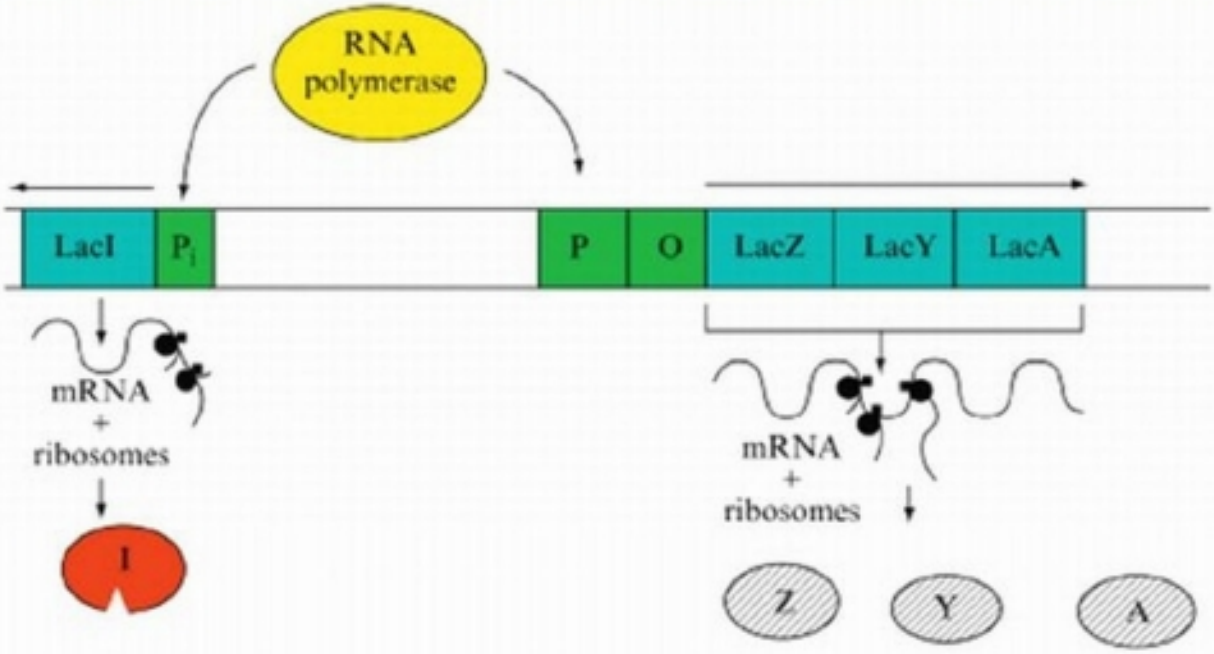
Reliable Lab Solutions

Easy-to-use

Dependable

Respected

ESSENTIAL NUMERICAL COMPUTER METHODS



Edited by
Michael Johnson

AP

Essential Numerical Computer Methods

Reliable Lab Solutions

This page intentionally left blank

Essential Numerical Computer Methods

Reliable Lab Solutions

Edited by

Michael L. Johnson

Departments of Pharmacology and Medicine
University of Virginia Health System
Charlottesville, VA, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
32 Jamestown Road, London NW1 7BY, UK
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

Material in the work originally appeared in Volumes 210, 240, 321, 383, 384, 454, and 467 of *Methods in Enzymology* (1992, 1994, 2000, 2004, 2009, Elsevier Inc.)

ISBN: 978-0-12-384997-7

For information on all Academic Press publications
visit our website at elsevierdirect.com

Printed and bound in USA

10 11 12 10 9 8 7 6 5 4 3 2 1

CONTENTS

Contributors xiii

Preface xv

1. Use of Least-Squares Techniques in Biochemistry 1

Michael L. Johnson

I. Update 1
II. Introduction 2
III. Nonlinear Least-Squares 4
IV. Why Use NLLS Analysis Procedures? 6
V. When to Use NLLS Analysis Procedures 7
VI. What Confidence Can Be Assigned to Results of NLLS Analysis? 14
VII. Conclusions 20
References 22

