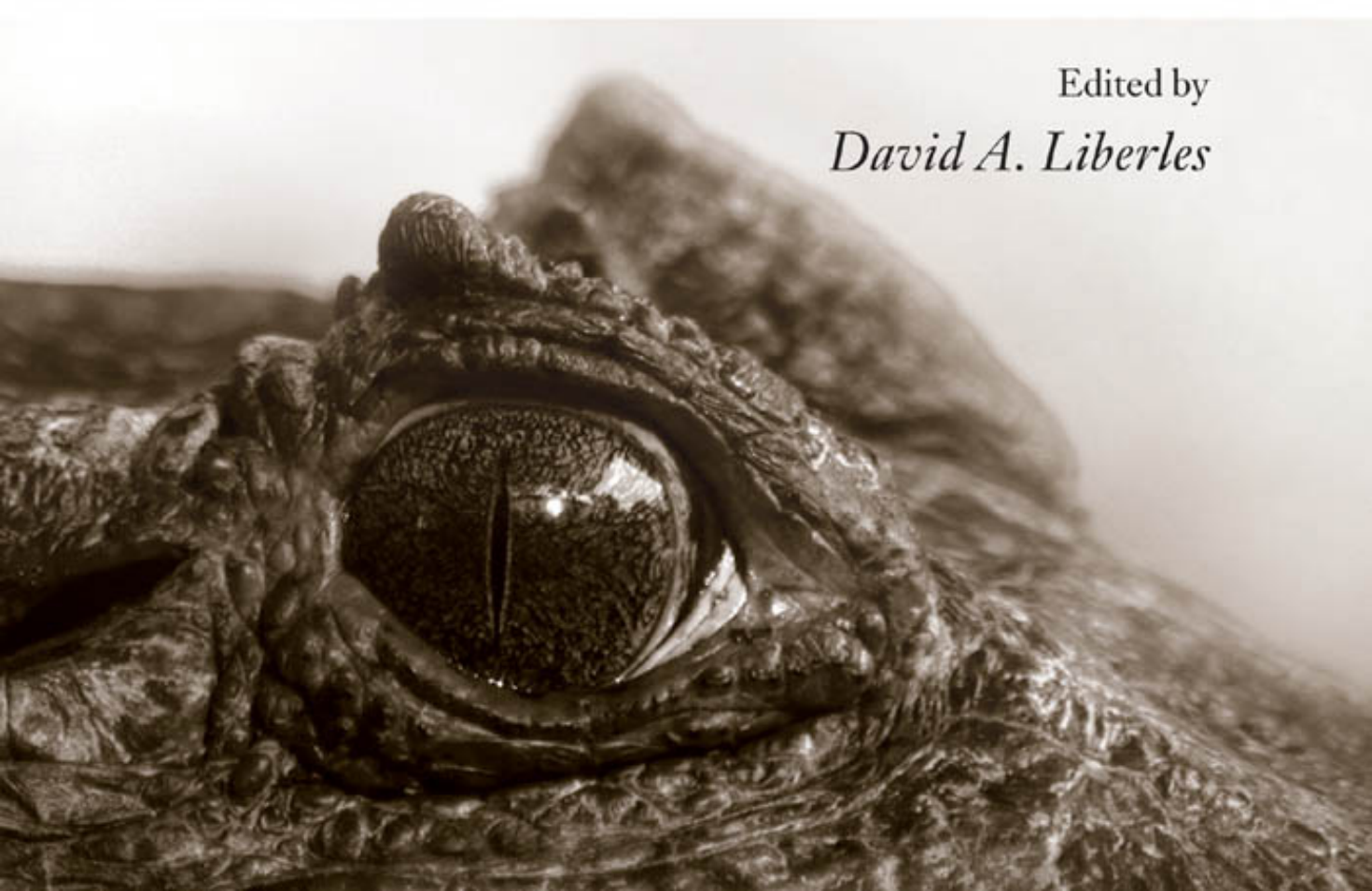
ANCESTRAL SEQUENCE RECONSTRUCTION

Edited by

David A. Liberles



Ancestral Sequence Reconstruction

This page intentionally left blank

Ancestral Sequence Reconstruction

EDITED BY

David A. Liberles

University of Wyoming, Laramie, WY, USA

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2007

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Newgen Imaging Systems (P) Ltd.,
Printed in Great Britain
on acid-free paper by
Antony Rowe, Chippenham, Wiltshire

