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

Advances in Enzymology and Related Areas of Molecular Biology, Volume 75: Protein Evolution Edited by Eric J. Toone

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 poven 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

TABLE 1
Examples of Molecular Resurrections[a]

Extant Genes	Ancestral Gene Resurrected	Approximate Age (million years)	References
Digestive ribonucleases	Ancestor of buffalo and ox	5	Benner, 1988; Stackhouse et al., 1990
Digestive ribonucleases	Digestive RNases in the first ruminants	40	Jermann et al., 1995
Lysozyme	Ancestral bird lysozyme	10	Malcolm et al., 1990
L1 retroposons in mouse	Ancestral rodent retrotransposon	6	Adey et al., 1994
Chymase proteases	Ancestral ortholog in LCA of mammals	80	Chandrasekaran et al., 1996
Sleeping Beauty transposon	Active ancestral transposon from fish	10	Ivics et al., 1997
Tc1/mariner transposons	Ancestral paralog genomes of eight salmonids	10	Ivics et al., 1997
Immune RNases	Ancestral ortholog LCA of higher primates	31	Zhang and Rosenberg, 2002
Pax transcription factors	Ancestral paralog	600	Sun et al., 2002
SWS1 visual pigment	Ortholog in LCA of bony vertebrates	400	Shi and Yokoyama, 2003
Vertebrate rhodopsins	Archosaur opsins	240	Chang et al., 2002
Fish opsins (blue, green)	Fish opsins	30–50	Chinen et al., 2005b
Steroid hormone receptors	Ancestral paralog	600	Thornton et al., 2003
Yeast alcohol dehydrogenase	Enzyme at origin of fermentation	80	Thomson et al., 2005
Green fluorescent proteins	Ancient fluorescent proteins	ca. 20?	Ugalde et al., 2004
Isocitrate dehydrogenase	Ancestral eubacteria	2500	Zhu et al., 2005
Isopropylmalate dehydrogenase	Ancestral archaebacteria	2500	Miyazaki et al., 2001
Isocitrate dehydrogenase	Ancestral archaebacteria	2500	Iwabata et al., 2005
Elongation factors	LCA of eubacteria	3500	Gaucher et al., 2003

[a]LCA, last common ancestor. Ages are approximate, and in some cases conjectural.

from very ancient bacteria, helping to define environments where the earliest forms of bacterial life lived.

To date, approximately 20 narratives have emerged where specific molecular systems from extinct organisms been resurrected for study in the laboratory. In general, understanding delivered by experimental paleogenetics was not accessible in other ways. After a brief introduction of the strategies and problems in experimental paleoscience, we review each of these narratives.

Our goal here is also to strengthen awareness among molecular and biomedical scientists of the ability of experimental paleogenetics to place meaning on biological data. We believe that current efforts falling under the rubric of ''systems biology'' can be complemented and strengthened when combined with a historical prespective. Understanding will not, we suspect, arise from still more, and still more quantitative, molecular, chemical, and geometric characterization of cellular, organ, and organism-defined systems. At the very least, the analysis must go further, to include the organism, the ecosystem, and the physical environment, which extends from the local habitat to the planet and the cosmos (Feder and Mitchell-Olds, 2003). Without an understanding of the history, we expect that efforts in reductive systems biology are likely to fall short of their promise to deliver understanding.

The same applies to the broader scientific community. Reviewing the first paleomolecular resurrections (Stackhouse et al., 1990; Jermann et al., 1995) a decade ago, Nicholas Wade, writing in the *New York Times Magazine* (Wade, 1995), expressed displeasure. ''The stirring of ancient artiodactyl ribonucleases,'' he wrote, ''is a foretaste of biology's demiurgic powers.'' He then suggested that biomolecular ''resurrection [remain] an unroutine event.'' Given the absence of hazard presented by paleomolecular resurrections (*pace Jurassic Park*), it seems unwise to forgo understanding that comes from bringing into the laboratory for study biomolecules from the past. As the size of the genome sequence database grows, and the gap between chemical data compilations and biological understanding increases, we suspect that experimental paleogenetics will be the key tool to bridge the gap.

B. EVOLUTIONARY ANALYSIS AND THE ''JUST SO'' STORY

After a half century of reductionist biology, evolutionary analysis must struggle to enter the mainstream of discussion within the molecular

sciences. As Adey noted a decade ago, many scientists view evolutionary hypotheses as being inherently resistant to experimental test and therefore fundamentally nonscientific (Adey et al., 1994). Certainly, neither the sequence nor the behavior of a protein from an organism that went extinct a billion years ago can be known with the same precision as the sequence or behavior of a descendent living today. But much in science is useful even though it is not known with certainty. Our job in paleogenetics, as with all science, is to manage whatever uncertainty we encounter.

The arcane nature of the most prominent debate in molecular evolution has also led many molecular biologists to shun evolutionary discussions. Many biomedical researchers do not understand how their view of biology might be different if it turns out that dog diverged from humans and rodents before or after humans and rodents themselves diverged. Similarly, it has always been unproductive in chemistry to ask whether alterations in chemical structure *generally* have any specific impact on behavior; productive discussions in chemistry focus on specific chemical structures and specific changes. Yet the neutralist–selectionist debate that consumed molecular evolution for nearly a decade asked precisely such questions (Hey, 1999).

Further, the accidents that shape molecular biology cannot be reproduced in the laboratory; they may, in fact, never be known in detail. Complex biological systems are chaotic. Small differences in input can have large impacts on the output. The notion that biology would be rather different had a fly flapped his wings differently in the Triassic is a compelling reason to marginalize historical narratives (Lorenz, 1963, 1969).

Further difficulties are encountered with historical narratives, even when they do explain a biological fact. Here, they can easily be viewed as ''just so'' stories. This epithet is pejorative; it indicates that the narrative is constructed ad hoc to explain a specific fact (how the zebra got his stripes), makes no reference to facts verifiable outside the fact being explained (we have no way to verify that an ancestral zebra took a nap under a ladder), and could easily be replaced by a different story, just as compelling, explaining a different observation (if the modern zebra had spots, the story might be that the ancestral zebra took a nap beneath a philodendron such as *Monstera friedrichsthalii*).

Any narrative can be made more compelling by bringing many different types of data to bear on a single system. A recent example of this use of multiple lines of evidence attempted to bring biological meaning to the fact

that modern swine have not one, but rather three, different genes for the enzyme *aromatase* (Gaucher et al., 2004). Aromatases use cytochrome P450 and molecular oxygen to convert androgenic steroids into estrogenic steroids. These steroids play roles throughout vertebrate reproductive biology.

Most mammals (including humans) have only a single aromatase gene. Pigs, for some reason, have three. Here, a complete molecular characterization of the three pig aromatases (the sequences of all three are known) did not address the question: Why do pigs have three genes for an enzyme, all catalyzing approximately the same reaction? A historical narrative was needed. To build this narrative, a cladistic analysis was used to suggest that the two duplications that created the three paralogous aromatase genes in swine occurred after the suids diverged from oxen about 60 million years ago (Ma) (Arnason et al., 1998; Kumar and Hedges, 1998; Foote et al., 1999). A silent transition redundant clock (Benner, 2003) was then applied to date the duplications at 39 to 26 Ma, in the late Eocene to mid-Oligocene. To help define further the timing of the duplications, gene fragments were sequenced from other close relatives of the pig, including the peccary and the babirousa.