2. Parameter Estimates from Nonlinear Models 23

Donald G. Watts

I. Introduction 24
II. Discussion 35
References 36

3. Analysis of Residuals: Criteria for Determining Goodness-of-Fit 37

Martin Straume and Michael L. Johnson

I. Update 37
II. Introduction 38
III. Scatter Diagram Residual Plots 40
IV. Cumulative Probability Distributions of Residuals 41
V. χ^2 Statistic: Quantifying Observed Versus Expected Frequencies of Residual Values 43
VI. Kolmogorov-Smirnov Test: An Alternative to the χ^2 Statistic 44
VII. Runs Test: Quantifying Trends in Residuals 45
VIII. Serial Lag_n Plots: Identifying Serial Correlation 47
IX. Durbin-Watson Test: Quantitative Testing for Serial Correlation 47

X. Autocorrelation: Detecting Serial Correlation in Time Series Experiments 49
XI. χ^2 Test: Quantitation of Goodness-of-Fit 50
XII. Outliers: Identifying Bad Points 51
XIII. Identifying Influential Observations 52
XIV. Conclusions 53
References 54

4. Monte Carlo Method for Determining Complete Confidence Probability Distributions of Estimated Model Parameters 55
Martin Straume and Michael L. Johnson

I. Update 55
II. Introduction 56
III. Monte Carlo Method 58
IV. Generating Confidence Probability Distributions for Estimated Parameters 58
V. Implementation and Interpretation 60
VI. Conclusion 65
References 66

5. Effects of Heteroscedasticity and Skewness on Prediction in Regression: Modeling Growth of the Human Heart 67
Robert D. Abbott and Howard P. Gutgesell

I. Introduction 67
II. Example from Modeling Growth of the Human Heart 68
III. Methods of Estimation 70
IV. Discussion 78
References 79

6. Singular Value Decomposition: Application to Analysis of Experimental Data 81
E. R. Henry and J. Hofrichter

I. Update 82
II. Introduction 83
III. Definition and Properties 91
IV. Singular Value Decomposition of Matrices Which Contain Noise 93
V. Application of Singular Value Decomposition to Analysis of Experimental Data 107
VI. Simulations for a Simple Example: The Reaction A → B→ C 120
VII. Summary 133
References 138

7. Irregularity and Asynchrony in Biologic Network Signals 141
Steven M. Pincus

I. Update 142
II. Introduction 142
III. Quantification of Regularity 146
IV. Implementation and Interpretation 148
V. Representative Biological Applications 152
VI. Relationship to Other Approaches 154
VII. Mechanistic Hypothesis for Altered Regularity 158
VIII. Cross-ApEn 158
IX. Toward More Faithful Network Modeling 166
X. Spatial (Vector) ApEn 167
XI. Summary and Conclusion 169
References 170

8. Distinguishing Models of Growth with Approximate Entropy 173
Michael L. Johnson, Michelle Lampl, and Martin Straume

I. Update 173
II. Introduction 174
III. Definition and Calculation of ApEn 175
IV. Modifications of ApEn Calculation for this Application 176
V. Growth Models 177
VI. Expected Model-Dependent Distribution of ApEn 178
VII. Example of this Use of ApEn 179
VIII. Conclusion 182
References 184

9. Application of the Kalman Filter to Computational Problems in Statistics 187
Emery N. Brown and Christopher H. Schmid

I. Introduction 187
II. Evaluating Gaussian Likelihood Using the Kalman Filter 189
III. Computing Posterior Densities for Bayesian Inference Using the Kalman Filter 192
IV. Missing Data Problems and the Kalman Filter 193
V. Extensions of the Kalman Filter Algorithm 196
References 196

10. Bayesian Hierarchical Models 199
Christopher H. Schmid and Emery N. Brown

I. Introduction 199
II. The Gaussian Model 203
III. Computation 208

IV. Example: Meta-Regression 211
V. Example: Saltatory Model of Infant Growth 216
VI. Incorporation of Variance Components 219
VII. Model Checking 220
VIII. Conclusion 221
References 222

11. Mixed-Model Regression Analysis and Dealing with Interindividual Differences 225
Hans P. A. Van Dongen, Erik Olofsen, David F. Dinges, and Greg Maislin

I. Introduction 226
II. Experiment and Data 227
III. Repeated-Measures ANOVA 228
IV. Mixed-Model Regression Analysis 235
V. An Alternative Linear Mixed-Effects Model 241
VI. Nonlinear Mixed-Model Regression Analysis 245
VII. Correlation Structures in Mixed-Model Regression Analysis 253
VIII. Conclusion 254
References 255

