prompts for a reaction and estimates its maximum rate.</td>
</tr>
<tr>
<td>Change Configuration</td>
<td>Prompts for the name of one of the available configurations, and installs it as the current configuration of the program frame.</td>
</tr>
<tr>
<td>Change Current Object</td>
<td>Prompts for the name of an object, and installs it as the current object being edited.</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>COMMAND NAME</th>
<th>EXPLANATION OF THE COMMAND'S ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Check Everything</td>
<td>Checks for simple inconsistencies, such as duplicate names of objects in the database.</td>
</tr>
<tr>
<td>Copy Into New Object</td>
<td>Prompts for a new name, creates an object with that name, copies the attributes current object into the new object, and installs the newly created object as the current object.</td>
</tr>
<tr>
<td>Create Solo Molecule</td>
<td>Prompts for a name and creates a molecule, with that name, which will not be broken down into groups.</td>
</tr>
<tr>
<td>Defining Composition</td>
<td>Displays the type of objects that can be used in the composition (stoichiometry) of the current object being edited.</td>
</tr>
<tr>
<td>Delete Object</td>
<td>Prompts for the name of an object, and deletes it from the database.</td>
</tr>
<tr>
<td>Empty Composition</td>
<td>Empties (nullifies) the stoichiometry (composition) of the current object.</td>
</tr>
<tr>
<td>COMMAND NAME</td>
<td>EXPLANATION OF THE COMMAND'S ACTION</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Initialize Everything</td>
<td>Kills the in-memory structure of the database and reads the version that is stored in the files.</td>
</tr>
<tr>
<td>Install Mouse Sensitivity</td>
<td>Makes all objects mouse sensitive.</td>
</tr>
<tr>
<td>Reaction From EC Number</td>
<td>Prompts for an Enzyme Commission number and displays the name of the reaction that corresponds to that number.</td>
</tr>
<tr>
<td>Read All Summaries</td>
<td>Reads the database from the files.</td>
</tr>
<tr>
<td>Show All Object Names</td>
<td>Prompts for a name of an object, and displays all the names of that object.</td>
</tr>
<tr>
<td>Show Composition</td>
<td>Prompts for a name of an object, and displays the (primary) composition of that object</td>
</tr>
<tr>
<td>Show EC Number</td>
<td>Prompts for a name of a reaction, and displays its Enzyme Commission number</td>
</tr>
<tr>
<td>Subtract From Composition</td>
<td>Prompts for an object, and subtracts it from the stoichiometry (composition) of the current object being edited</td>
</tr>
</tbody>
</table>

---

continued
Table A.6, continued

<table>
<thead>
<tr>
<th>COMMAND NAME</th>
<th>EXPLANATION OF THE COMMAND'S ACTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthesize Pathways</td>
<td>Provides a menu for the selection of synthesis specifications, and synthesizes all pathways that satisfy the specifications</td>
</tr>
<tr>
<td>Write All Summaries</td>
<td>Saves the in-memory version database into files — in effect saving all the modifications that have been made since the last time the database was read from files.</td>
</tr>
</tbody>
</table>
A.4.6. CREATING AND EDITING DATABASE OBJECTS

In this section we will discuss how the interface can be used to display information on objects that already exist in the database, edit the objects, and create new objects. The last task is accomplished either manually (in most cases) or, for biochemical pathways only, automatically through pathway synthesis.

A.4.6.1. Data Attributes

For each object, a series of attributes are stored in the database, and can be edited through the interface. We list and explain data attributes of groups in Table A.7, data attributes of molecules in Table A.8, data attributes of reactions in Table A.9, and data attributes of pathways in Table A.10.
Table A.7:
Data Stored for Groups.

<table>
<thead>
<tr>
<th>Data Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BONDING-COMPOSITION</td>
<td>free bonds of the group</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>Gibbs Energy contribution</td>
</tr>
<tr>
<td>ELEMENTARY-COMPOSITION</td>
<td>elements contained in the group</td>
</tr>
<tr>
<td>GC-PIVOTING-COMPOSITION</td>
<td>composition of the group in terms of simpler groups — only for composite groups</td>
</tr>
<tr>
<td>GENERALIZED</td>
<td>more general groups</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>list of all names</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name</td>
</tr>
<tr>
<td>SOLITARYYP</td>
<td>boolean indicating whether the group is contained in only one molecule</td>
</tr>
<tr>
<td>SPECIALIZED</td>
<td>specializations of the group</td>
</tr>
</tbody>
</table>
Table A.8:  
Data Stored for Molecules.

<table>
<thead>
<tr>
<th>Data Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAD-DATA-POINTS</td>
<td>data points that are not considered reliable</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>Gibbs Energy data</td>
</tr>
<tr>
<td>GROUP-COMPOSITION</td>
<td>groups contained in the molecule</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>all available names</td>
</tr>
<tr>
<td>OTHER-CONFORMATIONS</td>
<td>other conformations of the molecule</td>
</tr>
<tr>
<td>OTHER-DATA</td>
<td>miscellaneous data (other than Gibbs Energies)</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name used for the molecule</td>
</tr>
<tr>
<td>Data Attribute</td>
<td>Explanation</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>BACKWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the reaction in the direction opposite to its nominal direction</td>
</tr>
<tr>
<td>BAD-DATA-POINTS</td>
<td>data considered unreliable</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>data on the Gibbs Energy of reaction</td>
</tr>
<tr>
<td>EC-NUMBER</td>
<td>the Enzyme Commission number for the reaction</td>
</tr>
<tr>
<td>FORWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the reaction in its nominal direction</td>
</tr>
<tr>
<td>K-EQ</td>
<td>data on the equilibrium constant of the reaction</td>
</tr>
<tr>
<td>MOLECULAR-COMPOSITION</td>
<td>molecules participating in the reaction</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>all recognizable names for the reaction</td>
</tr>
<tr>
<td>OTHER-DATA</td>
<td>miscellaneous data (other than Gibbs Energies or equilibrium constants)</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name used for the reaction</td>
</tr>
<tr>
<td>TRUE-ENZYME-P</td>
<td>boolean indicating whether the reaction describes a true enzyme or just a transformation</td>
</tr>
</tbody>
</table>
Table A.10:

Data Stored for Pathways.