ISBN 978 0 19 929918 8

10 9 8 7 6 5 4 3 2 1

Contents

Foreword and introduction

With the realization that the combination of computational reconstruction of ancestral protein sequences and the experimental synthesis of these proteins could be used to test specific molecular, biomedical, ecological, and evolutionary hypotheses, this methodological combination has been used with increasing popularity. Because a number of scientific issues surround the use of ancestral sequence reconstruction that need to be fleshed out, a scientific meeting was organized to discuss the use of ancestral sequence reconstruction. Beyond procedures and pitfalls, a number of new applications of ancestral sequence reconstruction have begun to emerge and a presentation of several of these was deemed valuable.

With funding from the European Science Foundation (ESF), Vetenskapsrådet (the Swedish Research Council), and the Linnaeus Centre for Bioinformatics (Uppsala University, Sweden), David Ardell (Uppsala University, Sweden), Giorgio Matassi (University of Paris VI, France), and I organized a meeting entitled, "Using Ancestral Sequence Reconstruction to Understand Protein Function" in Kristineberg, Sweden, on 30–31 March 2005. The meeting consisted of 38 participants from 12 different countries attending 18 scientific presentations. Following the meeting and the vibrant discussion, it was decided that a book involving chapters by those attending the meeting and others in the field would be worthwhile, which was the origin of this project.

One philosophical discussion that emerged was on the true meaning of homology and what a homologous site is when a sequence slides through a structure generating diverging alignments from sequence and structure-based methods. David Ardell (in his lecture in Sweden) presented examples of cases where the two diverged and recommended sequence analysis using a DNA-based view of homology, where a single position within a gene represented the homologous site, where substitution models should then be applied to characterize its evolution. This is the traditional view of homology as embodied in the vast literatures of molecular evolution and population genetics. However, Richard Goldstein (who has in the past generated substitution matrices that characterize substitution differentially between different structural elements) and David Pollock took a structural perspective on homology, arguing that a homologous position might sometimes be better defined by the structural attributes constraining it in a three-dimensional structure rather than the position within a gene sequence. For example, position 3 in an α-helix could be aligned with position 3 in the homologous α-helix of another protein, even if they represent positions 65 and 68 in the gene sequence and no insertion or deletion events have occurred. This latter view requires the use of different types of substitution models than the former view, so the divergence of opinion has practical as well as philosophical concerns.

Another active area of discussion involved sources of bias and an ongoing discussion of the validity of using maximum-likelihood or -parsimony ancestral sequence reconstructions compared with a sampling from the posterior distribution of a Bayesian ancestral sequence reconstruction. The discussion at the meeting (in addition to Chapter 8 in this volume) has spawned an active discussion in the peer-reviewed scientific literature. The argument is that the maximum-likelihood or maximum-parsimony ancestral sequence is under-represented by rare variants, such as hydrophobic residues on the surface, that ultimately attribute overly stable or overly active properties to the reconstructed ancestor. It is argued that this is avoided by sampling from the posterior distribution, even if sampling from the posterior results in less accurate

reconstruction at the sequence level. The experimental implications of this proposal are presented in both Chapters 8 and 15. A brief rebuttal to this view and defense of maximum likelihood is presented by Eric Gaucher in Chapter 2. Further analysis and discussions of this topic are sure to appear in the literature over the coming years.

A third topic raised for discussion at the meeting by Giorgio Matassi was, "Are all proteins reconstructable?" Clearly some proteins, like the green fluorescent protein-like proteins worked on by Mikhail Matz and colleagues (and presented in Chapter 15) are more amenable to experimental study than other proteins. However, functional assays *in vitro* or *in vivo* are indeed available for a great many proteins. Chapter 17 presents a reconstruction back to the last universal ancestor and other chapters deal with various complexities in sequence evolution that will enable more accurate reconstruction.

The first two chapters provide a historical and scientific overview of ancestral sequence reconstruction, and Chapter 3 extends the use of the technique to applications of drug design and mentions the companion technique of substitutional mapping. A discussion of standard approaches for ancestral sequence reconstruction is presented in Chapters 4 and 5, with Chapter 6 presenting a method (with a companion software package) for substitutional mapping.