12. Distribution Functions from Moments and the Maximum-Entropy Method 257
Douglas Poland

I. Introduction 257
II. Ligand Binding: Moments 259
III. Maximum-Entropy Distributions 265
IV. Ligand Binding: Distribution Functions 273
V. Enthalpy Distributions 279
VI. Self-Association Distributions 287
References 291

13. The Mathematics of Biological Oscillators 293
G. Bard Ermentrout

I. Introduction 293
II. Oscillators and Excitability 294
III. Perturbations of Oscillators 299
IV. Coupled Oscillators 304
References 308

14. Modeling of Oscillations in Endocrine Networks with Feedback 309
Leon S. Farhy

I. Introduction 309
II. General Principles in Endocrine Network Modeling 310

III. Simulating the Concentration Dynamics of a Single Hormone 312
IV. Oscillations Driven by a Single System Feedback Loop 315
V. Networks with Multiple Feedback Loops 330
VI. Summary and Discussion 333
References 336

15. Boundary Analysis in Sedimentation Velocity Experiments 337
Walter F. Stafford

I. Update 338
II. Introduction 338
III. Methods of Data Acquisition 340
IV. Measurement of Transport in Analytical Ultracentrifuge 340
V. Traditional Methods of Analysis 341
VI. Transport Method 342
VII. Smoothing and Differentiating 344
VIII. Computation of Apparent Sedimentation Coefficient Distribution Functions 349
IX. Weight Average Sedimentation Coefficient from $g(s^\star)$ 354
X. Methods of Correcting Distribution Functions for Effects of Diffusion 357
XI. Discussion 358
References 358

16. Statistical Error in Isothermal Titration Calorimetry 361
Joel Tellinghuisen

I. Update 361
II. Introduction 364
III. Variance-Covariance Matrix in Least Squares 366
IV. Monte Carlo Computational Methods 372
V. Van't Hoff Analysis of $K^\circ(T)$: Least-Squares Demonstration 373
VI. Isothermal Titration Calorimetry 377
VII. Calorimetric Versus Van't Hoff ΔH° from ITC 390
VIII. Conclusion 394
References 397

17. Physiological Modeling with Virtual Cell Framework 399
James C. Schaff, Boris M. Slepchenko, and Leslie M. Loew

I. Update 400
II. Introduction 401
III. Modeling Abstractions for Cellular Physiology 402
IV. Application to Existing Model 412
V. Conclusions 418
References 424

18. Fractal Applications in Biology: Scaling Time in Biochemical Networks 425
F. Eugene Yates

I. Update 426
II. Introduction 427
III. Fractal Morphology in Mammals: Some Branchings 430
IV. Chaos in Enzyme Reactions 437
V. Practical Guide to Identification of Chaos and Fractals in Biochemical Reactions 438
VI. Summary, or What Does It All Mean? 440
VII. Glossary 442
References 458

19. Analytical Methods for the Retrieval and Interpretation of Continuous Glucose Monitoring Data in Diabetes 461
Boris Kovatchev, Marc Breton, and William Clarke

I. 2010 Update of Developments in the Field 462
II. Introduction 463
III. Decomposition of Sensor Errors 466
IV. Measures of Average Glycemia and Deviation from Target 467
V. Risk and Variability Assessment 469
VI. Measures and Plots of System Stability 472
VII. Time-Series-Based Prediction of Future BG Values 473
VIII. Conclusions 476
References 476

20. Analyses for Physiological and Behavioral Rhythmicity 479
Harold B. Dowse

I. Introduction 479
II. Types of Biological Data and Their Acquisition 480
III. Analysis in the Time Domain 482
IV. Analysis in the Frequency Domain 488
V. Time/Frequency Analysis and the Wavelet Transform 496
VI. Signal Conditioning 502
VII. Strength and Regularity of a Signal 505
VIII. Some Practical Considerations on Statistical Comparisons of Analytical Results 506
IX. Conclusions 507
References 508

21. Evaluation and Comparison of Computational Models 511
Jay I. Myung, Yun Tang, and Mark A. Pitt

I. Update 512
II. Introduction 513
III. Conceptual Overview of Model Evaluation and Comparison 513
IV. Model Comparison Methods 516
V. Model Comparison at Work: Choosing Between Protein Folding Models 521
VI. Conclusions 525
References 526