<table>
<thead>
<tr>
<th>Data</th>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>BACKWARD-MAXIMUM-RATE</strong></td>
<td>maximum rate estimate for the pathway in the direction opposite to its nominal direction</td>
</tr>
<tr>
<td></td>
<td><strong>COMMENT</strong></td>
<td>comment</td>
</tr>
<tr>
<td></td>
<td><strong>FORWARD-MAXIMUM-RATE</strong></td>
<td>maximum rate estimate for the pathway in its nominal direction</td>
</tr>
<tr>
<td></td>
<td><strong>NAME-LIST</strong></td>
<td>all recognizable names for the pathway</td>
</tr>
<tr>
<td></td>
<td><strong>REACTION-COMPOSITION</strong></td>
<td>reactions participating in the pathway</td>
</tr>
<tr>
<td></td>
<td><strong>SINGLE-NAME</strong></td>
<td>default name used for the pathway</td>
</tr>
</tbody>
</table>
A.4.6.2. Display of Information on Objects

In order to edit an object, we first select the appropriate mode of operation; at this point the List Chooser pane changes its display, as discussed in Section A.4.4. Then, we use the command "Change Current Object" and name the object we want to inspect or edit. Immediately, the Object Values and Display panes change, as shown in Figure A.3 (for editing the reaction \textit{Fructose-Diphosphate-Aldolase}) and Figure A.4 (for editing the pathway of glycolysis). Specifically:

- The Display pane shows:
  - Weight (relevant only for groups and molecules)
  - Charge (relevant only for groups and molecules)
  - Stoichiometries of the object

- The Object Values pane shows:
  - Names
  - Preferred Name
  - Comment
  - True Enzyme (relevant only for reactions)
  - EC numbers (relevant only for reactions)
  - Gibbs Energy data (relevant only for molecules, reactions, and pathways)
  - Equilibrium Constant data (relevant only for reactions and pathways)
  - Other data
If we want a pathway drawn on the screen, we issue, from the keyboard, the command "Draw Pathway". A pop-up graphic display of the pathway appears; the display vanishes as soon as we press any key.

A.4.6.3. Editing of Objects

Now we can edit all the attributes of the object, as follows:

- Maximum rate data can be recalculated and stored through the *Calculate Rate* command.

- The stoichiometries can be changed through four related commands:
  ◇ Add To Composition
  ◇ Subtract From Composition
  ◇ Add Set of Species
  ◇ Empty Composition

- One can edit any of the other attributes by clicking on their values of the Object Values pane, and typing in new values.
Figure A.3:
Displaying and Editing the reaction Fructose-Diphosphate-Aldolase
Figure A.4:
Displaying and editing the glycolytic pathway
A.4.6.4. Creation of New Objects

In order to create new objects we can use one of three commands. Specifically, we can:

- Invoke the Change Current Object command and give a name that is not currently used by any object.

- Invoke the Copy Into New Object command to copy the current object into a new object.

- Invoke the Synthesize Pathways command and provide stoichiometric specifications (by filling in the pop-up menu) to create all pathways satisfying a set of specifications.

A.4.7. ORDER-OF-MAGNITUDE REASONING

The implementation of O[M] is not accessible through any graphic interface. Thus, its use requires significant knowledge of LISP and the Symbolics Genera environment.

After the MAGNITUDES module is loaded, use of O[M] entails preparation (in Zmacs) of the specification of the problem, which is then loaded or compiled so that O[M] can perform the reasoning.

The results cannot be displayed because a large number of relations is typically produced. To explore the results, the user must type appropriate LISP forms or use the Inspector of the Symbolics Genera environment.

The use of O[M] will be clarified in the description of its internal structure (Section A.5.2).
A.5. INSIDE THE PROGRAMS

In this section we will describe the internal representations used in the implementation. As our programming style is Object Oriented Programming, we will describe the attributes (also called slots or instance variables) of the each kind of objects, the interconnections among various kinds of objects, and the methods that operate on each kind of objects.

A.5.1. BIOCHEMICAL DATABASE OBJECTS

The instance variables (attributes) of the objects are described in:

- Table A.11 for groups
- Table A.12 for molecules
- Table A.13 for bioreactions
- Table A.14 for pathways

The methods operating on the objects are described in:

- Table A.15 for groups
- Table A.16 for molecules
- Table A.17 for bioreactions
- Table A.18 for pathways

To further clarify the meaning of the attributes of objects, we list the contents of the slots, for three example-objects:
- The slot values of the molecule Fructose-Diphosphate are described in Table A.19.

- The slot values of the bioreaction Fructose-Diphosphate-Aldolase are described in Table A.20.

- The slot values of the pathway of glycolysis are described in Table A.21.

Finally, the interrelations among objects are illustrated in Figure A.5
Table A.11:

Flavor Instance Variables for Groups. For each instance variable there are methods to read, write, and locate it. For example, for the instance variable DG-CONTRIBUTION there are three methods: DG-CONTRIBUTION, reading the instance variable, (LOCF DG-CONTRIBUTION), locating the instance variable, and (SETF DG-CONTRIBUTION), writing the instance variable.

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BONDING-COMPOSITION</td>
<td>free bonds of the group</td>
</tr>
<tr>
<td>CHARGE</td>
<td>normal charge of the group under aqueous conditions, pH=7</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>name of the class of the object (i.e. &quot;GROUP&quot;)</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>Gibbs Energy contribution</td>
</tr>
<tr>
<td>ELEMENTARY-COMPOSITION</td>
<td>elements contained in the group</td>
</tr>
<tr>
<td>GC-PIVOTING-COMPOSITION</td>
<td>composition of the group in terms of simpler groups — only for composite groups</td>
</tr>
<tr>
<td>GENERALIZED</td>
<td>more general groups</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LONG-NAME</td>
<td>most formal name</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>list of all names</td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>package in which the object belongs</td>
</tr>
<tr>
<td>PARTICIPATES-IN</td>
<td>molecules in which the group participates</td>
</tr>
<tr>
<td>FLAVOR:PROPERTY-LIST</td>
<td>miscellaneous properties</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>shortest name</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name</td>
</tr>
<tr>
<td>SOLITARYP</td>
<td>boolean indicating whether the group is contained in only one molecule</td>
</tr>
<tr>
<td>SPECIALIZED</td>
<td>specializations of the group</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>name generated by the system</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>weight of the group</td>
</tr>
</tbody>
</table>
Table A.12:

Flavor Instance Variables for Molecules. For each instance variable there are methods to read, write, and locate it. For example, for the instance variable DG-CONTRIBUTION there are three methods: DG-CONTRIBUTION, reading the instance variable, (LOCF DG-CONTRIBUTION), locating the instance variable, and (SETF DG-CONTRIBUTION), writing the instance variable.