Chapters 7 and 8 present some of the limitations and considerations that should go into computationally reconstructing ancestors, including methodological sources of bias and biophysical implications. Chapter 9 presents a discussion of covarion or heterotacheous processes, where sites shift rates due to intra- or intermolecular coevolution, and their effects on ancestral sequence reconstruction. Chapter 10 analyzes some controversies in our knowledge of the reference species tree and how different topologies can affect reconstructed ancestral sequences. The covarion processes discussed in Chapter 9, while sometimes neutral, are also sometimes linked to functional shifts. Chapter 11 discusses methodology for linking this process to functional shifts after gene duplication using ancestral sequences. Chapters 12 and 13 present computational strategies and applications of using ancestral sequence data to reconstruct entire proteomes. In work not presented in this book, David Haussler and colleagues have extended this type of approach to reconstructing the entire genome of the last common ancestor of mammals. The thoughtful introduction by Emile Zuckerkandl proposes further extension of the analysis from entire proteomes to interactomes and the field, although not there yet, will surely move in this direction.

Moving to experimental work to test computational hypotheses, Chapters 14 and 15 present strategies for converting computationally reconstructed ancestral sequences to proteins resurrected in the laboratory. Chapter 15 includes an expanded discussion of how to accommodate the controversial computational strategy suggested in Chapter 8. Chapters 16–19 then address various biological questions using ancestral sequence reconstruction and resurrection, across different evolutionary depths and drawing on widely different scientific disciplines.

Rather than presenting a view that is consistent from chapter to chapter, several contradictory views are presented differently by different authors to give readers a chance to appreciate ongoing debates in the field and formulate their own opinions. In the concluding section, I provide a list of several available software packages that are available to perform different analyses described in the book. I also attempt to tie together some of the discussion to present the experimental molecular biologist with a potential way forward in attempting these methods in their own laboratory.

The image of crocodilians on the book cover was generated with the enthusiastic help of John Brueggen at the St. Augustine Alligator Farm Zoological Park (http://www.alligatorfarm.us). The picture shows all 23 extant species of crocodilans and as a Postdoctoral Researcher at University of Florida, I always enjoyed visiting the alligator farm and comparing the species in my mind. While I have never worked with crocodilians, one of the constant battles that my lab has faced is the search for DNA from different closley related species. Ultimately in this process, we are interested in addressing the question, "What were

the molecular events that made each species unique from its closest relatives?'' So, as you look at the crocodilians on the cover of the book, ask yourself how these species are different, what the molecular underpinnings of this are, what the selective forces that drove this were, and how the techniques described in this book can help us answer these questions.

I am grateful to the external reviewers of chapters for this book, notably Aoife McLysaght (Trinity College, Ireland), Arthur Lesk (Pennsylvania State University, USA), my research group (especially Alexander Churbanov and Steven Massey), and my wife Jessica, as well as to authors who took the time to review other chapters in this effort (a special thanks for extra effort go to Eric Gaucher, Tal Pupko, and Denis Shields). I also need to thank Ian Sherman and Stefanie Gehrig at Oxford University Press for their patience. Thank you for your interest in the growing research field.

David A. Liberles
University of Wyoming,
Laramie, WY, USA

Introduction to the Meeting in Kristineberg, Sweden

This introduction was written by Emile Zuckerkandl and read at the meeting by Giorgio Matassi.

Learning about this meeting was exciting news. It is a great honor for me to be permitted to address to you all a wholehearted welcome. I very much regret not to be able to do so in person and to miss out on two full days of attractively diverse and promising contributions to the meeting's theme: ancestral sequence reconstruction and its use in the study of the evolution of protein function. We must hope that the approach referred to by the meeting's title will emerge strengthened rather than weakened, because it appears to be of irreplaceable value to the study of the evolution of informational macromolecules. This study indeed is obliged to resort largely to deductive methods, since a direct determination of ancestral macromolecules is impossible in most cases. It thus becomes very important maximally to clarify the limitations of the methods of inference of ancestral sequences and their higher-order structures as well as to try to devise ways of overcoming the limitations. This is one of the aims of the present meeting. Likewise, it is very important to devise new approaches to help the reconstruction process.

