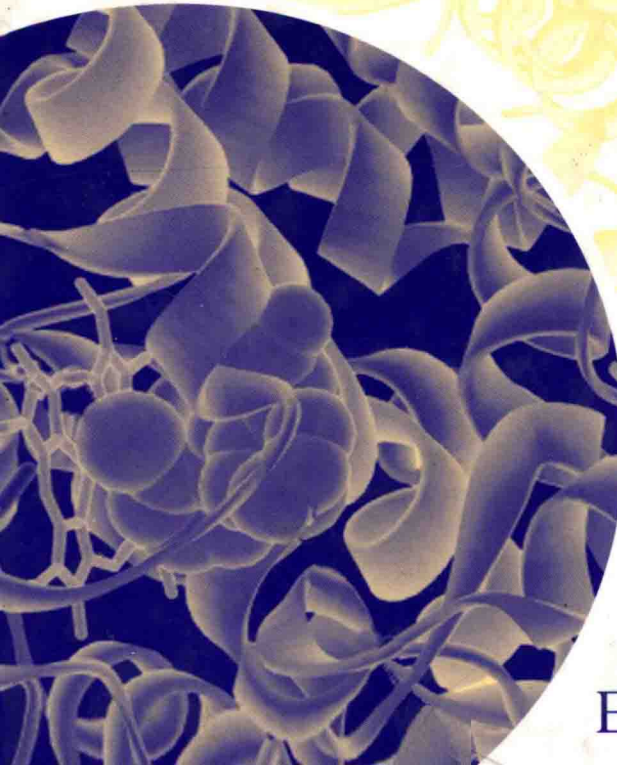


—●—

# Advances in Enzymology and Related Areas of Molecular Biology

—●—



Volume 75  
Protein  
Evolution

ERIC J. TOONE

# ADVANCES IN ENZYMOLOGY

*AND RELATED AREAS OF MOLECULAR BIOLOGY*

**Founded by F. F. NORD**

**Edited by ERIC J. TOONE**

DUKE UNIVERSITY, DURHAM, NORTH CAROLINA

**VOLUME 75 PROTEIN EVOLUTION**



WILEY-INTERSCIENCE  
A JOHN WILEY & SONS, INC. PUBLICATION

Copyright © 2007 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey  
Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at [www.copyright.com](http://www.copyright.com). Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, 201-748-6011, fax 201-748-6008, or online at <http://www.wiley.com/go/permission>.

**Limit of Liability/Disclaimer of Warranty:** While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at 877-762-2974, outside the United States at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at [www.wiley.com](http://www.wiley.com).

***Library of Congress Cataloging-in-Publication Data is available:***

ISBN-13: 978-0-471-20503-6  
ISBN-10: 0-471-20503-6

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

**ADVANCES IN ENZYMOLOGY**

AND RELATED AREAS OF  
MOLECULAR BIOLOGY

Volume 75



---

THE WILEY BICENTENNIAL—KNOWLEDGE FOR GENERATIONS

---

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

**WILLIAM J. PESCE**  
PRESIDENT AND CHIEF EXECUTIVE OFFICER

**PETER BOOTH WILEY**  
CHAIRMAN OF THE BOARD

## CONTRIBUTORS

- PATRICIA C. BABBITT, Department of Biopharmaceutical Sciences, University of California—San Francisco, San Francisco, CA 94143
- STEVEN A. BENNER, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601
- ERIC J. DEEDS, Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138
- ERIC A. GAUCHER, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601
- JOHN A. GERLT, Departments of Biochemistry and Chemistry, University of Illinois—Urbana, Urbana, IL 61801
- MARGARET E. GLASNER, Department of Biopharmaceutical Sciences, University of California—San Francisco, San Francisco, CA 94143
- DONALD HILVERT, Laboratory of Organic Chemistry, ETH—Zürich, CH-8093 Zürich, Switzerland
- SLIM O. SASSI, Foundation for Applied Molecular Evolution, 1115 NW 4th Street, Gainesville, FL 32601
- EUGENE I. SHAKHNOVICH, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138
- KATHERINA VAMVACA, Laboratory of Organic Chemistry, ETH—Zürich, CH-8093 Zürich, Switzerland
- KENNETH J. WOYCECHOWSKY, Laboratory of Organic Chemistry, ETH—Zürich, CH-8093 Zürich, Switzerland

## PREFACE

The nature of structure-function relationships in the context of protein catalysis is among the holy grails of molecular biological science. The second half of the 20th century heralded the arrival of an unprecedented volume of structural information, as both diffraction and resonance techniques became accessible to a large group of researchers. Almost simultaneously, the burgeoning toolkit of modern molecular biology facilitated the rapid redesign of existing protein scaffolds. Together, these tools would surely unlock the secrets of protein function.