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAD-DATA-POINTS</td>
<td>data points that are not considered reliable</td>
</tr>
<tr>
<td>BONDING-COMPOSITION</td>
<td>bonds connecting groups within the molecule</td>
</tr>
<tr>
<td>CHARGE</td>
<td>normal charge of the molecule, under aqueous conditions, pH=7</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>name of the class of the object (i.e., &quot;MOLECULE&quot;)</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DATA-POINTS</td>
<td>all thermodynamic data points</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>Gibbs Energy data</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>Gibbs Energy estimated from the contributions of groups</td>
</tr>
<tr>
<td>DG-VALUES</td>
<td>all Gibbs Energy values (from the literature as well as estimated)</td>
</tr>
<tr>
<td>ELEMENTARY-COMPOSITION</td>
<td>molecular formula of the molecule</td>
</tr>
<tr>
<td>GROUP-COMPOSITION</td>
<td>groups contained in the molecule</td>
</tr>
</tbody>
</table>

continued
Table A.12, continued

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LONG-NAME</td>
<td>most formal name</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>all available names</td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>package to which the object belongs</td>
</tr>
<tr>
<td>OTHER-CONFORMATIONS</td>
<td>other conformations of the molecule</td>
</tr>
<tr>
<td>OTHER-DATA</td>
<td>miscellaneous data (other than Gibbs Energies)</td>
</tr>
<tr>
<td>PARTICIPATES-IN</td>
<td>reactions in which the molecule participates</td>
</tr>
<tr>
<td>FLAVOR:PROPERTY-LIST</td>
<td>other properties</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>shortest name</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name used for the molecule</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>unique name for the molecule (generated by the system)</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>molecular weight</td>
</tr>
</tbody>
</table>
Table A.13:
Flavor Instance Variables for Reactions. For each instance variable there are methods to read, write, and locate it. For example, for the instance variable DG-CONTRIBUTION there are three methods:
DG-CONTRIBUTION, reading the instance variable, (LOCF DG-CONTRIBUTION), locating the instance variable, and (SETF DG-CONTRIBUTION), writing the instance variable.

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACKWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the reaction in the direction opposite to its nominal direction</td>
</tr>
<tr>
<td>BAD-DATA-POINTS</td>
<td>data considered unreliable</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>name of the class of the object (i.e., &quot;REACTION&quot;)</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DATA-POINTS</td>
<td>all thermodynamic data points</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>data on the Gibbs Energy of reaction</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>Gibbs Energy of reaction estimated from the contributions of groups</td>
</tr>
<tr>
<td>DG-VALUES</td>
<td>all Gibbs Energy values (from the literature as well as estimated)</td>
</tr>
</tbody>
</table>

continued
Table A.13, continued

<table>
<thead>
<tr>
<th>Instance</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC-NUMBER</td>
<td>the Enzyme Commission number for the reaction</td>
<td></td>
</tr>
<tr>
<td>FORWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the reaction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In its nominal direction</td>
<td></td>
</tr>
<tr>
<td>GROUP-COMPOSITION</td>
<td>groups with a net participation in the reaction</td>
<td></td>
</tr>
<tr>
<td>K-EQ</td>
<td>data on the equilibrium constant of the reaction</td>
<td></td>
</tr>
<tr>
<td>LONG-NAME</td>
<td>most formal name of the reaction</td>
<td></td>
</tr>
<tr>
<td>MOLECULAR-COMPOSITION</td>
<td>molecules participating in the reaction</td>
<td></td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>all recognizable names for the reaction</td>
<td></td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>package to which the object belongs</td>
<td></td>
</tr>
<tr>
<td>OTHER-DATA</td>
<td>miscellaneous data (other than Gibbs Energies or equilibrium constants)</td>
<td></td>
</tr>
<tr>
<td>PARTICIPATES-IN</td>
<td>pathways in which the reaction participates</td>
<td></td>
</tr>
<tr>
<td>FLAVOR:PROPERTY-LIST</td>
<td>other properties</td>
<td></td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>shortest name for the reaction</td>
<td></td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name used for the reaction</td>
<td></td>
</tr>
<tr>
<td>TRUE-ENZYME-P</td>
<td>boolean indicating whether the reaction describes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>a true enzyme or just a transformation</td>
<td></td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>unique name (generated automatically)</td>
<td></td>
</tr>
</tbody>
</table>
**Table A.14:**

Flavor Instance Variables for Pathways. For each instance variable there are methods to read, write, and locate it. For example, for the instance variable DG-CONTRIBUTION there are three methods: DG-CONTRIBUTION, reading the instance variable, (LOCF DG-CONTRIBUTION), locating the instance variable, and (SETF DG-CONTRIBUTION), writing the instance variable.

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACKWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the pathway in the direction opposite to its nominal direction</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>name of the class of the object (i.e., &quot;PATHWAY&quot;)</td>
</tr>
<tr>
<td>COMMENT</td>
<td>comment</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>Gibbs Energy for the transformation represented by the pathway, estimated from the contributions of groups</td>
</tr>
<tr>
<td>FORWARD-MAXIMUM-RATE</td>
<td>maximum rate estimate for the pathway in its nominal direction</td>
</tr>
<tr>
<td>GROUP-COMPOSITION</td>
<td>groups with a net participation in the pathway</td>
</tr>
<tr>
<td>LONG-NAME</td>
<td>most formal name of the pathway</td>
</tr>
<tr>
<td>MOLECULAR-COMPOSITION</td>
<td>molecules with a net participation in the pathway</td>
</tr>
<tr>
<td>MOLECULAR-INTERMEDIATES</td>
<td>intermediates of the pathway</td>
</tr>
</tbody>
</table>

*continued*
Table A.14, continued

<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAME-LIST</td>
<td>all recognizable names for the pathway</td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>package to which the object belongs</td>
</tr>
<tr>
<td>REACTION-COMPOSITION</td>
<td>reactions participating in the pathway</td>
</tr>
<tr>
<td>FLAVOR:PROPERTY-LIST</td>
<td>other properties</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>shortest name for the pathway</td>
</tr>
<tr>
<td>SINGLE-NAME</td>
<td>default name used for the pathway</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>unique name (generated automatically)</td>
</tr>
</tbody>
</table>
Table A.15:  
Flavor Methods for Groups.
In addition to the listed methods, there are methods to read, write, and locate each instance variable (Table A.11).