Beyond the methodologies, however, lies the *most* exciting: the analysis of their results. In fact, the greatest interest of the reconstruction of ancestral informational macromolecules may well lie in the reconstruction of their interactions. Let me refer to informational macromolecules as semantides, as proposed 40 years ago, and thus use three syllables for the concept in lieu of ten. The ambition to reconstruct semantide interactions poses great additional challenges. It can be successful only with the help of differing but convergent approaches. One such approach will probably take a large set of deduced structures of individual ancestral semantides and arrange this set according to an ancestral network of evolutionarily conserved semantide interactions. Such networks of ancestral semantide interactions would be inferred from comparisons of whole genomes over various—at times very wide—spans of the phylogenetic tree, taking into account the interactions as observable in contemporary organisms. New technologies will hopefully be developed to help carry out this gigantic task, and the fruits of the labor should justify the effort. By determining the macromolecular interactions that were conserved at different evolutionary times, one might achieve two remarkable feats. The first would be to establish a sort of skeleton of molecular evolution, represented by the conserved interactions among semantides along various evolutionary lines of descent. One would trace along distinct phylogenetic branches the different degrees of persistence of semantide interactions, taking into account variations in the mutual fit of many semantides that despite structural alterations continue to interact functionally. Such differences in interactions are often attributable to mutational damage followed by functional restoration through a modified fit. The second feat would consist of attempts to discover, on the flip side of conservation, the molecular pathways taken at various times by evolutionary novelties. Analyzing the molecular pathways of past evolutionary novelties in the making may well be the greatest challenge of all, but again contemporary organisms are likely, here, to extend a helping hand.

By applying the various methodologies to be further developed—and, in part, yet to be invented—the most fascinating general discovery to be made, it seems to me, is that of the evolution of gene regulation; namely, in particular, of

transcriptional regulation and its evolutionary history. Knowing about the precise pathways of the evolution of the various modes of transcriptional regulation would greatly assist another undertaking that may be considered one of the broadest and deepest aims of biology: understanding more fully the molecular nature and molecular evolution of development. The present meeting may well be considered, albeit indirectly, as a stepping stone in this direction, too.

One may think that all life in the universe has to be built on linear heteropolymers capable of forming higher-order structures and that these structures, possibly with the help of other structures, have to be in turn capable of copying the linear heteropolymers that they contain. Perhaps heteropolymer complementarity—with a sequence strand generating its complementary strand—is likewise an absolutely general aspect of living systems. In light of such a generalization, at any rate, the methods discussed or introduced at the present meeting are likely to be applicable to all forms of life in the universe.

I am glad, right now, that I am not standing on this podium so that I shall not myself receive the tomatoes that may very well be thrown at the speaker at this point. Perhaps, however, there is in this hall at least one scheduled speaker who might agree with the statement just made, since he has put astrobiology in his title. There is little doubt that the potential extensions from the topics of this meeting are vast, whether their vastness approaches infinity or not. Welcome then, once again, to a gathering that I am sure is going to be most substantive, in good measure by its intense and patient scrutiny of detailed mechanics, short of which there can never be a successful completion of any interstellar voyage.

Emile Zuckerkandl
Stanford University and Institute of
Molecular Medical Sciences,
Palo Alto, CA, USA

Contributors

Steven A. Benner, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601, USA

Jonathan P. Bollback, Center for Bioinformatics and Institute of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark

F. Nicholas Braun, Institute of Medical Biology, University of Tromsö, N-9037 Tromsö, Norway

Jamie T. Bridgham, Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403, USA

Dawn J. Brooks, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601, USA

Gina M. Cannarozzi, Institute of Computational Science, ETH Zurich, 8092 Zürich, Switzerland

Belinda S.W. Chang, Departments of Ecology and Evolutionary Biology, and Cell and Systems Biology, University of Toronto, 25 Harbord Street, Toronto, ON M5S 3G5, Canada

Lesley J. Collins, Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Etienne G.J. Danchin, Glycogenomics and Biomedical Strucural Biology, AFMB, UMR 6098 CNRS/Université de Provence/Université de la Méditerrannée, Marseilles, France