Given the timing of the duplications that generated the three genes, it was concluded that the three aromatases in pig were not needed to manage the fundamental reproductive endocrinology of mammals, which arose very early in vertebrates (perhaps earlier than 400 Ma). Nor did the three aromatase genes arise in response to the domestication of pigs, which occurred only a few thousand years ago.

Rather, the date of the duplication events correlated with the time in history when the size of pig litters increased (Figure 1). This timing was suggested by a cladistic analysis of reproductive physiology and litter size and the fossil record, which includes fossils of pregnant ancestral animals (O'Harra, 1930; Franzen, 1997). Since the period following the Eocene was a time of global cooling, it is possible that an increase in litter size was an adaptive change that contributed to the fitness of the pig lineage under a changing climate.

From this analysis, an explanatory narrative emerged. The narrative hypothesized that the gene duplications gave rise to functionally different isozymes of aromatase that played specific roles to establish and maintain large litter sizes in pigs.

At this point, most biologists would regard this as simply a ''just so'' story. To avoid this epithet, structural biology was then recruited. A crystal

FIGURE 1. Old World fossil suid shortly after the emergence of large litter sizes.

structure was not available for aromatase. Nevertheless, an approximate homology model could be built from a homologous P450-dependent enzyme. The sites in the aromatase protein that changed immediately after the aromatase genes had duplicated were then mapped onto the homology model. This map showed that many of the amino acid replacements occurring shortly after the duplications were in or near the substrate binding site. This, in turn, suggested that the substrate–product specificity of the aromatase enzyme changed at this time. This change was confirmed by experiment, where the ability of the different aromatases to catalyze the oxidation of different androgens was shown to differ (Corbin et al., 2004).

The diverging catalytic properties of the aromatase proteins were then observed to be consistent with the different physiologies of the three enzymes. One aromatase is expressed in ovaries, as are aromatases in other mammals (including human).

Another of the aromatases was expressed in the pig embryo, between days 11 and 13 following conception. This is approximately the time of implantation. It was proposed that estrogen created by this aromatase isoform helped space multiple embryos around the uterine wall.

The third aromatase duplicate was expressed primarily in the placenta. The estrogen that this aromatase releases was proposed to help the mother

determine whether the pregnancy should be terminated, which happens in pigs if too few embryos implant successfully.

Together, the narrative combined sequencing data, molecular evolution, and paleontology with the geological records, genomic sequence analysis, structural biology, experimental biochemistry, and reproductive physiology, to create an answer to the question: Why do pigs have three aromatases?

Some commentaries view this use of many lines of evidence of many different types as able to elevate this narrative beyond a ''just so'' story. One commentary even called the combination a *tour de force* (Faculty of 1000, 2004). Nevertheless, the argument embedded in the narrative remains correlative. Correlations do not compel causality. It would therefore be helpful to have an additional tool to expand such narratives further.

C. BIOMOLECULAR RESURRECTIONS AS A WAY OF ADDING TO AN EVOLUTIONARY NARRATIVE

Experimental paleogenetics provides such a tool. Here, the sequences of ancestral DNA or protein sequences are inferred from the sequences of their descendents. Then, using molecular biological methods, an ancestral gene sequence is synthesized in the laboratory and delivered to a biological host where its behavior can be studied. If the interesting ancestral biomolecule is a protein, the gene is expressed to deliver the corresponding ancestral protein, which is then isolated for study.

Biomolecular resurrections allow the scientist to partially relive past events. By doing so, we hope to generate evidence to confirm or deny a hypothesis about past or current function. For example, we might hypothesize that lysozymes or ribonucleases in modern oxen function to support ruminant digestion. This hypothesis might be supported by arguments analogous to those used for aromatase if the ribonucleases and lysozymes emerged at the time when ruminant digestion arose. But a paleobiochemical experiment with ribonucleases and lysozymes resurrected from an animal just beginning to create ruminant digestion may contribute more. If the experimental properties of these resurrected ancient enzymes show that particular digestive behaviors emerged in the ancient proteins at the time when ruminant digestion arose in the ancestral organisms, the narrative is strengthened.

Alternatively, we might suspect that ancestral bacteria lived at an elevated temperature. Resurrection of proteins from those ancestors, and

determining the temperature at which they function best, might confirm or deny that hypothesis.

II. PRACTICING EXPERIMENTAL PALEOBIOCHEMISTRY

We review next the process of drawing inferences about the past from information obtained from the present. More details surrounding the process can be found in the literature (Benner, 2003).

A. BUILDING A MODEL FOR THE EVOLUTION OF A PROTEIN FAMILY

1. Homology, Alignments, and Matrices

The basic element of an evolutionary analysis is the pairwise sequence alignment. Here, two gene or protein sequences are written next to each other so that their similarities are the most perspicuous.

Although any two sequences can be aligned, an alignment makes evolutionary sense only if the two sequences themselves share a common ancestor, that is, if the two sequences are *homologous*. Indeed, the goal of sequence alignment is often to determine whether or not two sequences are homologous. To make this determination, all possible alignments of two sequences are made, each is given a score, and the alignment with the highest score is chosen and examined statistically to determine if the score surpasses a threshold to answer the question: Are these sequences related by common ancestry? (yes or no).

An alignment is *correct* if it accurately represents the history of individual aligned sites. It is an assumption (often arguable) that the highest-scoring alignment is the correct alignment. If so, it is possible to map the sites in the sequence of two homologous descendents of an ancestral protein onto the sites in the sequence of the ancestor. Except for sites that are gained by insertions, each site in a descendent is mappable into a single site in the ancestral sequence. Except for sites that are lost by deletions, each site in an ancestor is mappable into a single site in the descendent sequence. Correct alignment involves a transitive mapping, aligning the sites in one descendent to sites in the other if the sites are mapped onto the same site in the ancestor.

We shall not review here the literature discussing the process of searching possible alignments to find the one with the best score. Calculating a score for a pairwise alignment requires, however, a theory of evolution that describes how biomolecular sequences divergently evolve

naturally. The details of the theory often have an impact on the sequences that are inferred for ancestral proteins. It therefore becomes important conceptually to understand how inferences about ancestral sequences are made. Thus, it is useful to summarize the use of scoring functions.

Nearly all strategies for scoring a pairwise alignment incorporate some notion of minimum motion. Thus, identities, where two biomolecules have the same nucleotide or amino acid at an aligned site, are considered to be stronger signals of homology than are nonidentities.

Most theories also include parameters that permit some nonidentities to be stronger indicators of homology than others, however. For example, two aligned hydrophobic amino acids (e.g., a Leu and a Val) are presumed to be stronger indicators of homology than is a hydrophobic amino acid aligned with a hydrophilic amino acid (e.g., a Leu and a Glu). Scores that incorporate such additional parameters are said to measure the ''similarity'' between two sequences.