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL-COMPOSITIONS</td>
<td>provides the elements and the free bonds of the group</td>
</tr>
<tr>
<td>ALL-GROUP-DATA</td>
<td>provides all thermodynamic data on molecules that include the group and reactions that transform the group</td>
</tr>
<tr>
<td>CONNECTING-P</td>
<td>checks whether the group contains free bonds</td>
</tr>
<tr>
<td>CONSUMED-BY-REACTIONS</td>
<td>returns all reactions that consume the group</td>
</tr>
<tr>
<td>:DESCRIBE</td>
<td>describes all attributes of the group</td>
</tr>
</tbody>
</table>

continued
Table A.15, continued

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>DO-GENERALIZED</td>
<td>updates all generalizations of this group,</td>
</tr>
<tr>
<td></td>
<td>to make them include this group as a specialization</td>
</tr>
<tr>
<td>DO-SPECIALIZED</td>
<td>updates all specializations of this group,</td>
</tr>
<tr>
<td></td>
<td>to make them include this group as a generalization</td>
</tr>
<tr>
<td>GET</td>
<td>reads a property</td>
</tr>
<tr>
<td>GET-A-NAME</td>
<td>returns a name of the object</td>
</tr>
<tr>
<td></td>
<td>(depending on what names are available — at worst, the UNIQUE-NAME)</td>
</tr>
<tr>
<td>NON-RINGP</td>
<td>checks whether the group is an open-chain group</td>
</tr>
<tr>
<td>NUMBER-OF-ATOMS</td>
<td>returns the number of atoms contained in the group</td>
</tr>
<tr>
<td>PARTICIPATES-IN-REACTIONS</td>
<td>returns the reactions in which the group</td>
</tr>
<tr>
<td></td>
<td>has a net participation (i.e., net consumption or production)</td>
</tr>
<tr>
<td>PRIMARY-COMPOSITION</td>
<td>returns the elementary-composition of the group</td>
</tr>
<tr>
<td>(SETF PRIMARY-COMPOSITION)</td>
<td>writes the elementary-composition</td>
</tr>
<tr>
<td>PRODUCED-BY-REACTIONS</td>
<td>returns the reactions which involve</td>
</tr>
<tr>
<td></td>
<td>a net production of the group</td>
</tr>
<tr>
<td>:PUTPROP</td>
<td>sets a property in the property-list</td>
</tr>
<tr>
<td>:REMPROP</td>
<td>removes a property from the property-list</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>REPLACE-OBJECT-BY-COMPOSITION</td>
<td>substitute a set of other groups for the current group, wherever the current group participates</td>
</tr>
<tr>
<td>RINGP</td>
<td>checks whether the group is a ring group</td>
</tr>
<tr>
<td>SECONDARY-COMPOSITION</td>
<td>returns the bonding-composition of the group</td>
</tr>
<tr>
<td>(SETF SECONDARY-COMPOSITION)</td>
<td>sets the bonding-composition of the group</td>
</tr>
<tr>
<td>SORT-ITEM</td>
<td>given another group, it decides whether the current group should be placed before the other group in ordered lists</td>
</tr>
<tr>
<td>SPECIALIZE-GROUP</td>
<td>creates a specialization of the group</td>
</tr>
<tr>
<td>TERMINATORP</td>
<td>checks whether the group contains exactly one free bond</td>
</tr>
<tr>
<td>TOTAL-BONDS</td>
<td>returns the total number of free bonds of the group</td>
</tr>
<tr>
<td>UNBIND-NAMES</td>
<td>unnames the group — if it is to be deleted</td>
</tr>
<tr>
<td>UNDO-GENERALIZED</td>
<td>disconnects a group from its generalizations</td>
</tr>
<tr>
<td>UNDO-SPECIALIZED</td>
<td>disconnects a group from its specializations</td>
</tr>
</tbody>
</table>
Table A.16:
Flavor Methods for Molecules.
In addition to the listed methods, there are methods to read, write, and locate each instance variable (Table A.12).

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD-TERM</td>
<td>add another group to the molecule</td>
</tr>
<tr>
<td>ALL-COMPOSITIONS</td>
<td>returns all compositions of the molecule (i.e., groups, molecules, or bonds contained)</td>
</tr>
<tr>
<td>CHANGE-CONFORMATION</td>
<td>choose a new conformation for the molecule</td>
</tr>
<tr>
<td>CONSUMED-BY</td>
<td>returns the reactions that consume the molecule</td>
</tr>
<tr>
<td>CONTAINS-DATA-P</td>
<td>checks whether the molecule contains any data</td>
</tr>
<tr>
<td>DELETE-CONFORMATION</td>
<td>removes one of the possible conformations</td>
</tr>
<tr>
<td>:DESCRIBE</td>
<td>provides a description of the molecule</td>
</tr>
<tr>
<td>:GET</td>
<td>retrieves a property from the property-list</td>
</tr>
<tr>
<td>GET-A-NAME</td>
<td>returns a name of the molecule</td>
</tr>
<tr>
<td></td>
<td>(depending on what names are available — at worst, the UNIQUE-NAME)</td>
</tr>
<tr>
<td>NON-RINGP</td>
<td>checks whether the molecule is an open-chain molecule</td>
</tr>
<tr>
<td>NUMBER-OF-ATOMS</td>
<td>returns the total number of atoms</td>
</tr>
<tr>
<td>PARTICIPATES-AS-INTERMEDIATE</td>
<td>returns the pathways in which the molecule is an intermediate (but not a net reactant or product)</td>
</tr>
</tbody>
</table>

continued
<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTICIPATES-IN-PATHWAYS</td>
<td>returns the pathways in which the molecule participates as a net reactant or product</td>
</tr>
<tr>
<td>PRIMARY-COMPOSITION</td>
<td>returns the group-composition of the molecule</td>
</tr>
<tr>
<td>(SETF PRIMARY-COMPOSITION)</td>
<td>sets the group-composition of the molecule</td>
</tr>
<tr>
<td>PRODUCED-BY</td>
<td>returns the reactions which produce the molecule</td>
</tr>
<tr>
<td>:PUTPROP</td>
<td>sets a property in the property-list</td>
</tr>
<tr>
<td>:REMPROP</td>
<td>removes a property from the property-list</td>
</tr>
<tr>
<td>REPLACE-OBJECT-BY-COMPOSITION</td>
<td>replaces the molecule, in the reactions and pathways in which it participates, by another set of molecules</td>
</tr>
<tr>
<td>RINGP</td>
<td>checks whether the molecule contains a ring</td>
</tr>
<tr>
<td>SORT-ITEM</td>
<td>given another molecule, the method decides whether the current molecule should precede the other molecule in ordered-lists</td>
</tr>
<tr>
<td>SUBSTITUTE-COMPOSITE-GROUPS-IN-OBJECT</td>
<td>breaks down any composite groups that the molecule contains into their constituent simpler groups</td>
</tr>
<tr>
<td>TOTAL-BONDS</td>
<td>total number of bonds interconnecting groups in the molecule</td>
</tr>
<tr>
<td>UNBIND-NAMES</td>
<td>cancel the names given to the molecule</td>
</tr>
<tr>
<td>UPDATE-DG-CONTRIBUTION</td>
<td>recalculate the Gibbs Energy from the contributions of groups and store it into dg-contribution</td>
</tr>
</tbody>
</table>
Table A.17:
Flavor Methods for Reactions.
In addition to the listed methods, there are methods to read, write, and locate each instance variable (Table A.13).