But the path to the grail has been arduous. Despite tremendous advances in both protein folding and the development of protein catalysts, our abilities to design protein sequence with a priori predictable function remain largely rudimentary. Against this backdrop, evolutionary approaches emerged as a promising tool for the *de novo* creation of catalytic activity. Originally conceived in the context of antibodies, the concepts and approaches were later applied to the evolution — or redesign — of existing enzyme scaffolds. And while the development of synthetically useful entities remains an important end unto itself, the opportunity to uncover fundamental aspects of the structural basis of protein catalysis is perhaps the most powerful motivation for the study of enzyme evolution.

In this volume, four strikingly different perspectives on protein evolution are presented. The first chapter by Benner and coworkers offers a retrospective approach to protein evolution. The revival of sequence data from historical sources, dating back millions of years, provides a unique opportunity to examine the course of “successful” evolutionary excursions. In the second chapter, Deeds and Shakhnovich consider the larger question of protein architectures, and explore attempts to provide a theoretical basis for the existence of a relatively small number of protein folds and processes by which new folds arise.

The last two chapters consider more immediate aspects of protein evolution. In the first, Glasner, Babbitt, and Gerlt explore the notion of structural conservation in enzyme catalysis and approaches to the selection of a catalyst scaffold during evolutionary selection. These authors also

explore the hypothesis that new catalytic activities arise through the intermediacy of promiscuous progenitors, rather than by the more direct path from one specificity to another. In the final chapter, Hilvert and coworkers review the various experimental approaches to novel enzyme activity, including the *de novo* development of catalytic activity and the redesign of existing activities, in both cases through random and guided approaches.

Together, the chapters of this volume consider both fundamental and practical issues associated with the development of novel catalytic activity and, more broadly, the development of robust structure-function relationships in protein catalysis. We hope that the ideas presented here further stimulate the evolution of protein redesign.

Eric J. Toone  
Durham, NC



## ABSTRACTS

### **Molecular Paleoscience: Systems Biology from the Past**

Experimental paleomolecular biology, paleobiochemistry, and paleogenetics are closely related emerging fields that infer the sequences of ancient genes and proteins from now-extinct organisms, and then resurrect them for study in the laboratory. The goal of paleogenetics is to use information from natural history to solve the conundrum of modern genomics: How can we understand deeply the function of biomolecular structures uncovered and described by modern chemical biology? Reviewed here are the first 20 cases where biomolecular resurrections have been achieved. These show how paleogenetics can lead to an understanding of the function of biomolecules, analyze changing function, and put meaning to genomic sequences, all in ways that are not possible with traditional molecular biological studies.

### **A Structure-Centric View of Protein Evolution, Design, and Adaptation**

Proteins, by virtue of their central role in most biological processes, represent one of the key subjects of the study of molecular evolution. Inherent in the indispensability of proteins for living cells is the fact that a given protein can adopt a specific three-dimensional shape that is specified solely by the protein's sequence of amino acids. Over the past several decades, structural biologists have demonstrated that the array of structures that proteins may adopt is quite astounding, and this has led to a strong interest in understanding how protein structures change and evolve over time. In this review we consider a large body of recent work that attempts to illuminate this structure-centric picture of protein evolution. Much of this work has focused on the question of how completely new protein structures (i.e., new folds or topologies) are discovered by protein sequences as they evolve. Pursuant to this question of structural innovation has been a desire to describe and understand the observation that certain types of protein structures are far more abundant than others and how this uneven

distribution of proteins implicates on the process through which new shapes are discovered. We consider a number of theoretical models that have been successful at explaining this heterogeneity in protein populations and discuss the increasing amount of evidence that indicates that the process of structural evolution involves the divergence of protein sequences and structures from one another. We also consider the topic of protein designability, which concerns itself with understanding how a protein's structure influences the number of sequences that can fold successfully into that structure. Understanding and quantifying the relationship between the physical feature of a structure and its designability has been a long-standing goal of the study of protein structure and evolution, and we discuss a number of recent advances that have yielded a promising answer to this question. Finally, we review the relatively new field of protein structural phylogeny, an area of study in which information about the distribution of protein structures among different organisms is used to reconstruct the evolutionary relationships between them. Taken together, the work that we review presents an increasingly coherent picture of how these unique polymers have evolved over the course of life on Earth.