A 20×20 matrix (such as the matrix in Figure 2) is generally used to score identities and nonidentities in an alignment. The diagonal elements of such matrices hold the scores for identities in the matched sites; the off-diagonal elements hold the scores for nonidentities. The matrix itself is defined for a particular overall divergence of the pair of sequences being aligned. The more divergent the pair of proteins that the matrix is intended to score, the smaller the on-diagonal terms are relative to the off-diagonal

```
C 11.5
S  0.1  2.2
T -0.5  1.5  2.5
P -3.1  0.4  0.1  7.6
A  0.5  1.1  0.6  0.3  2.4
G -2.0  0.4 -1.1 -1.6  0.5  6.6
N -1.8  0.9  0.5 -0.9 -0.3  0.4  3.8
D -3.2  0.5  0.0 -0.7 -0.3  0.1  2.2  4.7
E -3.0  0.2 -0.1 -0.5  0.0 -0.8  0.9  2.7  3.6
Q -2.4  0.2  0.0 -0.2 -0.2 -1.0  0.7  0.9  1.7  2.7
H -1.3 -0.2 -0.3 -1.1 -0.8 -1.4  1.2  0.4  0.4  1.2  6.0
R -2.2 -0.2 -0.2 -0.9 -0.6 -1.0  0.3 -0.3  0.4  1.5  0.6  4.7
K -2.8  0.1  0.1 -0.6 -0.4 -1.1  0.8  0.5  1.2  1.5  0.6  2.7  3.2
M -0.9 -1.4 -0.6 -2.4 -0.7 -3.5 -2.2 -3.0 -2.0 -1.0 -1.3 -1.7 -1.4  4.3
I -1.1 -1.8 -0.6 -2.6 -0.8 -4.5 -2.8 -3.8 -2.7 -1.9 -2.2 -2.4 -2.1  2.5  4.0
L -1.5 -2.1 -1.3 -2.3 -1.2 -4.4 -3.0 -4.0 -2.8 -1.6 -1.9 -2.2 -2.1  2.8  2.8  4.0
V  0.0 -1.0  0.0 -1.8  0.1 -3.3 -2.2 -2.9 -1.9 -1.5 -2.0 -2.0 -1.7  1.6  3.1  1.8  3.4
F -0.8 -2.8 -2.2 -3.8 -2.3 -5.2 -3.1 -4.5 -3.9 -2.6 -0.1 -3.2 -3.3  1.6  1.0  2.0  0.1  7.0
Y -0.5 -1.9 -1.9 -3.1 -2.2 -4.0 -1.4 -2.8 -2.7 -1.7  2.2 -1.8 -2.1 -0.2 -0.7  0.0 -1.1  5.1  7.8
W -1.0 -3.3 -3.5 -5.0 -3.6 -4.0 -3.6 -5.2 -4.3 -2.7 -0.8 -1.6 -3.5 -1.0 -1.8 -0.7 -2.6  3.6  4.1 14.2
    C    S    T    P    A    G    N    D    E    Q    H    R    K    M    I    L    V    F    Y    W
```

FIGURE 2. A 20×20 matrix whose elements are 10 times the logarithm of the probability of the index amino acids being aligned by reason of common ancestry, divided by the probability of their being aligned by random chance, collected from aligned protein sequence pairs having about 0.15 amino acid replacement per site, and extrapolated to protein pairs having 2.5 amino acid replacements per site, based on the assumption that accumulation of accepted point mutations is a Markovian process, where subsequent point mutations are not influenced by predecessor point mutations.

terms. Because neither of the sequences in a pair being aligned is distinctive, a Leu matched against a Val is given the same score as a Val matched against a Leu. Thus, the alignment scoring matrix is symmetrical across the diagonal. A rate matrix, which describes the probability of each of the 20 amino acids of being replaced by each of the 20 amino acids, is different from a matrix that is used to score a pairwise alignment. The rate of replacement of one amino acid by another can be described as a pseudo-first-order process having the units of changes per site per unit time. Alternatively, the rate can be normalized to remove the time dimension. This provides a unitless rate parameter that describes the rate of replacement of one amino acid by another relative to the rate of replacement of all amino acids by all others. In either case, there is no reason for particular amino acid (e.g., Leu) to be replaced by another (e.g., Val) with the same rate constant as a Val is replaced by a Leu. If the rate constant for the conversion of Leu to Val is higher than the rate constant for the conversion of Val to Leu, the equilibrium ratio of Leu to Val will be higher in the descendent than in the ancestor.

2. Trees and Outgroups

Given a family of proteins containing more than three members, it is possible to summarize their interrelationship using an evolutionary tree. The tree is a graphical model of the history of a family of proteins. The leaves of the tree represent modern sequences from contemporary organisms. The internal threefold nodes in the tree each represent a duplication in an ancestral gene to give rise to two descendant lineages. The lengths of the edges in the graph represent the distance between nodes, often expressed in the units of changes per site. The longer the branch, the more the sequences at its ends differ. This is exemplified in Figure 3 using four ribonucleases (RNases) from four closely related bovids: the eland, the ox, the river buffalo, and the swamp buffalo.

The root of the tree is the point that represents the oldest sequence. If the rate of sequence divergence has been constant throughout the period of divergent evolution, the root is represented by a point at the middle of the tree. In general, however, the rates of protein sequence divergence are not constant over time, meaning that the root of the tree cannot be placed easily if sequence data are the only input.

Fortunately, it is possible to use other, nonsequence information to identify the root of the tree. For example, in the tree in Figure 3, the root is

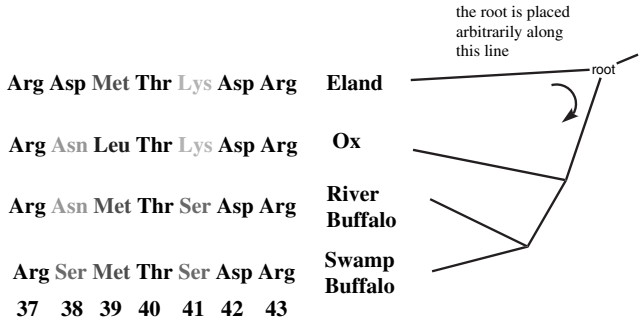


FIGURE 3. (Left) Part of the multiple sequence alignment relating the sequences of the ribonucleases from the eland, the ox, and the swamp and river buffaloes, covering sites 37 to 43. Note that the full sequence alignment has about 130 sites. Thus, the topology of the tree (right) is determined by the analysis of all of the sites, not just the seven sites shown here. This evolutionary tree (right) relates the sequences of the ribonucleases from the eland, the ox, and the swamp and river buffaloes. The eland sequence is, from paleontological and cladistic studies, chosen to be the outgroup. As such, it can provide a root for the tree containing just the sequences of the RNases from the ox, the swamp buffalo, and the river buffalo.

placed along the branch connecting the eland sequence to the point on the tree representing the protein sequence from the last common ancestor of the oxen and buffaloes. This placement is based on information from cladistics, which suggests that the eland diverged from the lineage leading to oxen and buffaloes before buffaloes diverged from the oxen. Thus, the eland is an *outgroup* for the tree that contains the oxen, the river buffalo, and the swamp buffalo sequences. An outgroup, however obtained, can be used to root the tree of the ''in group.''

3. Correlating the Molecular and Paleontological Records

Just as we might use cladistics to determine that the eland is the appropriate outgroup for a family of proteins obtained from the ox, the swamp buffalo, and the river buffalo, we might refer to the paleontological record to gain some insight into the animal that carried the ribonuclease that is represented by the threefold node in the tree connecting the branch to the ox ribonuclease and the branch to the RNases from the buffaloes. The paleontological record is incomplete. Therefore, one can never find in the paleontological record a fossil that truly corresponds to an individual, or even the population that held the individual, that generated two descendent

populations that eventually evolved to give two descendent species. Therefore, any assignment of a specific fossil to a specific point in an evolutionary tree can be only an approximation. Such assignments are often useful approximations, however, as they allow us to draw upon the paleontological and geological records to interpret the molecular record.