Adi Doron-Faigenboim, Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Richard J. Edwards, Bioinformatics, Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland

Steven F. Field, Whitney Laboratory for Marine Bioscience, University of Florida, 9505 Ocean Shore Blvd, Saint Augustine, FL 32080, USA

Toni Gabaldon, Bioinformatics Department, Centro de Investigación Príncipe Felipe, Autopista del Saler 16, 46013 Valencia, Spain

Paul P. Gardner, Center for Bioinformatics and Institute of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark

Eric A. Gaucher, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601, USA

Gaston H. Gonnet, Institute of Computational Science, ETH Zurich, 8092 Zürich, Switzerland

Xun Gu, Department of Genetics, Development and Cell Biology and Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Yong Huang, Department of Genetics, Development and Cell Biology and Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Martijn A. Huynen, Center for Molecular and Biomolecular Informatics and Nijmegen Center for Molecular Life Sciences, University Medical Center St. Radboud, Toernoooiveld 1, 6525 ED Nijmegen, The Netherlands

Catriona R. Johnston, Bioinformatics, Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland

Janos T. Kodra, Novo Nordisk A/S, Novo Alle, 2760 Måløv, Denmark

David A. Liberles, Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA

Peter J. Lockhart, Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand

Dennis Madsen, Novo Nordisk A/S, Novo Alle, 2760 Måløv, Denmark

Mikhail V. Matz, Whitney Section of Integrative Biology, University of Texas at Austin, 1 University Station C0930, Austin, TX 78712, USA

James O. McInerney, Department of Biology, Callan Building, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

Johannes Müller, Humboldt-Universität zu Berlin, Museum für Naturkunde, D-10099 Berlin, Germany

Rasmus Nielsen, Center for Bioinformatics and Institute of Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark

Mary J. O'Connell, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

David D. Pollock, Department of Biochemistry and Molecular Genetics, University of Colorado Health Sciences Center, Aurora, CO 80045, USA

Pierre Pontarotti, Phylogenomics Laboratory, EA 3781 Evolution Biologique, Université de Provence, Marseilles, France.

Tal Pupko, Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Slim O. Sassi, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601, USA

Adrian Schneider, Institute of Computational Science, ETH Zurich, 8092 Zürich, Switzerland

Denis C. Shields, Bioinformatics, Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland

Marie Skovgaard, Novo Nordisk A/S, Novo Alle, 2760 Måløv, Denmark

Joseph W. Thornton, Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, OR 97403, USA

Ilke van Hazel, Department of Ecology and Evolutionary Biology, University of Toronto, 25 Harbord Street, Toronto, ON M5S 3G5, Canada

Iain M. Wallace, Bioinformatics, Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Dublin 4, Ireland

Dongping Xu, Department of Genetics, Development and Cell Biology and Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Ying Zheng, Department of Genetics, Development and Cell Biology and Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA

Emile Zuckerkandl, Department of Biological Sciences, Stanford University and Institute of Molecular Medical Sciences, Palo Alto, CA, USA

This page intentionally left blank

I

Introductory scientific overview

This page intentionally left blank

CHAPTER 1

The early days of paleogenetics: connecting molecules to the planet

Steven A. Benner

1.1 Introduction

Anyone asked to write about the early days feels elderly. Fortunately, Emile Zuckerkandl's introduction shows that the ideas that led to this volume have been around for some time, at least in their basic form, and are rooted in ideas of many heroes of modern molecular biology, including Pauling, Anfinsen, and Zuckerkandl himself.

In 1980, my laboratory was unaware of the Pauling–Zuckerkandl paper (Pauling and Zuckerkandl, 1963; see Chapter 2 in this volume for a fuller discussion of the implications of this paper) when we set out to resurrect ancient proteins from extinct organisms. My group, then consisting of only Krishnan Nambiar and Joseph Stackhouse, was trained to describe the chemical structures and behaviors of enzymes. In those days technology was allowing molecular scientists to extend these descriptions to atomic resolution, the picosecond time scale, and the microscopic rate constant.