22. Algebraic Models of Biochemical Networks 529
Reinhard Laubenbacher and Abdul Salam Jarrah

I. Introduction 530
II. Computational Systems Biology 531
III. Network Inference 541
IV. Reverse-Engineering of Discrete Models: An Example 545
V. Discussion 553
References 556

23. Monte Carlo Simulation in Establishing Analytical Quality Requirements for Clinical Laboratory Tests: Meeting Clinical Needs 561
James C. Boyd and David E. Bruns

I. Update 562
II. Introduction 563
III. Modeling Approach 566
IV. Methods for Simulation Study 567
V. Results 568
VI. Discussion 580
References 582

24. Pancreatic Network Control of Glucagon Secretion and Counterregulation 585
Leon S. Farhy and Anthony L. McCall

I. Update 586
II. Introduction 588
III. Mechanisms of Glucagon Counterregulation (GCR) Dysregulation in Diabetes 589

IV. Interdisciplinary Approach to Investigating the Defects in the GCR 591
V. Initial Qualitative Analysis of the GCR Control Axis 592
VI. Mathematical Models of the GCR Control Mechanisms in STZ-Treated Rats 595
VII. Approximation of the Normal Endocrine Pancreas by a Minimal Control Network (MCN) and Analysis of the GCR Abnormalities in the Insulin-Deficient Pancreas 598
VIII. Advantages and Limitations of the Interdisciplinary Approach 609
IX. Conclusions 612
References 612

Index 619

CONTRIBUTORS

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Robert D. Abbott (67), Division of Biostatistics, Department of Internal Medicine, University of Virginia School of Medicine, Charlottesville, Virginia, USA

G. Bard Ermentrout (293), Department of Mathematics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

James C. Boyd (561), Department of Pathology, University of Virginia Health System, Charlottesville, Virginia, USA

Marc Breton (461), University of Virginia Health System, Charlottesville, Virginia, USA

Emery N. Brown (187, 199), Statistics Research Laboratory, Department of Anesthesia and Critical Care, Massachusetts General Hospital, Harvard Medical School/MIT Division of Health Sciences and Technology, Boston, Massachusetts, USA

David E. Bruns (561), Department of Pathology, University of Virginia Health System, Charlottesville, Virginia, USA

William Clarke (461), University of Virginia Health System, Charlottesville, Virginia, USA

David F. Dinges (225), Unit for Experimental Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA

Harold B. Dowse (479), School of Biology and Ecology, University of Maine, Orono, Maine, USA; Department of Mathematics and Statistics, University of Maine, Orono, Maine, USA

Leon S. Farhy (309, 585), Department of Medicine, Center for Biomathematical Technology, University of Virginia, Charlottesville, Virginia, USA

Howard P. Gutgesell (67), Division of Pediatric Cardiology, Department of Pediatrics, University of Virginia School of Medicine, Charlottesville, Virginia, USA

E. R. Henry (81), Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda MD, USA

J. Hofrichter (81), Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda MD, USA

Abdul Salam Jarrah (529), Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, Virginia, USA; Department of Mathematics and Statistics, American University of Sharjah, Sharjah, United Arab Emirates

Michael L. Johnson (1, 37, 55, 173), University of Virginia Health System, Charlottesville, VA, USA

Boris Kovatchev (461), University of Virginia Health System, Charlottesville, Virginia, USA