In the case of the tree modeling, the evolutionary history of the ribonucleases from eland, ox, swamp buffalo, and river buffalo, the fossil record contains several bovids that might correspond approximately to the last common ancestor of ox, swamp buffalo, and river buffalo. Stackhouse et al. (1990) chose *Pachyportax* (Figure 4) as the relevant genus from the

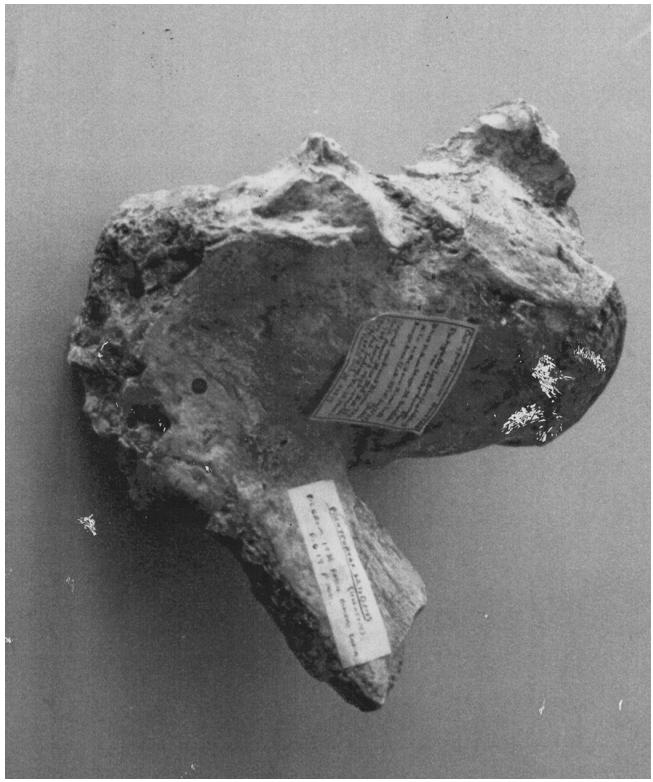


FIGURE 4. Fragmentary specimen of *Pachyportax latidens*, from (presently) Pakistan, representing part of the horn core and skull of a bovid that approximates the ancestor of the buffaloes and modern oxen. (From the British Museum of Natural History; photograph courtesy of Steven Benner.)

paleontological record to represent this node; this genus is known from a fossil record from the Indian subcontinent. The *Pachyportax* is represented on the tree using the script letter P (Figure 13).

B. HIERARCHY OF MODELS FOR MODELING ANCESTRAL PROTEIN SEQUENCES

1. Assuming That the Historical Reality Arose from the Minimum Number of Amino Acid Replacements

In a world without knowledge, we would know nothing about sequences of any ancestral protein that lived in any ancient organism that went extinct millions of years ago. In particular, we would know nothing about the sequence of the ribonuclease that was biosynthesized by *Pachyportax*.

But we have some knowledge. In particular, we know the sequences of some of the descendents of the ancient RNase as well as the sequences of descendents of its relatives. Further, from sequence data generally, we might derive a theory describing the replacement of amino acids generally during the divergent evolution of protein sequences. We might assume that the general pattern of amino acid sequence evolution might be a good approximation of the pattern in the protein family of interest. Axioms commonly incorporated into theories of protein sequence evolution are as follows:

1. Site *i* suffers replacement independently of site *j*.
2. Future replacements at site *i* are independent of past replacements.
3. The rate of replacement at each site is the same.
4. Patterns of replacement of amino acids at site *i* are the same as those patterns in the average site.

To understand how one can draw inferences about ancestral states from the sequences of descendents, let us consider a hierarchy of models that begins with just two homologous protein sequences, using ribonucleases as an example. We begin with the ribonucleases from the swamp and river buffaloes. These two homologs are represented by the leaves, or endpoints, of a simple tree consisting of a single line connecting the two endpoints. The points within the line represent evolutionary intermediates between the two sequences. Evolution is, of course, discrete, meaning that it is not represented perfectly by a continuous line with an infinite number of points, but the approximation is serviceable.

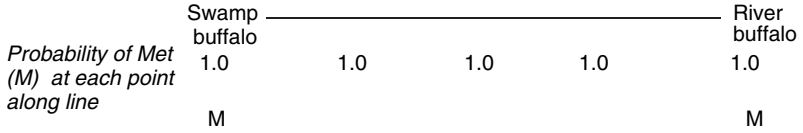


FIGURE 5. The amino acid at site 39 is methionine (Met, or M) in the RNase from both the swamp and the river buffalo. Therefore, a simple theory of evolution that holds that no change is substantially more likely than one change can account for the commonality of the amino acid residues by assuming that site 39 held a Met in all of the evolutionary intermediates between swamp and river buffalo. The probability that site 39 holds a Met is unity at every point along the line.

Consider now just two types of sites in the pairwise alignment of the two ribonucleases. In the first, the amino acids occupying the two aligned sites are identical. Site 39 is an example of such a site in these two ribonucleases. In the ribonuclease from both the swamp buffalo and the river buffalo, site 39 holds a methionine (abbreviated Met, or the single letter M). Figure 5 shows the simple tree with a M at each leaf, represented by the points at the end of the line. Patterns of replacement of amino acids at site *i* are the same as those patterns in the average site.

If our theory of evolution holds that the absence of change is substantially more likely than change (and this site does not contradict that theory), a simple model infers a Met at site 39 in all proteins that are evolutionary intermediates between the swamp and river buffalo RNases. This inference can be described using the language of probability. We simply say that the probability of finding a Met at site 39 in every ribonuclease represented by each point in the tree is unity (Figure 5).

The same is true in a rooted tree. As an approximation, we place the root midway between the two leaves of the tree, as in Figure 6. The oldest sequence is now at the top of the tree, with the direction of time being positive (toward the future) as one proceeds from the node at the top to the leaves at the bottom on the tree. The inference that is now drawn is that the ribonuclease found in the last common ancestor of swamp and river buffalo had a Met at position 39.

What if the amino acids occupying the two aligned sites are not identical in the two buffalo RNase sequences? This is the case for site 38 in the pairwise alignment (Figure 7). Here, the RNase from swamp buffalo holds a serine (Ser, or S), while the RNase from river buffalo holds an asparagine (Asn, or N). At this site, no change is not an option. If the history that we are to narrate

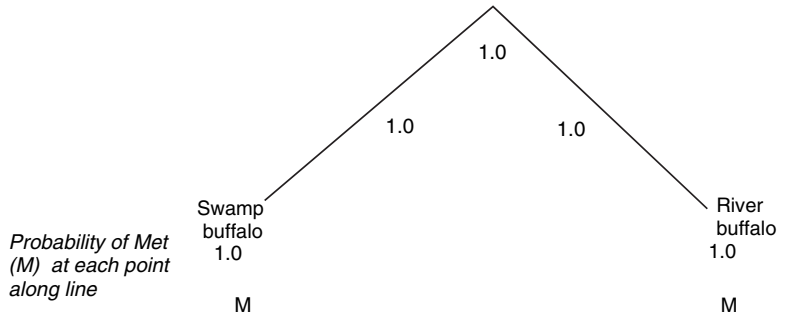


FIGURE 6. The inferences drawn from the RNAse sequences from the swamp and river buffalo does not differ if the tree is rooted. The probability of Met being at position 39 in the last common ancestor of the buffalos is still unity.

is to account for the presence of two different amino acids at this site in the contemporary sequences, at least one amino acid replacement must have occurred along the line connecting the two contemporary proteins.