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD-TERM</td>
<td>adds another molecule to the stoichiometry of the reaction</td>
</tr>
<tr>
<td>ALL-COMPOSITIONS</td>
<td>provides all compositions of the reaction</td>
</tr>
<tr>
<td>ALL-DATA</td>
<td>all thermodynamic data available on the reaction</td>
</tr>
<tr>
<td>CONTAINS-DATA-P</td>
<td>checks whether the reaction contains any data</td>
</tr>
<tr>
<td>:DESCRIBE</td>
<td>describes the reaction</td>
</tr>
<tr>
<td>ESTIMATE-MAXIMUM-RATE</td>
<td>estimates the maximum rate of the reaction</td>
</tr>
<tr>
<td></td>
<td>in its nominal (forward) direction</td>
</tr>
<tr>
<td>ESTIMATE-REVERSE-MAXIMUM-RATE</td>
<td>estimates the maximum rate of the reaction in its reverse (backward) direction</td>
</tr>
<tr>
<td>:GET</td>
<td>retrieves a property of the reaction</td>
</tr>
<tr>
<td>GET-A-NAME</td>
<td>returns a name of the reaction</td>
</tr>
<tr>
<td></td>
<td>(depending on what names are available — at worst, the UNIQUE-NAME)</td>
</tr>
</tbody>
</table>

...continued
Table A.17, continued

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAXIMUM-RATE-VALUE</td>
<td>retrieves the value of the maximum rate of the reaction</td>
</tr>
<tr>
<td>PRIMARY-COMPOSITION</td>
<td>returns the molecular-composition of the reaction</td>
</tr>
<tr>
<td>(SETF PRIMARY-COMPOSITION)</td>
<td>sets the molecular-composition</td>
</tr>
<tr>
<td>:PUTPROP</td>
<td>sets a property in the property-list</td>
</tr>
<tr>
<td>:REMPROP</td>
<td>removes a property from the property-list</td>
</tr>
<tr>
<td>REPLACE-OBJECT-BY-COMPOSITION</td>
<td>replaces the reaction, in the pathways in which it participates, by another set of reactions</td>
</tr>
<tr>
<td>SMOOTH-PLUS-AND-MINUS-CHARGES</td>
<td>adjusts various attributes to account for the fact positive and negative charges can cancel out</td>
</tr>
<tr>
<td>SORT-ITEM</td>
<td>given another reaction, the method decides whether the current reaction should precede the other reaction in ordered-lists</td>
</tr>
<tr>
<td>SUBSTITUTE-COMPOSITE-GROUPS-IN-OBJECT</td>
<td>breaks down any composite groups that the reaction contains into their constituent simpler groups</td>
</tr>
<tr>
<td>UNBIND-NAMES</td>
<td>cancel the names given to the reaction</td>
</tr>
<tr>
<td>UPDATE-ALL-MAXRAT-EQUATIONS</td>
<td>constructs the equations describing the maximum rate of the reaction</td>
</tr>
<tr>
<td>UPDATE-DG-CONTRIBUTION</td>
<td>recalculates the Gibbs Energy from the contributions of groups and store it into dg-contribution</td>
</tr>
</tbody>
</table>
Table A.18:
Flavor Methods for Pathways.
In addition to the listed methods, there are methods to read, write, and locate each instance variable (Table A.14).

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD-TERM</td>
<td>adds another reaction to the pathway</td>
</tr>
<tr>
<td>ALL-COMPOSITIONS</td>
<td>provides the composition of the pathway at all levels, i.e., in terms of reactions participating and molecules and groups transformed</td>
</tr>
<tr>
<td>:DESCRIBE</td>
<td>describes the pathway</td>
</tr>
<tr>
<td>ESTIMATE-MAXIMUM-RATE</td>
<td>estimates the maximum rate of the whole pathway in its nominal (i.e., forward) direction</td>
</tr>
<tr>
<td>ESTIMATE-REVERSE-MAXIMUM-RATE</td>
<td>estimates the maximum rate of the whole pathway in its reverse (i.e., backward) direction</td>
</tr>
<tr>
<td>ESTIMATE-UNCOPLED-MAXIMUM-RATE</td>
<td>estimates the maximum rate of the pathway (in both directions) by combining independent estimations for each reaction, i.e., neglecting the coupling of reactions by intermediates</td>
</tr>
<tr>
<td>:GET</td>
<td>retrieves a property</td>
</tr>
<tr>
<td>GET-A-NAME</td>
<td>returns a name of the pathway</td>
</tr>
<tr>
<td>(depending on what names are available — at worst, the UNIQUE-NAME)</td>
<td></td>
</tr>
<tr>
<td>MAXIMUM-RATE-VALUE</td>
<td>returns the value of the maximum rate</td>
</tr>
<tr>
<td>NUMBER-OF-REACTIONS</td>
<td>returns the total number of reactions of the pathway</td>
</tr>
</tbody>
</table>

continued
Table A.18, continued

<table>
<thead>
<tr>
<th>METHOD NAME</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRESENT-MAXIMUM-RATE</td>
<td>displays detailed maximum-rate results</td>
</tr>
<tr>
<td>PRIMARY-COMPOSITION</td>
<td>returns the pathway's reaction-composition</td>
</tr>
<tr>
<td>(SETF PRIMARY-COMPOSITION)</td>
<td>sets the pathway's reaction-composition</td>
</tr>
<tr>
<td>:PUTPROP</td>
<td>sets a property of the pathway</td>
</tr>
<tr>
<td>:REMPROP</td>
<td>removes a property of the pathway</td>
</tr>
<tr>
<td>REPORT-ON-MAXIMUM-RATE</td>
<td>displays detailed maximum-rate results</td>
</tr>
<tr>
<td>SECONDARY-COMPOSITION</td>
<td>retrieves the molecular-composition</td>
</tr>
<tr>
<td>(SETF SECONDARY-COMPOSITION)</td>
<td>sets the molecular-composition</td>
</tr>
<tr>
<td>SORT-ITEM</td>
<td>given another pathway, the method decides whether the current pathway</td>
</tr>
<tr>
<td></td>
<td>should precede the other pathway in ordered-lists</td>
</tr>
<tr>
<td>SUBSTITUTE-COMPOSITE-GROUPS-IN-OBJECT</td>
<td>breaks down any composite groups that the pathway contains into their</td>
</tr>
<tr>
<td></td>
<td>constituent simpler groups</td>
</tr>
<tr>
<td>UNBIND-NAMES</td>
<td>cancel the names given to the pathway</td>
</tr>
<tr>
<td>UPDATE-ALL-MAXRAT-EQUATIONS</td>
<td>constructs the equations that can be used for the estimation of the</td>
</tr>
<tr>
<td></td>
<td>maximum rate of the whole pathway</td>
</tr>
<tr>
<td>UPDATE-DG-CONTRIBUTION</td>
<td>recalculate the Gibbs Energy from the contributions of groups and store</td>
</tr>
<tr>
<td></td>
<td>it into dg-contribution</td>
</tr>
<tr>
<td>UPDATE-MOLECULAR-INTERMEDIATES</td>
<td>reconstructs the list of the molecular intermediates of the pathway</td>
</tr>
</tbody>
</table>
Table A.19:

Example of an instantiated molecule.