Michelle Lampl (173), Emory University, Atlanta, GA, USA
Reinhard Laubenbacher (529), Virginia Bioinformatics Institute at Virginia Tech, Blacksburg, Virginia, USA
Leslie M. Loew (399), R. D. Berlin Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA
Greg Maislin (225), Biomedical Statistical Consulting, Wynnewood, Pennsylvania, USA
Anthony L. McCall (585), Department of Medicine, Center for Biomathematical Technology, University of Virginia, Charlottesville, Virginia, USA
Jay I. Myung (511), Department of Psychology, Ohio State University, Columbus, Ohio, USA
Erik Olofsen (225), Department of Anesthesiology, P5Q, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands
Steven M. Pincus (141), Independent Mathematician, Guilford, CT, USA
Mark A. Pitt (511), Department of Psychology, Ohio State University, Columbus, Ohio, USA
Douglas Poland (257), Department of Chemistry, The Johns Hopkins University, Baltimore, Maryland, USA
James C. Schaff (399), R. D. Berlin Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA
Christopher H. Schmid (187, 199), Biostatistics Research Center, Division of Clinical Care Research, New England Medical Center, Tufts University, Boston, Massachusetts, USA
Boris M. Slepchenko (399), R. D. Berlin Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT, USA
Walter F. Stafford III (337), Analytical Centrifugation Research Laboratory, Boston Biomedical Research Institute, Boston, MA, USA; Department of Neurology, Harvard Medical School, Boston, MA, USA
Martin Straume (37, 55, 173), COBRA, Inc.
Yun Tang (511), Department of Psychology, Ohio State University, Columbus, Ohio, USA
Joel Tellinghuisen (361), Department of Chemistry, Vanderbilt University, Nashville, Tennessee, USA
Hans P. A. Van Dongen (225), Unit for Experimental Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA
Donald G. Watts (23), Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6
F. Eugene Yates (425), The John Douglas French Alzheimer's Foundation, 11620 Wilshire Blvd, Suite 270, Los Angeles, CA, USA

PREFACE

Many of the chapters within this volume were first published almost two decades ago. Since then, these basic algorithms have not changed. However, what has changed is the huge increase in computer speed and ease of use along with the corresponding decrease in the costs. The increase in computer speed has made the use of some algorithms common that were almost never used in biochemistry laboratories two decades ago. During the past two decades, the training of the majority of senior M.D.s and Ph.D.s in clinical or basic disciplines at academic research and medical centers has not kept pace with advanced coursework in mathematics, numerical analysis, statistics, or computer science.

Nevertheless, the use of computers and computational methods has become ubiquitous in biological and biomedical research. One primary reason is the emphasis being placed on computers and computational methods within the National Institutes of Health (NIH) Roadmap. Another factor is the increased level of mathematical and computational sophistication among researchers, particularly among junior scientists, students, journal reviewers and NIH Study Section members. Yet another is the rapid advances in computer hardware and software that make these methods far more accessible to the rank-and-file members of the research community.

There exists a general perception that the applications of computers and computer methods in biological and biomedical research are either basic statistical analysis or the searching of DNA sequence data bases. While these are important applications, they only scratch the surface of the current and potential applications of computers and computer methods in biomedical research. The various chapters within this volume include a wide variety of applications that extend far beyond this limited perception. The chapters within this volume are basically in chronological order of original publication in Methods in Enzymology volumes 210, 240, 321, 383, 384, 454, and 467. This chronological order also provides a general progression from basic numerical methods to more specific biochemical and biomedical applications.

This page intentionally left blank

CHAPTER 1

Use of Least-Squares Techniques in Biochemistry[1]

Michael L. Johnson

University of Virginia Health System
Charlottesville, VA, USA

I. Update
II. Introduction
III. Nonlinear Least-Squares
IV. Why Use NLLS Analysis Procedures?
V. When to Use NLLS Analysis Procedures
  A. Assumption 1: No Experimental Uncertainty
  B. Assumption 2: Gaussian Uncertainties
  C. Assumption 3: Independent Observations
  D. Assumption 4: Large Number of Data Points
VI. What Confidence Can Be Assigned to Results of NLLS Analysis?
VII. Conclusions
References

I. Update

This chapter was originally published (Johnson, 1992) under the title "Why, When, and How Biochemists Should Use Least-Squares" and this descriptive title clearly states the purpose of this chapter. This chapter emphasizes the underlying assumptions and philosophy of least-squares fitting of nonlinear equations to experimental data.

In the last two decades, the basic algorithms have not changed. Most are based on the Gauss-Newton algorithm. However, what has changed is the huge increase

[1] This article was originally published as "Why, When, and How Biochemists Should Use Least Squares" in *Analytical Biochemistry*, Volume 206 (1992). Reprinted with permission from Academic Press.

DOI: 10.1016/B978-0-12-384997-7.00001-7

in computer speed and ease of use along with the corresponding orders of magnitude decrease in cost. These factors have combined to make the least-squares fitting of nonlinear equations to experimental data a common task in all modern research laboratories. Many of the available software packages will perform the required least-squares fits, but fail to address subjects such as goodness-of-fit, joint confidence intervals, and correlation within the estimated parameters. This chapter provides an introduction to these subjects.