If we assume that one replacement is substantially more likely than two, a simple model represents the amino acid at site 38 as a linear function of the distance along the line connecting the two leaves. Thus, an ancestral protein represented by a point on the line near the swamp buffalo leaf is more likely to have a Ser than an Asn at site 38, while an ancestral protein represented by a point on the line nearer the river buffalo leaf is more likely to have an Asn than an Ser. The amino acid from the swamp buffalo sequence ''morphs'' into the sequence from the river buffalo (Figure 7).

The same is true for the tree arbitrarily rooted at the midpoint of the line (Figure 8). The rooted tree illustrates the inference from the model that the last common ancestor of the RNases from swamp and river buffalo has a 50% chance of holding a Ser at position 38 and a 50% chance of holding an Asn at position 38.

	Swamp buffalo				River buffalo
Probability of Ser	1.0	0.75	0.5	0.25	0.0
Probability of Asn	0.0	0.25	0.5	0.75	1.0

FIGURE 7. The probabilistic values for the amino acid residue at site 38 of the RNase system morph smoothly along an evolutionary tree from 1.0 (for Ser) at the swamp buffalo leaf to 1.0 (for Asn, N) at the river buffalo leaf.

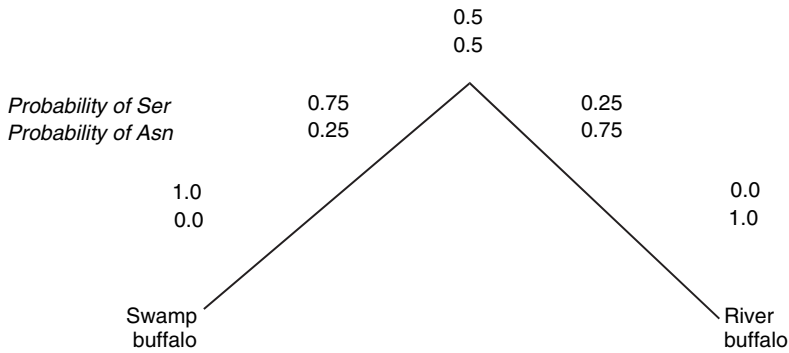


FIGURE 8. A rooted tree for site 38 shows probabilistic changes down each branch, with the change in time being positive as one proceeds from nods higher in the tree to nodes lower in the tree. Thus, in the branch leading from the last common ancestor to the modern swamp buffalo, there is 0.5 change of an ancestral Asn to a Ser. In the branch leading from the last common ancestor to the modern river buffalo, there is 0.5 change of an ancestral Ser to an Asn.

This simple analysis introduces two concepts that are key when using information from modern sequences to make inferences about ancestral sequences. The first is the concept of a *probabilistic ancestral sequence*. Every site in a sequence at every point along the branch(es) of the tree can be represented as a 20×1 probability matrix, where each element is the probability that each of the 20 amino acids was found at that site in the ancestral protein represented by the point. These probabilities sum to unity.

The example also illustrates the application of the notion of *parsimony*, or minimum change, when making inferences about ancestral sequences using information from modern sequences. Central to this model is that no amino acid replacement is more probable than one replacement, and that one replacement is more probable than two. Thus, if we can account for the contemporary sequences by a historical model that involves no changes, no changes are inferred. If we can account for the contemporary sequences by a historical model that involves one change, one change (not two changes) is inferred. These inferences are said to have been made using an evolutionary theory that is known as *maximum parsimony.*

2. Allowing the Possibility That the History Actually Had More Than the Minimum Number of Changes Required

More sophisticated theories describing the divergence of protein sequences allow for the possibility that more amino acid replacements have

occurred in the history of a site than the minimum required to account for the sequences derived. For example, we may incorporate into our theory the possibility that a site contained amino acids that are not found in *any* of the sequences derived, even though this requires some histories to have more than the minimum number of changes absolutely required to account for the amino acids in the sequences derived.

Consider again site 39 of the ribonucleases from the swamp and river buffaloes. Even though the sequences of both RNases hold a Met at this site, it is conceivable that another amino acid (let us say Ile) occupied site 39 in the last common ancestor. If this had been the case, at least two independent events in the history of the protein family would be required to account for the fact that Met occupies site 39 in both of the proteins derived. One Ile-to-Met replacement must have occurred at site 39 during evolution of the modern swamp buffalo sequence from the common ancestor. Another Ile-to-Met replacement must have occurred at site 39 during the evolution of the modern river buffalo sequence from the common ancestor. Because such a history requires two changes at site 39, and because the probability of a change is assumed to be a small fraction of the probability of no change, the likelihood that this ancestor has an Ile at site 39 is considerably less than the likelihood that the ancestor had a Met there. But this likelihood is not zero.

Further, we can build a theory that assigns a numerical likelihood based on this model. We might consider empirical data or theory to weight probabilities that various other amino acids intruded at site 39 during this history. For example, Lys is (on various grounds) chemically more dissimilar (compared to Met) than is Ile. Thus, the theory might give a lower probability to a Lys-to-Met replacement than an Ile-to-Met replacement. Accordingly, the chances of site 39 having suffered *two* Lys-to-Met replacements is considerably lower than the chances of site 39 having suffered two Ile-to-Met replacements. This means that the chance of the ancestral sequence having had a Lys is lower than the chance of it having had an Ile, given that the two derived sequences both have Met at site 39. Here, a probabilistic ancestor might have nonzero values for all amino acids at site 39. For example, here are the occupancies for site 39 for the 20 amino acids calculated at the midpoint of the tree (Figure 9).

This is illustrated graphically in Figure 10. Here, the probability of Met at site 39 does not remain unity, even though both leaves have Met at this site. Rather, the probability of Met at site 39 drops linearly until the midpoint of the line and then rises back to unity as the other leaf is approached. At the same time, the probability of all of the other amino

A	C	D	E	F	G
0.006	0.001	0.002	0.002	0.011	0.006

H	I	K	L	M	N	P
0.004	0.015	0.004	0.012	0.894	0.004	0.003

Q	R	S	T	V	W	Y
0.004	0.002	0.003	0.004	0.011	0.002	0.010

FIGURE 9. Probabilistic model for site 39 in the last common ancestor of swamp and river buffalo. Note how the values sum to unity. Note also that the likelihood of different amino acids reflects the alignment probabilities of the scoring matrix in Figure 2.

acids being present at site 39 increases from zero at the leaves to a maximum number at the midpoint and then drops back to zero as the other leaf is approached.

This analysis is sometimes called *maximum likelihood*, because it integrates features, including the distance of the node from the leaves (where the sequence is perfectly defined), into a probabilistic model. Thus, the probability of two changes occurring along a branch of a tree is lower if the line is shorter (i.e., if the distance between the ends of the branch is shorter) and higher if the branch is longer. This follows directly from the fact that the length of the branch is the number of changes per site. For the tree holding two leaves, the ML formalism says that the certainty that a residue found in common at both leaves of a tree was present throughout the history of the site diminishes as the branch length becomes longer.

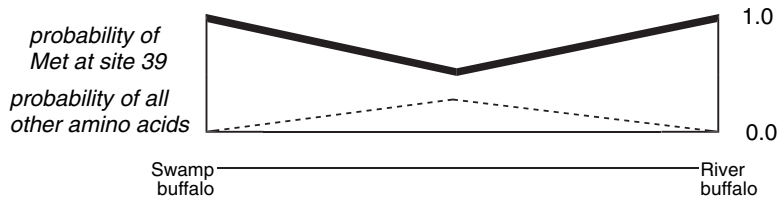


FIGURE 10. In a maximum likelihood model, the probability of finding a Met at site 39, which is unity at the ends of the line (as we know that Met occupies site 39 in the sequences from modern swamp and river buffalo), decreases toward the middle, with the probability of all other amino acids being present at that site increasing from zero to a maximum at the midpoint of the tree.