### **Mechanisms of Protein Evolution and Their Applications to Protein Engineering**

Protein engineering holds great promise for the development of new biosensors, diagnostics, therapeutics, and agents for bioremediation. Despite some remarkable successes in experimental and computational protein design, engineered proteins rarely achieve the efficiency or specificity of natural enzymes. Current protein design methods utilize evolutionary concepts, including mutation, recombination, and selection, but the inability to fully recapitulate the success of natural evolution suggests that some evolutionary principles have not been fully exploited. One aspect of protein engineering that has received little attention is how to select the most promising proteins to serve as templates, or scaffolds, for engineering. Two evolutionary concepts that could provide a rational basis for template selection are the conservation of catalytic mechanisms and functional promiscuity. Knowledge of the catalytic motifs responsible for conserved aspects of catalysis in mechanistically diverse superfamilies could be used to identify promising templates for protein engineering. Second, protein evolution often proceeds through promiscuous intermediates, suggesting that templates which are naturally promiscuous for a target

reaction could enhance protein engineering strategies. This review explores these ideas and alternative hypotheses concerning protein evolution and engineering. Future research will determine if application of these principles will lead to a protein engineering methodology governed by predictable rules for designing efficient, novel catalysts.

### **Novel Enzymes Through Design and Evolution**

The generation of enzymes with new catalytic activities remains a major challenge. So far, several different strategies have been developed to tackle this problem, including site-directed mutagenesis, random mutagenesis (directed evolution), antibody catalysis, computational redesign, and *de novo* methods. Using these techniques, a broad array of novel enzymes has been created (aldolases, decarboxylases, dehydratases, isomerases, oxidases, reductases, and others), although their low efficiencies (10 to 100  $M^{-1} s^{-1}$ ) compared to those of the best natural enzymes ( $10^6$  to  $10^8$   $M^{-1} s^{-1}$ ) remains a significant concern. Whereas rational design might be the most promising and versatile approach to generating new activities, directed evolution seems to be the best way to optimize the catalytic properties of novel enzymes. Indeed, impressive successes in enzyme engineering have resulted from a combination of rational and random design.

## CONTENTS

Contributors . . . . .	vii
Preface . . . . .	ix
Abstracts . . . . .	xi
Molecular Paleoscience: Systems Biology from the Past . . . . .	1
<i>Steven A. Benner, Slim O. Sassi, and Eric A. Gaucher</i>	
A Structure-Centric View of Protein Evolution, Design, and Adaptation . . . . .	133
<i>Eric J. Deeds and Eugene I. Shakhnovich</i>	
Mechanisms of Protein Evolution and Their Application to Protein Engineering . . . . .	193
<i>Margaret E. Glasner, John A. Gerlt, and Patricia C. Babbitt</i>	
Novel Enzymes Through Design and Evolution . . . . .	241
<i>Kenneth J. Woycechowsky, Katherina Vamvaca, and Donald Hilvert</i>	
Author Index . . . . .	295
Subject Index . . . . .	301

# MOLECULAR PALEOSCIENCE: SYSTEMS BIOLOGY FROM THE PAST

By STEVEN A. BENNER, SLIM O. SASSI, and  
ERIC A. GAUCHER, *Foundation for Applied Molecular Evolution,*  
1115 NW 4th Street, Gainesville, FL 32601

## CONTENTS

- I. Introduction
  - A. Role for History in Molecular Biology
  - B. Evolutionary Analysis and the “Just So” Story
  - C. Biomolecular Resurrections as a Way of Adding to an Evolutionary Narrative
- II. Practicing Experimental Paleobiochemistry
  - A. Building a Model for the Evolution of a Protein Family
    - 1. Homology, Alignments, and Matrices
    - 2. Trees and Outgroups
    - 3. Correlating the Molecular and Paleontological Records
  - B. Hierarchy of Models for Modeling Ancestral Protein Sequences
    - 1. Assuming That the Historical Reality Arose from the Minimum Number of Amino Acid Replacements
    - 2. Allowing the Possibility That the History Actually Had More Than the Minimum Number of Changes Required
    - 3. Adding a Third Sequence
    - 4. Relative Merits of Maximum Likelihood Versus Maximum Parsimony Methods for Inferring Ancestral Sequences
  - C. Computational Methods
  - D. How Not to Draw Inferences About Ancestral States
- III. Ambiguity in the Historical Models
  - A. Sources of Ambiguity in the Reconstructions
  - B. Managing Ambiguity
    - 1. Hierarchical Models of Inference
    - 2. Collecting More Sequences

---

*Advances in Enzymology and Related Areas of Molecular Biology, Volume 75:*  
*Protein Evolution* Edited by Eric J. Toone  
Copyright © 2007 John Wiley & Sons, Inc.