Instance variable values of FRUCTOSE-DIPHOSPHATE.

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGLE-NAME</td>
<td>FRUDP</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>FRUDP</td>
</tr>
<tr>
<td>LONG-NAME</td>
<td>ALPHA-D-FRUCTOSE-1-6-DIPHOSPHATE</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>(ALPHA-D-FRUCTOSE-1-6-DIPHOSPHATE</td>
</tr>
<tr>
<td></td>
<td>A-D-FRU-1-6DIP FRUCTOSE-16DIPHOSPHATE</td>
</tr>
<tr>
<td></td>
<td>D-FRUCTOSE-16DIPHOSPHATE FRUC16DP</td>
</tr>
<tr>
<td></td>
<td>FRUDP FRU16DP FRUCTOSE-DIPHOSPHATE)</td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>&quot;CHUNKS&quot;</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>MOLECULE-574</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>&quot;MOLECULE&quot;</td>
</tr>
<tr>
<td>PARTICIPATES-IN</td>
<td>(FRUDP-ALDOLASE PFK)</td>
</tr>
</tbody>
</table>

continued
Table A.19, continued

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMMENT</td>
<td>&quot;Alternative conformations are Ignored&quot;</td>
</tr>
<tr>
<td>WEIGHT</td>
<td>336</td>
</tr>
<tr>
<td>CHARGE</td>
<td>-4</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>-911682.94 kJ/mol</td>
</tr>
<tr>
<td>BAD-DATA-POINTS</td>
<td>NIL</td>
</tr>
<tr>
<td>OTHER-CONFORMATIONS</td>
<td>NIL</td>
</tr>
<tr>
<td>BONDING-COMPOSITION</td>
<td>((SINGLE-BOND 14)</td>
</tr>
<tr>
<td></td>
<td>(RING-SINGLE-BOND 10))</td>
</tr>
<tr>
<td>ELEMENTARY-COMPOSITION</td>
<td>((PHOSPHOROUS-ATOM 2)</td>
</tr>
<tr>
<td></td>
<td>(HYDROGEN-ATOM 10)</td>
</tr>
<tr>
<td></td>
<td>(CARBON-ATOM 6)</td>
</tr>
<tr>
<td></td>
<td>(OXYGEN-ATOM 12)</td>
</tr>
<tr>
<td></td>
<td>(EXTRA-ELECTRON 4))</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>NIL</td>
</tr>
<tr>
<td>DG-VALUES</td>
<td>(-911682.94), in kJ/mol</td>
</tr>
<tr>
<td>GROUP-COMPOSITION</td>
<td>((RESIDUAL 1)</td>
</tr>
<tr>
<td></td>
<td>(-OH/SEC 2)</td>
</tr>
<tr>
<td></td>
<td>(-CH2- 2)</td>
</tr>
<tr>
<td></td>
<td>(-OPO3@-2 2)</td>
</tr>
<tr>
<td></td>
<td>(2R-O- 1)</td>
</tr>
<tr>
<td></td>
<td>(2R-CH-- 3)</td>
</tr>
<tr>
<td></td>
<td>(-OH/TERT 1)</td>
</tr>
<tr>
<td></td>
<td>(2R--C-- 1))</td>
</tr>
</tbody>
</table>
Table A.20:

Example of an instantiated reaction.

Instance variable values of FRUCTOSE-DIPHOSPHATE-ALDOLASE.

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGLE-NAME</td>
<td>FRUDP-ALDOLASE</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>ALDOLASE</td>
</tr>
<tr>
<td>LONG-NAME</td>
<td>FRUCTOSE-1-6-DIPHOSPHATE- D-GLYCERALDEHYDE-3-PHOSPHATE-LYASE</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>(FRUCTOSEDIPHOSPHATE-ALDOLASE</td>
</tr>
<tr>
<td></td>
<td>FRUCTOSE-1-6-DIPHOSPHATE- D-GLYCERALDEHYDE-3-PHOSPHATE-LYASE</td>
</tr>
<tr>
<td></td>
<td>FRUCTOSE-BIPHOSPHATE-ALDOLASE</td>
</tr>
<tr>
<td></td>
<td>FRUCTOSE-DIPHOSPHATE-ALDOLASE ALDOLASE</td>
</tr>
<tr>
<td></td>
<td>FRU-DP-ALDOLASE FRU16DP-ALDOLASE</td>
</tr>
<tr>
<td></td>
<td>FRUDP-ALDOLASE FRUC-DP-ALDOLASE)</td>
</tr>
</tbody>
</table>

continued
Table A.20, continued

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJECT-PACKAGE</td>
<td>&quot;CHUNKS&quot;</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>REACTION-764</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>&quot;REACTION&quot;</td>
</tr>
<tr>
<td>PARTICIPATES-IN</td>
<td>(GLC→PYR GLC→PEP→PYR→OXAC</td>
</tr>
<tr>
<td></td>
<td>GLC→PEP→PYR GLC→PEP→LYS GLC→LYS)</td>
</tr>
<tr>
<td>COMMENT</td>
<td>&quot;This cannot be a rate-limiting step of glycolysis&quot;</td>
</tr>
<tr>
<td>DG-CONTRIBUTION</td>
<td>20379.14</td>
</tr>
<tr>
<td>BAD-DATA-POINTS</td>
<td>NIL</td>
</tr>
<tr>
<td>OTHER-DATA</td>
<td>NIL</td>
</tr>
<tr>
<td>DELTA-G</td>
<td>((5.73 kcal/mol ±0.01, from LEHNINGER))</td>
</tr>
<tr>
<td>K-EQ</td>
<td>((8.1e-5 M, from BARMAN))</td>
</tr>
<tr>
<td>DG-VALUES</td>
<td>(20379.14 23744.74 23985.78), in J/mol</td>
</tr>
</tbody>
</table>