The increase in computer speed has made some algorithms possible that were almost never used in biochemistry laboratories two decades ago. One such algorithm is the use of bootstraps (Chernick, 1999) for the determination of parameter confidence intervals. If this chapter were being written today (Johnson, 2008), I would have stressed their use.

II. Introduction

There are relatively few methods available for the analysis of experimental data in the biochemical laboratory. Graphical methods and least-squares (regression) methods are by far the most common. Unfortunately, both classes of analysis methods are commonly misused. The purpose of this chapter is to explain why, when, and how a biochemist should use least-squares techniques and what confidence can be assigned to the resulting estimated parameters.

One classic group of biochemical experiments involves measuring the response of a system to an external perturbation. Temperature-jump experiments perturb the chemical equilibrium of a solution by rapidly increasing the temperature of the solution and subsequently monitoring an observable, like absorbance, as a function of time. Here, the absorbance is the observable (i.e., the variable) that is dependent on the experiment, and time is the variable that can be independently controlled by the experimental protocol.

Another example of this general class is the ligand-binding titration experiment. The investigator measures the amount of a ligand bound (the dependent variable) by fluorescence, absorbance, or radioactive counting. To do so, the investigator titrates the ligand concentration (the independent variable). Note that the ligand concentrations might be either the total or the free ligand concentration, depending on the experimental protocol.

In these examples, and all others of this class, the investigator has measured a response caused by a perturbation of the system. The next step is to obtain the parameters of the system that characterize the chemical processes by ''analyzing'' the data. In the above examples, these parameters, the desired answers, might be the relaxation half-lives or macroscopic binding constants. Alternatively, the desired parameters might be the microscopic forward and reverse reaction rates of the biochemical system.

Analysis of these data requires that the biochemist assume a mathematical relationship between the observed quantities, the dependent variables, and the

independent variables. This relationship is the fitting function. In the past, analysis of relaxation experiments, such as temperature jump, assumed that the mathematical relationship was a single exponential decay. Based on this assumption, the investigator would commonly perform a logarithmic transformation of the dependent variable and create a graph of, for example, the logarithm of absorbance as a function of time. If the original assumption of a single exponential process is correct, then the graph will be a straight line with a slope related to the relaxation rate of the chemical process. A single class of binding sites is a common assumption for ligand binding experiments. This, in turn, implied that the mathematical relationship for the amount bound as a function of free, or unbound, ligand was a rectangular hyperbola. A consequence of this mathematical relationship is that various transformations of the data, such as a Scatchard plot, will yield a straight line with a slope related to the binding affinity. It was quickly realized that the assumption of a single biochemical process was generally not valid. Generalizations of these graphical procedures for consideration of multiple processes were attempted but with generally poor results.

The desired result of the analysis of any experimental data is to obtain the set of parameters of the biochemical reaction with the maximum likelihood (ML), highest probability, of being correct. This is the most critical lesson of this review. We do not care what the slope of a log plot is; we want the relaxation rate constants with the ML of being correct. We do not care what the slope of a Scatchard plot is; we want the ligand binding constants with the highest probability of being correct.

Does a Scatchard plot, or a logarithmic plot, yield parameter values with the ML of being correct? Generally, they do not (Johnson and Frasier, 2010). These methods are mathematically correct if the experimental data contain no experimental uncertainties. They fail because they do not correctly consider the experimental uncertainties present in all experimental data. Why then were these graphical methods developed and commonly reported? The evaluation of the parameters with the ML of being correct requires a high-speed digital computer to perform the calculations, but the development of the graphical methods occurred before high-speed digital computers were commonly available to the biochemical researcher. At that stage graphical methods were the only practical ones for the analysis of the experimental data. Should these methods still be used? They may aid the investigator in visualizing the data, but the methods should not be used for determining parameter values.

The most common alternative to graphical analysis in use in the biochemical laboratory today is nonlinear least-squares (NLLS). To use a NLLS method, an investigator must assume a functional form for the mathematical relationship between the dependent and independent variables of the experiments in terms of a series of desired parameters. This functional form is not restricted to a form that can be transformed into a straight line, as with the graphical procedures. NLLS is a process of ''fitting'' the experimental data to almost any functional form by evaluating an optimal set of parameters for the fitting function.

Does a NLLS method yield parameter values with the highest probability of being correct? It may, if the NLLS analysis procedure is correctly formulated and