3. Selecting Sites Considered to Be Important and Ignoring Ambiguity Elsewhere
  4. Synthesizing Multiple Candidate Ancestral Proteins That Cover, or Sample, the Ambiguity
- C. Extent to Which Ambiguity Defeats the Paleogenetic Paradigm
- IV. Examples
- A. Ribonucleases from Mammals: From Ecology to Medicine
1. Resurrecting Ancestral Ribonucleases from Artiodactyls
  2. Understanding the Origin of Ruminant Digestion
  3. Ribonuclease Homologs Involved in Unexpected Biological Activities
  4. Paleobiochemistry with Eosinophil RNase Homologs
  5. Paleobiochemistry with Ribonuclease Homologs in Bovine Seminal Fluid
  6. Lessons Learned from Ribonuclease Resurrections
- B. Lysozymes: Testing Neutrality and Parallel Evolution
- C. Ancestral Transposable Elements
1. Long Interspersed Repetitive Elements of Type I
  2. Sleeping Beauty Transposon
  3. Frog Prince
  4. Biomedical Applications of Transposons
- D. Chymase–Angiotensin Converting Enzyme: Understanding Protease Specificity
- E. Resurrection of Regulatory Systems: The Pax System
- F. Visual Pigments
1. Rhodopsins from Archaeosaurs: An Ancestor of Modern Alligators and Birds
  2. History of Short Wavelength–Sensitive Type 1 Visual Pigments
  3. Green Opsin from Fish
  4. Blue Opsins
  5. Planetary Biology of the Opsins
- G. At What Temperature Did Early Bacteria Live?
1. Elongation Factors
  2. Isopropylmalate and Isocitrate Dehydrogenases
  3. Conclusions from “Deep Time” Paleogenetic Studies
- H. Alcohol Dehydrogenase: Changing Ecosystem in the Cretaceous
- I. Resurrecting the Ancestral Steroid Receptor and the Origin of Estrogen Signaling
- J. Ancestral Coral Fluorescent Proteins
- K. Isocitrate Dehydrogenase
- V. Global Lessons
- References

## I. INTRODUCTION

### A. ROLE FOR HISTORY IN MOLECULAR BIOLOGY

The structures that we find in living systems are the outcomes of random events. These are filtered through processes described by population dynamics and through natural selection to generate macroscopic, microscopic, and molecular physiology. The outcomes are, of course, constrained by physical and chemical law. Further, the outcome is limited by the Darwinian strategy by which natural selection superimposed on random variation searches for solutions to biological problems. The Darwinian strategy need not deliver the best response to an environmental challenge; indeed, it may deliver no response that allows a species to avoid extinction. The outcomes of evolutionary mechanisms therefore reflect history as much as optimization.

It is therefore not surprising that biology finds its roots in natural history. The classical fields in this classical discipline include systematic zoology, botany, paleontology, and planetary science. Here, seemingly trivial details (such as the physiology of the panda's thumb) have proven enlightening to naturalists as they attempt to understand the interplay of chance and necessity in determining the outcome of evolution (Gould, 1980; Glenner et al., 2004). These roots in natural history are not felt as strongly in modern molecular biology, however. Molecular biology emerged in the twentieth century as an alliance between biology and chemistry. The alliance has been enormously productive, but largely without reference to systematics, history, or evolution. Today, we have the chemical structures of millions of biomolecules and their complexes: as small as glucose and as large as the human genome (Venter et al., 2001). X-ray crystallography and nuclear magnetic resonance spectroscopy locate atoms within biomacromolecules with precisions of tenths of nanometers. Biophysical methods measure the time course of biological events on a microsecond scale (Buck and Rosen, 2001). These and other molecular characterizations, written in the language of chemistry, have supported industries such as drug design and foodstuff manufacture, all without any apparent need to make reference to the history of their molecular components or the evolutionary processes that generated them.