3. Adding a Third Sequence

Adding a third sequence to the original tree creates a trifurcated tree. The third sequence is tied into the line separating the first two sequences by a branch at the point where the probabilistic ancestral sequence along the line resembles the third sequence the most. In the example we use the oxen sequence as the third.

We might say that the sequence of the RNase from oxen serves as an outgroup for the two RNase sequences from swamp and river buffalo. Alternatively, if we assume that the ox diverged from the lineage leading to the two buffalos before the buffalos themselves diverged, we may say that the ox sequence roots the swamp buffalo–river buffalo tree. The oldest point on the swamp buffalo–river buffalo tree, given this additional information, is the point where the oxen sequence is tied to that tree. The root of the trifurcation is not known but lies somewhere along the line from the central threefold node and the ox sequence.

The third sequence contributes more prior information that can be used to infer the amino acids occupying various sites at points internal to the tree. Consider again site 38 (Figure 11), where the RNase from swamp buffalo holds a Ser, while the RNase from river buffalo holds an Asn. With just two sequences, the ancestral sequences morphed smoothly from Ser to Asn from left to right along the line. With the ox sequence, the available

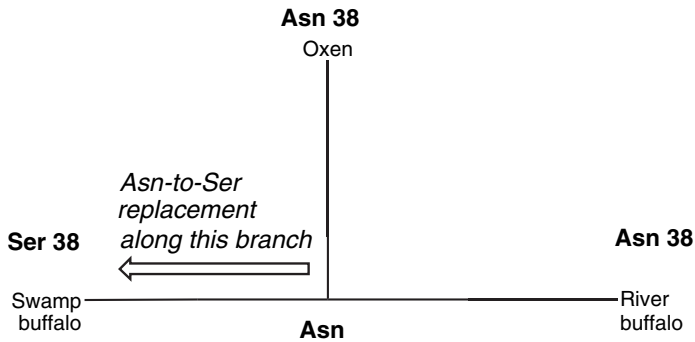
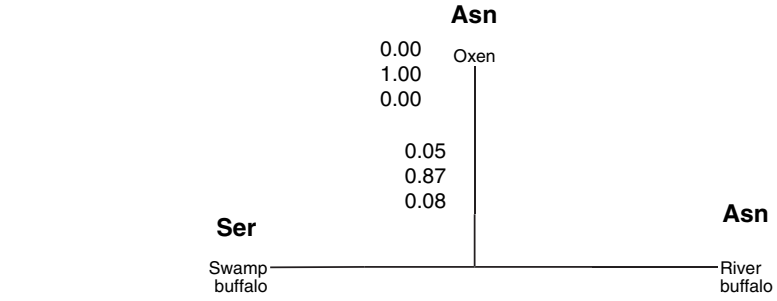


FIGURE 11. A parsimony analysis infers an Asn at site 38, as this inference permits the occupancy of site 38 in all of the proteins to be accounted for by a single change. Any other inference for site 38 would require our evolutionary model to assume more changes than the minimum absolutely required to account for the amino acids observed at site 38.



Probability of Ser	1.0	0.48	0.10	0.05	0.0
Probability of Asn	0.0	0.48	0.84	0.87	1.0
Probability of others	0.0	0.04	0.06	0.08	0.0

FIGURE 12. A maximum likelihood analysis infers an Asn at site 38 as the most probable residue. Note how the additional "prior" information alters the probabilities assigned to different amino acids at points throughout the tree. Note that the sum of the probabilities of Ser, Asn, and all other amino acids is unity at all points in the tree.

information has increased. Now, a simple parsimony analysis assigns an Asn at the central node, with the model for the history of site 38 incorporating an Asn-to-Ser replacement along the branch leading from the central node to the RNase from swamp buffalo. The probabilities along the line are also changed dramatically by the introduction of the third sequence. This is illustrated in Figure 12.

It requires no conceptual leap to extend this analysis to all of the sites in the threefold alignment of the ox, swamp buffalo, and river buffalo RNase sequences, or to trees that include the sequence of the RNase from eland, added to the tree as an outgroup. The eland sequence roots the (ox, swamp buffalo, river buffalo) tree and allows us to infer the maximum parsimony sequence for the RNase from *Pachyportax*. This sequence was the first for an ancestral biomolecule to have been resurrected experimentally (Stackhouse et al., 1990).

As an exercise, it might be useful to see how a parsimony analysis infers the amino acid replacements that occur along each branch of the tree. For example, in the time during which the RNase from ox evolved from the RNase in *Pachyportax*, the methionine at site 39 was replaced by a leucine. In the time during which the RNase from the last common ancestor of the swamp and river buffalos evolved from the RNase in *Pachyportax*, the Lys at site 41 was replaced by a Ser. In the time during which the RNase from

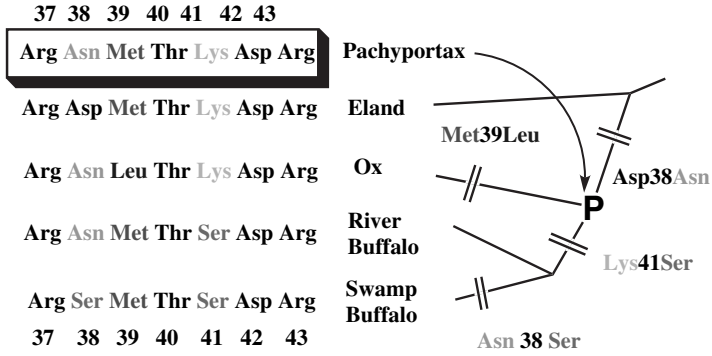


FIGURE 13. The ancestral residues inferred for specific sites at internal nodes in the tree also assign changes along individual branches in the tree.

swamp buffalo evolved from the last common ancestor of the swamp and river buffaloes, the Asn at site 38 was replaced by a Ser (Figure 13).

Because the maximum likelihood analysis infers fractional residues at nodes, it also assigns fractional changes along branches. These are not different conceptually from integral changes.

4. Relative Merits of Maximum Likelihood Versus Maximum Parsimony Methods for Inferring Ancestral Sequences

In much of the literature on experimental paleogenetics, it is argued that the maximum likelihood methods are preferred to infer ancestral sequences over maximum parsimony methods (Yang et al., 1995; Zhang and Nei, 1997; Pagel, 1999a; Nielsen, 2002). Although this argument is true, especially to the extent that maximum likelihood methods capture more of what we know about protein sequence evolution, it is important to realize that the impact of replacing a maximum parsimony analysis by a maximum likelihood analysis is small when the tree is highly articulated, the branching topology is secure, and the overall extent of sequence divergence is small. This is the case for the ribonucleases that we have just discussed.