But what good were clever experiments to determine, for example, which of two hydrogens was removed by a dehydrogenase (Allemann *et al.*, 1988), or whether the replacement of carbon dioxide by a proton on acetoacetate proceeded with retention or inversion of stereochemical configuration (Benner *et al.*, 1981)? It occurred to us that we might be doing the biochemical equivalent of studying a Picasso with an electron microscope. Were we not describing biomolecular systems to resolutions far greater than they were designed? Biomolecules are not designed, however. They are the products of natural selection imposed upon random variation in their chemical structures. As the result of a combination of historical accident, selective pressures, and vestigiality, all constrained by physical and chemical law, different behaviors must be interesting at different levels. Biomolecular behaviors that influenced the ability of a host organism to survive, mate, and reproduce were especially interesting, as these had been fashioned by natural selection. Behaviors that did not, were not, because they had not. As a criterion for selecting interesting chemical features of a biomolecule to study in detail, an understanding of the relation between biomolecular structure and behavior and fitness was important.

It did not take long at Harvard to realize that this relation was going to be difficult to understand. There, Martin Kreitman, Robert Dorit, and others, including some very dialectical biologists (Levins and Lewontin, 1985), were struggling to make this connection starting from the side of biology (Kreitman and Akashi, 1995). Despite this interest, it was proving difficult to connect *any* biomolecular structure or behavior with the survival of an organism, at least in a way that would be compelling to those who chose to deny it (Lewontin, 1974; Clarke, 1975; Gillespie, 1984, 1991; Somero, 1995; Powers and Schulte, 1998). In fact, the discussion was central to the most hotly disputed dispute in molecular evolution, between neutralists and selectionists, where both sides of the dialectic were populated by individuals who were professionally intent on showing how any data interpretable in favor of one side could equally well support the other.

As chemists, we had no part in this fight. However, a review of the contending sides of these disputes (Benner and Ellington, 1988) reminded us of analogous disputes in organic chemistry.

These were often Seinfeld arguments about nothing. For example, chemists had for years discussed the non-classical carbocation problem (Brown, 1977). This was a disagreement about whether the structures of positively charged organic molecules, in general, were better modeled by a formula with dotted lines, or by two formulas without dotted lines. Rational observers realized that one model was undoubtedly better for some molecules, whereas the other was better for others. After all, similar issues had been addressed and resolved in many molecular systems. For example, the structures of benzene and many boron-containing compounds both contained dotted lines. Which model was best undoubtedly depended on the exact structure of the molecule being discussed. By 1980, this dispute had forced chemists to appreciate a certain truism about molecules: organic molecules are never productively discussed in terms of a general molecular structure; they must always be considered individually. This truism, of course, recognized that the discussion of models for the structure of *individual* molecules could nevertheless be interesting.

To chemists, the neutralist/selectionist dispute was directly analogous. This was essentially a disagreement about whether changes in the chemical structure of the generic protein would, in general, change its behavior enough to change its contribution to the fitness of the generic organism. Again, the rational answer was in some cases yes, and in other cases no, depending on the exact structure of the system. Proteins are, after all, organic molecules, suggesting that they must be considered individually. As expected by those who understood this truism, the neutralist/selectionist dispute, in its general form, melted away as soon as our ability to analyze the behavior of individual proteins improved (Hey, 1999).

Even in 1980, however, it was clear that connecting fitness to the behavior of *individual* biomolecules would always remain interesting, for many reasons. First, that understanding would certainly help us select behaviors of those biomolecules to study in detail. If a behavior was important to fitness, it might be highly optimized. Detailed study might therefore instruct us about the interaction between chemical structure and biomolecular behavior, instruction worthy of the growing armamentarium of biophysics and molecular biology.

1.2 History as an essential tool to understand chemistry

It was clear, however, that Structure Theory in chemistry would not support a deep understanding of biological molecules. With simple molecules, like methane, one does not ask about its purpose. This is not true about complex systems, or living systems, where it is appropriate to ask: *why* does it exist? History can be key to any answer to why? questions. Any system, natural or human-made, can be understood better if we understand *both* its structure *and* its history. We would not understand the QWERTY computer keyboard, the Microsoft Windows operating system, or the US Federal Reserve Bank (for example) if we simply deconstructed each into its parts. An understanding of the history of each is essential to an understanding of the systems themselves.