The success of this reductionist approach has caused many molecular biologists to place a lower priority on historical biology. Indeed, the archetypal molecular biologist has never studied systematics, paleontology, or Earth science. The combination of chemistry and biology has generated

so much excitement that history seems no longer to be relevant, and certainly not necessary, to the practice of life science or the training of life scientists.

Nearly overlooked in the excitement, however, has been the failure of molecular characterization, even the most detailed, to generate something that might be called “understanding.” The human genome provides an example of this. The genome is itself nothing more (and nothing less) than a collection of natural product structures. Each structure indicates how carbon, hydrogen, oxygen, nitrogen, and phosphorus atoms are bonded within a molecule that is special only in that it is directly inherited. It has long been known to natural product chemists that such biomolecular structures need not make statements about the function of the biomolecule described, either in its host organism or as the host organism interacts with its environment to survive and reproduce. This has proven to be true for genomic structures as well.

Genomic sequences do offer certain opportunities better than other natural product structures when it comes to understanding their function. Comparisons of the structures of genes and proteins can offer models for their histories better than comparisons of the structures of other natural products (Hesse, 2002). As was recognized nearly a half century ago by Pauling and Zuckerkandl (1963), a degree of similarity between two gene or protein sequences indicates, to a degree of certainty, that the two proteins share a common ancestor. Two homologous gene sequences may be aligned to indicate where a nucleotide in one gene shares common ancestry with a nucleotide in the other, both descending from a single nucleotide in an ancestral gene. An evolutionary tree can be built from an alignment of many sequences to show their familial relationships. The sequences of ancestral genes represented by points throughout the trees can be inferred, to a degree of certainty, from the sequences of the descendent sequences at the leaves of the tree.

The history that gene and protein sequences convey can then be used to understand their function. In its most general form, the strategy exploits the truism that any system, natural or human-made, from the QWERTY keyboard to the Federal Reserve banking system, can be better understood if one understands *both* its structure *and* its history.

Much understanding can come first by analyzing the sets of homologous sequences themselves. Thus, credible models for the folded structure of a protein can be predicted from a detailed analysis of the patterns of variation and conservation of amino acids within an evolutionary family



(Benner et al., 1997a; Gerloff et al., 1999; Rost, 2001), if these are set within a model of the history of the family (Thornton and DeSalle, 2000). The quality of these predictions has been demonstrated through their application to protein structure prediction contests as well as through the use of predicted structures to detect distant protein homologs (Benner and Gerloff, 1991; Gerloff et al., 1997; Tauer and Benner, 1997; Dietmann and Holm, 2001). More recently, analysis of patterns of variation and conservation in genes is used to determine whether the gross function of a protein is changing and which amino acids are involved in the change (Gaucher et al., 2001, 2002; Bielawski and Yang, 2004).

Computer analysis of protein sequences from an evolutionary perspective has emerged as a major activity in the past decade. Here, sets of protein sequences are studied computationally within the context of an evolutionary model in an effort to better connect evolving sequences with changing function. Our purpose is to review strategies that go *beyond* simple computational manipulation of gene and protein sequences. In this review we explore *experiment* as a way to exploit the history captured within the chemical structures of DNA and protein molecules.

Our focus will be the emerging field known variously as *experimental paleogenetics*, *paleobiochemistry*, *paleomolecular biology*, and *paleosystems biology*. Practitioners of the field resurrect ancient biomolecular systems from now-extinct organisms for study in the laboratory. The field was started 20 years ago (Nambiar et al., 1984; Presnell and Benner, 1988; Stackhouse et al., 1990) for the specific purpose of joining information from natural history, itself undergoing a surge of activity, to the chemical characterization of biomolecules, with the multiple intents of helping molecular biologists select interesting research problems, generating hypotheses and models to understand the molecular features of biomolecular systems, and providing a way of experimentally testing historical models.

The field has now explored approximately a dozen biomolecular systems (Table 1). These include digestive proteins (ribonucleases, proteases, and lysozymes) in ruminants to illustrate how digestive function arose from nondigestive function in response to a changing global ecosystem, fermentive enzymes from fungi to illustrate how molecular adaptation supported mammals as they displaced dinosaurs as the dominant large land animals, pigments in the visual system adapting to function optimally in different environments, steroid hormone receptors adapting to changing function in steroid-based regulation of metazoans, and proteins