Nevertheless, it is worth using this example to illustrate how the nature of an inferred ancestral sequence might change if elements of the evolutionary model were to change. In particular, the ancestral sequences inferred via a parsimony analysis are often extremely sensitive to changes in the topology of the tree. Consider, for example, how the sequence

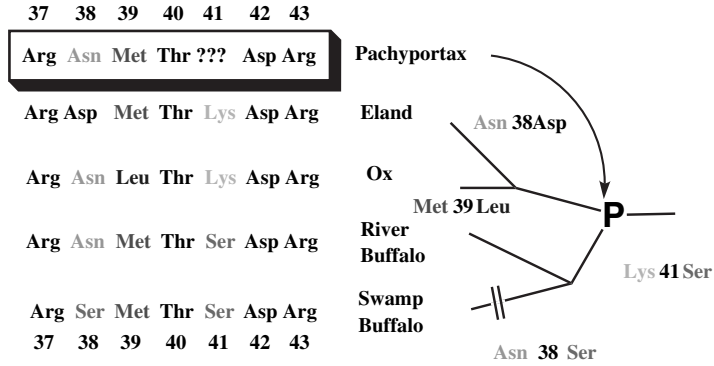


FIGURE 14. With a different tree, the amino acid inferred at site 41 changes in a parsimony analysis.

inferred for the ancestor P would change if the position of the eland on the tree were moved so that it was not an outgroup, but rather, was a sister group of the oxen. This change is illustrated in Figure 14. Here, the amino acid at site 41 is no longer defined by a parsimony analysis. Site 41 could with equal probability hold a Lys or a Ser.

Because many trees inferred from sequence data are not known with certainty, where much of the ambiguity involves swapping around short branches, maximum likelihood inferences have significant advantages over maximum parsimony inferences. Here, maximum likelihood considers the lengths of branches when assigning probabilities for different amino acids being present. Changes in tree topology around short branches therefore do not change greatly the probabilities assigned by maximum likelihood tools.

At the same time, because an experimental paleogeneticist cannot construct a protein having fractional occupancies of single sites, the uncertainty in the sequence represented by the probabilities of "all the other" amino acids is generally ignored if the probability assigned to those other amino acids falls below a threshold. There is no consensus in the community for where that threshold should be. It is clear, however, that if the probability of site 39 holding a Met is greater than 0.81 while the probability of any of the other amino acids at that site is 0.01, one can make a convincing (to most) argument that the only ancestral protein that needs to be resurrected is the one holding a Met at site 39.

C. COMPUTATIONAL METHODS

A variety of computer programs are now available to construct multiple sequence alignments, to build trees, and to infer ancestral sequences. Clustal W is widely used for the first, although nearly all practitioners adjust the multiple sequence alignments that Clustal W produces by hand. Further, most practitioners adjust the trees that are generated by automated computer tools by hand, with the goal of having the tree topology conform to the topology expected based on information independent of the sequences in the family of interest.

The identification of the optimal (best scoring and, we hope, corresponding best to the historical reality) of multiple sequence alignments and evolutionary trees is computationally expensive. The number of possible alternative trees scales severely with increasing numbers of sequences. A large literature, not reviewed here, discusses heuristics that generate trees with many leaves more efficiently (Huelsenbeck et al., 2001).

Once a multiple sequence alignment and a tree are in hand, several computer programs are available to infer ancestral sequences. Parsimony methods are implemented, for example, in programs such as MacClade and PAUP* (Maddison and Maddison, 1989; Swofford et al., 1996; Swofford, 2001). MacClade allows a user to alter the tree manually, interrelating the input sequences to find the most parsimonious tree. This is a powerful tool, as the interaction between the experimentalist and his or her intuition with the computational support provided by the program allows the user to ask ''what if?'' and ''why not?'' questions with ease.

PAUP* does not have this interactive feature. It has the advantage, however, of calculating a maximum likelihood tree as well as the most parsimonious tree. Maximum likelihood analysis for the inference of ancestral sequences is implemented in programs such as Darwin (Gonnet and Benner, 1991), PHYLIP (Felsenstein, 1989), MOLPHY (Adachi and Hasegawa, 1996), PAML (Yang, 1997), and NHML (Galtier and Guoy, 1998). These all make accessible one or more formal models for evolution and use a likelihood score as an optimality criterion (Felsenstein, 1981). Optimization of the likelihood score can be used to specify topology and parameters such as branch lengths, character state frequencies, and ancestral states (Zhang and Nei, 1997; Pupko et al., 2000; Cai et al., 2004; Thornton, 2004).

In the examples discussed below, these and other methods are used to infer the sequences of ancient proteins that are to be resurrected. We

comment on the methods used throughout the discussion, to allow the reader to better understand the uncertainties that the methods generated and the ambiguity of the resurrection achieved.

D. HOW NOT TO DRAW INFERENCES ABOUT ANCESTRAL STATES

Each of the methods described above exploits the information in sequence according to the position of the sequence within a tree, which models the familial relationships of the protein sequences. The tree weights the sequences so that two nearly identical sequences do not contribute twice as much as one distant sequence to the inferences of the ancestral sequences.

For this reason, consensus tools are not preferred as ways to infer ancestral character states. Consensus tools allow each member of the family to "vote," and build a consensus sequence by "majority rule." Thus, if six proteins hold a Met at site 39, and 18 hold a Leu at site 39, the consensus holds a Leu at site 39. This approach makes no sense if all of the 18 proteins holding Leu are from the same breed of ox, and the six holding Met are from buffalo, eland, deer, sheep, impala, and camel.

Several groups, especially in the earliest literature in experimental paleoscience, used a consensus sequence to approximate an ancestral sequence. We include these in this review despite this defect. Even if a tree is not known precisely, an approximate tree can better weight different sequences to give better inferred ancestral sequences than can a consensus tool.

III. AMBIGUITY IN THE HISTORICAL MODELS

A. SOURCES OF AMBIGUITY IN THE RECONSTRUCTIONS

As with any inference about the past, ancestral sequences are not inferred with absolute certainty. Mistakes in the sequence database, failure of approximations built into the evolutionary theory, uncertainty in the multiple sequence alignment, and uncertainty in the tree topology all contribute to the ambiguity, even though these contributions are not captured in the formalism of a maximum likelihood analysis based on a single multiple sequence analysis and a single tree.

In addition to ambiguity in the database, theory, alignment, and tree, a measure of the uncertainty of the inferences is formally offered by a maximum likelihood analysis. The extent of this ambiguity is captured in the fractional probabilities of each of the amino acids inferred for each site.

While different maximum likelihood (ML) tools give different levels of ambiguity in any given case (and sometimes infer different preferred amino acids at a site), it is difficult to know which ML tool is most likely to capture the historical reality for any particular protein family. Methods for performing statistical tests on models, such as the likelihood ratio test and Bayes factors, may be used to identify the best model to fit the data (Huelsenbeck et al., 2004).

Such tests do not, of course, capture the ambiguity that arises from defects in the underlying data or approximations in the theory. The principal source of formal ambiguity in the inference by ML tools of the amino acid at an ancestral site arises simply because the history of a site contains too many replacements relative to the degree of articulation of the tree. Further ambiguity arises if that history includes cases of homoplasy (e.g., parallel or convergent evolution), where the same replacement occurs on different branches of a tree with a higher-than-random frequency. Such events are not captured within the formalism of the theory. These issues are illustrated by two sites, 168 and 211, from the alcohol dehydrogenases from various taxa of yeast (Figure 15). In this example, the topology of the tree was relatively secure, as were the sequences at the leaves, which were checked in many ways (Thomson et al., 2005). The sites had suffered too many replacements, however, for any method (parsimony or ML) to strongly infer the presence of one amino acid at a key node in the tree (the node at the right end of the red branch). The corresponding ML analysis gave posterior probabilities substantially less than 0.5 for all 20 amino acids at that node, with no individual amino acid having a clear preference.