Structure Theory from chemistry had absolutely no historical component. Methane is how it is because of its structure. It always has been this way, and always will be. Where the methane came from and how it got to us was fully irrelevant to our understanding of this molecule. This raised the next in a series of questions leading to experimental paleogenetics: how were Structure Theory and natural history to be combined to better understand biomolecules? Fragments of the history of life on Earth are found in the geological strata, of course. But the fossil record is notoriously incomplete, and would not provide information about proteins even were it not. Molecular fossils (such as those found in petroleum) can be informative, but generally not about individual protein function. Further, any analysis of molecular function must recognize that the behaviors that confer fitness are determined by the system, including other organisms (ecology), the physical environment (planetary biology), and even the cosmos (astrobiology). This level of complexity defeats most theoretical contexts.

It was clear, however, that the chemical structures of proteins themselves contain historical

information. The historical relationships between proteins related by common ancestry can be inferred by comparing their amino acid sequences, a theme that was already well developed by 1980 (Dayhoff *et al*., 1978). Analysis of protein sequences could generate the basic elements of an evolutionary model: a multiple sequence alignment, a tree, and sequences of ancestral protein sequences inferred from these. From these, it might be possible to construct narratives connecting biomolecular structure to fitness.

This process was analogous to processes well known in the field of historical linguistics (Lehman, 1973), which Robert Breedlove had described to me when I was an undergraduate. This field infers the features in ancestral languages by analyzing the features of their descendent languages. For example, the Proto-Indoeuropean word for snow (**sneig^{w}h-*) can be reconstructed from the descendant words for snow in the descendant Indoeuropean languages (German *schnee*, French*neige*, Irish *sneachta*, Russian *sneg*, Sanskrit *snihyati*, and so on). Other features of the histories of these languages, such as the universal replacement of *sn-* by *n-* in the Romance languages, can also be inferred from this analysis. The analogous inferences about ancestral structures could also be done for proteins.

The reconstruction of ancestral languages provides paleoanthropological information as well. From the ancestral features of reconstructed ancestral languages, one can extract information about the people who spoke them. For example, the ease with which we reconstructed the Proto-Indoeuropean word for snow (with some concessions; the Sanskrit word cited above actually means "he gets wet") tells the story that the Proto-Indoeuropeans themselves lived in a locale where it snowed. In 1980, we hoped to tell analogous stories using proteins inferred to have been present in ancestral forms of life on Earth.

1.3 Swapping places: biologists become chemists and statisticians, just as chemists become natural historians

But would these be only just-so stories? The just-so story is one of the worst insults that a biologist can direct at another. This epithet accuses a professional adversary of building *ad hoc* explanations for specific facts (how the zebra got his stripes). The events behind a just-so story (an ancestral zebra took a nap under a ladder) cannot be independently verified, and are not mathematically modelable. Further, the story could easily be replaced by a different story, just as compelling, had the observations been the opposite. It was clear in 1980 that once the insult stuck, papers would be rejected, grant applications would be turned down, and tenure would be denied.

Curiously, this issue also had a parallel in organic chemistry. Chemists are well known for their ability to use Structure Theory to explain a set of facts, only to be told that the facts are opposite, and then to explain the counter-facts using the same theory. Chemists are rarely defensive about this. In part, this is because Structure Theory as a heuristic has been so successful. If one can make petrochemicals and pharmaceuticals (and much in between) using a theory based on plastic tinker toy models, who can argue?

The success of non-mathematical Structure Theory from chemistry makes a larger point about human knowledge; that it is intrinsically heuristic and intuitive. This is true even for knowledge that is cast in the language of mathematics. This conclusion had been reached by the last century of epistemology as well (Suppe, 1977). Nothing is "proven" (Galison, 1987); the perception of proof is only a function of the number of logical steps that must be taken to premises that are intuitive and heuristic. Experiments end when a burden of proof is met, where that burden is defined by the culture, not by logic.

This point is not fully appreciated by many modern biologists. Many modern biologists seek to avoid the just-so story epithet, and the perception of theirs being a heuristic and/or intuitive theory, by placing a mathematical formalism on top of their models. This drives them towards statistics, which analyzes collections of things. Statistics, in turn, nearly always requires the statistician to deny the truism in chemistry that there is no such thing as general molecular behavior. This, in turn, means that statisticians, in their pursuit of general models framed in mathematical language, are not able to