continued
Table A.20, continued

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP-COMPOSITION</td>
<td></td>
</tr>
<tr>
<td>((-OH/PRIM 1) (-OH/SEC -1)</td>
<td></td>
</tr>
<tr>
<td>(-OH/TERT -1) (-CHO 1) (-CH2-1)</td>
<td></td>
</tr>
<tr>
<td>(-CO- 1) (-CH-- 1) (2R-O- -1)</td>
<td></td>
</tr>
<tr>
<td>(2R-CH-- -3) (2R--C-- -1)</td>
<td></td>
</tr>
<tr>
<td>(/RESIDUAL 1))</td>
<td></td>
</tr>
<tr>
<td>MOLECULAR-COMPOSITION</td>
<td></td>
</tr>
<tr>
<td>((DHAP 1) (GAP 1) (FRUDP -1))</td>
<td></td>
</tr>
<tr>
<td>FORWARD-MAXIMUM-RATE</td>
<td></td>
</tr>
<tr>
<td>(7.996 d-5 obtained with:</td>
<td></td>
</tr>
<tr>
<td>DISSOCIATION-PARAMETER for DHAP =29120)</td>
<td></td>
</tr>
<tr>
<td>BACKWARD-MAXIMUM-RATE</td>
<td></td>
</tr>
<tr>
<td>(1.816 d-5 obtained with:</td>
<td></td>
</tr>
<tr>
<td>DISSOCIATION-PARAMETER for DHAP =16437)</td>
<td></td>
</tr>
<tr>
<td>EC-NUMBER</td>
<td></td>
</tr>
<tr>
<td>(4 1 2 13)</td>
<td></td>
</tr>
<tr>
<td>TRUE-ENZYME-P</td>
<td>T</td>
</tr>
</tbody>
</table>
Table A.21:
Example of an instantiated pathway.
Instance variable values of GLC→PYR,
which represents the glycolytic pathway.

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINGLE-NAME</td>
<td>GLC→PYR</td>
</tr>
<tr>
<td>SHORT-NAME</td>
<td>GLC→PYR</td>
</tr>
<tr>
<td>LONG-NAME</td>
<td>GLUCOSE→PYRUVATE</td>
</tr>
<tr>
<td>NAME-LIST</td>
<td>(GLC→PYR GLUCOSE→PYRUVATE GLYCOLYSIS GLYCOLYTIC-PATHWAY)</td>
</tr>
<tr>
<td>OBJECT-PACKAGE</td>
<td>&quot;CHUNKS&quot;</td>
</tr>
<tr>
<td>UNIQUE-NAME</td>
<td>PATHWAY-906</td>
</tr>
<tr>
<td>CLASS-STRING</td>
<td>&quot;PATHWAY&quot;</td>
</tr>
<tr>
<td>COMMENT</td>
<td>&quot;Main pathway for the catabolism of glucose&quot;</td>
</tr>
</tbody>
</table>

_continued_
<table>
<thead>
<tr>
<th>Instance Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DG-CONTRIBUTION</strong></td>
<td>-112302.82 kJ/mol</td>
</tr>
<tr>
<td><strong>DG-VALUES</strong></td>
<td>{-112302.82}</td>
</tr>
<tr>
<td><strong>FORWARD-MAXIMUM-RATE</strong></td>
<td>0.0022353, uncoupled</td>
</tr>
<tr>
<td><strong>BACKWARD-MAXIMUM-RATE</strong></td>
<td>:INFEASIBLE</td>
</tr>
<tr>
<td><strong>GROUP-COMPOSITION</strong></td>
<td>((/RESIDUAL 3) (NAD-OX-GROUP -2)</td>
</tr>
<tr>
<td></td>
<td>(NAD-RED-GROUP 2) (H@+1-GROUP 2)</td>
</tr>
<tr>
<td></td>
<td>(H2O-GROUP 2) (PI@-2-GROUP -2)</td>
</tr>
<tr>
<td></td>
<td>(-OH/PRIM -1) (-OH/SEC -4)</td>
</tr>
<tr>
<td></td>
<td>(-COO@-1 2) (-CH2- -1) (-CO- 2)</td>
</tr>
<tr>
<td></td>
<td>(-OPO2@-1- 2) (2R-CH2- 2)</td>
</tr>
<tr>
<td></td>
<td>(2R-CH-- -5) (2R-N-- 2)</td>
</tr>
<tr>
<td></td>
<td>(2R-CH= -2) (2R=N-@+1 -2)</td>
</tr>
<tr>
<td></td>
<td>(2R-O- -1) (-CH3 2))</td>
</tr>
</tbody>
</table>

continued
Table A.21, continued

<table>
<thead>
<tr>
<th>INSTANCE VARIABLE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOLECULAR-COMPOSITION</td>
<td>((H@+1 \ 2) \ (H_2O \ 2) \ (Pi@-2 \ -2) )</td>
</tr>
<tr>
<td></td>
<td>((\text{PYR} \ 2) \ (\text{GLC} \ -1) \ (\text{ADP} \ -2) )</td>
</tr>
<tr>
<td></td>
<td>((\text{ATP} \ 2) \ (\text{NAD}+ \ -2) \ (\text{NADH} \ 2) )</td>
</tr>
<tr>
<td>REACTION-COMPOSITION</td>
<td>((\text{GAP-DE-H-ASE} \ 2) \ (\text{HEXOKINASE} \ 1) )</td>
</tr>
<tr>
<td></td>
<td>((\text{PYR-KINASE} \ 2) \ (\text{PG-MUTASE} \ -2) )</td>
</tr>
<tr>
<td></td>
<td>((\text{FRUDP-ALDOLASE} \ 1) \ (\text{ENOLASE} \ 2) )</td>
</tr>
<tr>
<td></td>
<td>((\text{PG-KINASE} \ 2) \ (\text{GLC6P-ISOMERASE} \ 1) )</td>
</tr>
<tr>
<td></td>
<td>((\text{TRIOSE-P-ISOMERASE} \ 1) \ (\text{PFK} \ 1) )</td>
</tr>
<tr>
<td>MOLECULAR-INTERMEDIATES</td>
<td>((\text{PYR} \ \text{NADH} \ \text{NAD}+ \ Pi@-2 \ \text{GLC} )</td>
</tr>
<tr>
<td></td>
<td>((13\text{DPG} \ 3\text{PG} \ 2\text{PG} \ \text{PEP} \ \text{H}_2\text{O} \ \text{GAP} )</td>
</tr>
<tr>
<td></td>
<td>((\text{DHAP} \ \text{GLC6P} \ \text{ATP} \ \text{ADP} \ \text{FRUDP} )</td>
</tr>
<tr>
<td></td>
<td>((\text{FRU6P} \ H@+1) )</td>
</tr>
</tbody>
</table>
Figure A.5:
Interrelationships among Biochemical Objects
A.5.2. ORDER-OF-MAGNITUDE OBJECTS

The O[M] system contains objects that are much simpler than the biochemical database objects, because the tasks of O[M] objects are much more confined and well-defined. We will describe O[M] objects and their attributes and methods in in-text tables in the following paragraphs. The interconnections among Order-of-Magnitude objects are shown in Figure A.6, at the end of this section.

A.5.2.1. Relators

Relators are the operators of the O[M] relations. Thus, there are 7 primitive relators (<<, <=, ~<, ==, ~>, >, >>) and 21 compound relators.

The attributes of relators are:

- Expanded-Range (from the heuristic interpretation)
- Shrank-Range (from the heuristic interpretation)
- Epsilon, which is the accuracy-parameter of the interpretation
- Primitives

The relators are predefined and cannot be modified. The only operation that can be performed on them is a change in the accuracy parameter, accomplished through the method new-epsilon.

A.5.2.2. Landmarks

The attributes of landmarks are:
• Physical-dimension

• Links, in which the landmark participates

• Value (numerical)

• Units (for the value)

Landmarks are constructed by calling the function `get-landmark`, with the value and the units as arguments. Landmarks can be retracted through the method `retract`.

A.5.2.3. Variables

The attributes of variables are:

• Physical-Dimension

• Links, in which the variable participates

• Name, which is a symbol

Variables are constructed by calling the function `make-variable`, with the dimension as argument.

A.5.2.4. Links

Some of the instance variables of links contain objects in which relations from the link can be used. Other instance variables contain objects from which relations for the link can be inferred. The instance variables are:

• Operands, describing the quantities that the link interrelates. The two quantities are also listed as quantity-1 and quantity-2.
• Relations, referring to the link

• Constraints, involving the link

• From-Assignments, i.e., assignments that contain the link as their destination

• To-Assignments, i.e., assignments that contain the link in their right hand side

• From-Rules, i.e., rules that contain the link in their consequents

• To-Rules, i.e., rules that contain the link in their antecedents

Links can be constructed by calling the function `get-link`, with the operands as arguments.

A.5.2.5. Relations

Relations, such as "A << B," have the following instance variables:

• Relator

• Link

• Assumptions, i.e., the assumption set on which the validity of the relation depends

• Justification, which is a construct describing how the relation was derived

• Nogoodp, a boolean which has the value t if the relation has been retracted due to inconsistencies
Relations are constructed by calling the function `construct-relation`, which takes as keyword arguments the relator, the link, the assumptions, and the justification of the relation. Relations can be retracted by the method `retract`. The method `employ` derives all possible inferences from a newly defined relation.

### A.5.2.6. Rules

The instance variables of rules are:

- Antecedents
- Consequents
- Antecedent-links, i.e., links of the antecedent relations
- Consequent-links, i.e., links of the consequent relations

Rules are constructed by the function `make-rule`, with the antecedents and consequents as keywords arguments. The method `employ`, derives all possible inferences from a newly defined rule.

### A.5.2.7. Assignments

The instance variables of assignments are:

- Landmarks, involved in the right hand side of the assignment
- Links, involved in the right hand side of the assignment
- Destination
- Equation, describing the right hand side of the assignment
The function `construct-assignment`, with the destination and equation as keyword arguments, makes an assignment. The method `employ`, derives all possible inferences from a newly defined assignment.

### A.5.2.8. Constraints

The instance variables of constraints are:

- Landmarks, involved in the constraint
- Links, involved in the constraint
- Equation, describing the constraint

The function `make-constraint`, with the equation as an argument, makes a constraint. The method `employ`, derives all possible inferences from a newly defined constraint.

### A.5.2.9. Use of the System

To use the O[M] system (after the program module has been loaded, as described in Section A.4.1), we follow these steps:

- We construct all the variables, landmarks, and links that we will use, using the functions prescribed for each type of object in the previous paragraphs.
- We construct relations, constraints, assignments, rules, using the functions prescribed in the previous paragraphs. After each object is constructed, the method `employ` is called on the object, to
instruct O[M] to infer all possible results from the new piece of knowledge.

• We check for any inconsistencies that may have been uncovered. Inconsistencies appear as nogood relations and assumptions.

• We inspect the variables and links of interest, to find all the relations that were derived for them.

• If the reasoning has not yielded sufficient results, we use the function target, with a link as its argument, to state goals for the reasoning, and repeat the above procedure beginning from the second step.
Figure A.6:

Interconnections among Order-of-Magnitude objects
A.5.3. IMPLEMENTATION EFFICIENCY ISSUES

LISP offers significant advantages in applications whose nature entails symbolic manipulation rather than numerical operations. In our case, numerical operations were an integral part of a largely symbolic computation because of the numerical interpretation of the Order-of-Magnitude relations. Contrary to common belief, the LISP language and "specialized" LISP computers are perfectly capable of performing number manipulation quite efficiently, and the programmer still gains from the flexible development environment.

Objects are constructed and demolished rapidly during Order-of-Magnitude reasoning, as new relations are derived and redundant or false relations are discarded. Handling all these objects is the most expensive part of the computation, and, in order to keep it manageable, special attention was given to the construction and demolition of objects:

- Before actually creating a new object, we perform as many checks as possible to make sure it is not redundant. Thus, if we know that $A \ll B$, and some inference yields $A \ll..=\ll B$, the latter relation will be discarded, without ever creating an object to represent it.

- We get rid of any objects that become unnecessary, to free up space. If the relation $A \ll..=\ll B$ is shown redundant because of a new relation $A \prec B$, we remove from other data structures all references (pointers) to the object $A \ll..=\ll B$ and redirect the references to the object $A \prec B$, so that the old object becomes garbage and is swept by the garbage collector. This has the added benefit of reducing the volume of the results, so that they are easier for the user to scan.
In O[M], the use of LISP and OOP in a procedural fashion, is much more suitable than a rule-based approach (like FOG) to Order-of-Magnitude Reasoning. With our procedural method we can not only use a variety of knowledge-representation means (such as relations, constraints, assumptions, rules, and numerical values), but also operate on these items with a powerful set of inference methods, and take advantage of available numerical knowledge. Since we have explicit control over every kind of search the procedures perform, we can achieve the computational efficiency crucial for large problems.

In the synthesis of biochemical pathways we have again a phenomenon of rapid construction and demolition of objects during the reasoning, as new pathways are constructed and pathways violating the specifications are discarded. Hence, careful construction and demolition is required in the implementation of the synthesis algorithm as well.