Sites such as these are often discussed when different maximum likelihood tools are compared for their ability to infer ancestral sequences. ML tools generate significantly different inferences at such sites than those of parsimony tools. Different ML tools often generate considerably different probabilistic ancestral sequences (e.g., DNA-, codon-, or amino acid–based models). Extensive discussion is emerging as to which tool makes ‘‘better’’ inferences. In such discussions it is important to recognize that the systematic errors introduced by incomplete evolutionary theories, including those arising from the assumptions that amino acid replacements at individual sites reflect replacements at the average site, might be more significant than the formal uncertainty expressed by ML probabilities.

The experience in experimental paleogenetics over the past 20 years suggests that even with abundant data and highly articulated trees,

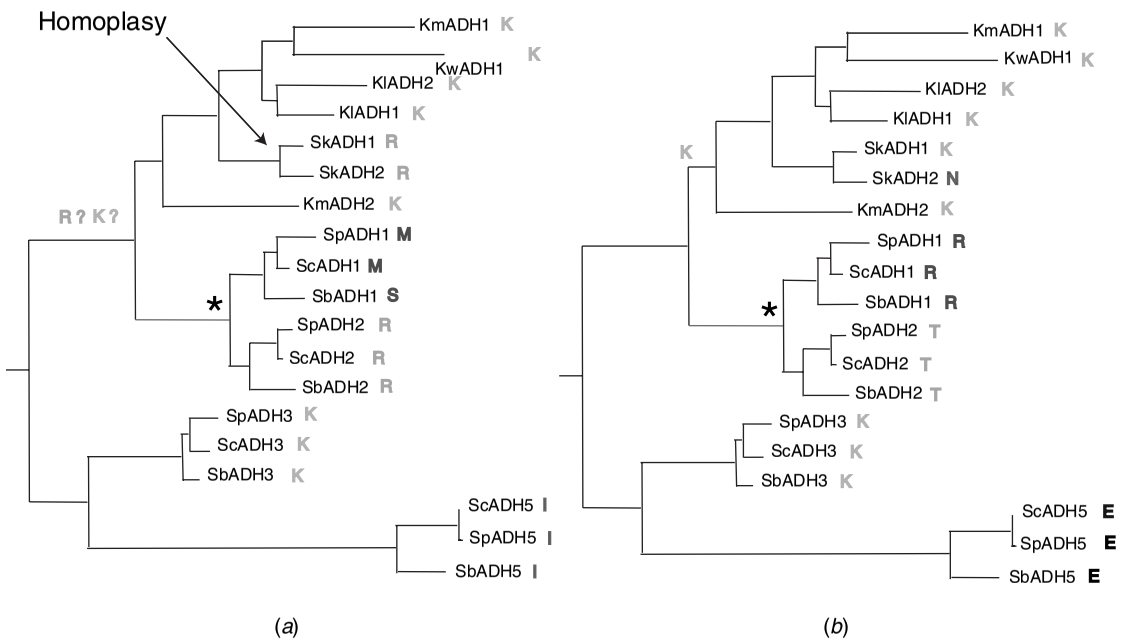


FIGURE 15. Distribution of amino acids at site 168 (*a*) and site 211 (*b*) in a set of 19 fungal alcohol dehydrogenases. The node of interest is at the right end of the red branch, marked with an asterisk. Note the difficulty in reconstructing the amino acid at these sites at the node at the right end of the red branch.

ambiguity will remain. In those cases where it remains, the experimental paleogeneticist must consider how to manage any ambiguity in an ancestral sequence that remains.

B. MANAGING AMBIGUITY

Four ways are commonly used to manage the ambiguity in a set of input sequences. The first relies on statistical models of sequence evolution wherein ''optimized'' parameters are used as input to estimate ancestral character states. Here, inferred character states are often ''resolved'' from ambiguity using the branch lengths interconnecting nodes of the phylogenetic tree. The second involves collecting more sequences in the hope of eliminating the ambiguity. The third ignores the ambiguity based on an argument that the ambiguity occurs only at sites that are not critical for the biological interpretation. The fourth involves synthesizing and studying many candidate ancestral sequences to cover all plausible alternative reconstructions or to sample among the plausible alternative reconstructions.

1. Hierarchical Models of Inference

Frequent ambiguity in the estimation of ancestral character states is a weakness of the parsimony approach (*vide supra*) and can be rampant even when the relationships among the sequences are known. Individual efforts to resolve this weakness have attempted to accommodate branch length information and empirical frequencies associated with amino acid replacements to infer ancestral states (Schluter, 1995; Yang et al., 1995; Koshi and Goldstein, 1996). A concerted effort to unite tools from various laboratories was made at a symposium entitled ''Reconstructing Ancestral Character States'' in 1999, which focused on phenotypic and ecological characters (Cunningham, 1999; Martins, 1999; Mooers and Schluter, 1999; Omland, 1999; Pagel, 1999b; Ree and Donoghue, 1999; Schultz and Churchill, 1999).

This was followed up six years later by the inaugural conference on ''Ancestral Sequence Reconstruction'' (www.cbu.uib.no/asr), which focused on ancestral nucleic and amino acid states. A common discussion presented at both symposia addressed uncertainties in evolutionary models and their impact on inferred ancestral states (Schultz et al., 1996; Schultz and Churchill, 1999; Huelsenbeck and Bollback, 2001; Krishnan et al., 2004).

Central in these discussions are *hierarchical bayesian* approaches. A Bayesian analysis begins by constructing a formal relationship between an unknown, a data set, and one or more inference rules. Adding inference rules can create a hierarchy of analyses. These attempt to accommodate uncertainty in optimized parameters such as tree topology, branch lengths, and rate heterogeneity, *inter alia*, by adding stepwise sophistication to the models (Huelsenbeck and Bollback, 2001; Pagel et al., 2004).

The phylogenetic accuracy, consistency, and congruence of these models remain to be determined. We expect that the more uncertainties in the estimates of the parameters accepted by a model, the more the ambiguous the ancestral states will appear. Thus, this approach does not solve the key problem in experimental resurrections: One needs to actually resurrect an ancestral sequence, not be paralyzed by the perception of ambiguity.

The challenge will require a balance between acceptable uncertainty and acceptable parameter estimation. To this end, researchers at the Foundation for Applied Molecular Evolution and the University of Florida are using computer simulations to evaluate the performance of hierarchical Bayesian approaches. Nevertheless, for practicing experimental paleogenetics, it is clear that purely mathematical tools may not resolve issues of ambiguity.

2. *Collecting More Sequences*

The least controversial way to manage ambiguity in an ancient reconstruction is simply to collect more sequences. If chosen strategically, additional sequences can articulate a tree in a way that resolves ambiguity in a parsimony analysis or alters the posterior probabilities in a maximum likelihood analysis to increase confidence in the inference.

Depending on the taxa involved, collecting more sequences might be a simple task, requiring only that the scientist obtain biological tissue specimens of organisms that branch from the tree at the positions where further articulation of the tree might resolve ambiguities. Extinctions of lineages obviously can prevent this from being done, of course. To the extent that extinctions have removed information from the biosphere, it may be impossible to find an extant organism that branches from the tree at a strategic point useful to resolve an ambiguity in ancestral reconstruction.

Of course, it is always possible that collecting additional sequences may not resolve ambiguities. Indeed, additional sequences might create new

ambiguities, especially when long branches are being articulated. This is not bad, as it means that additional sequences have discovered ambiguities that exist but were not revealed by the previous, smaller data set. Although it is obvious, a general rule is worth stating: The more sequences in the data set, and the more broadly the relevant history is sampled, the more reliable the reconstructed ancestral sequences will be.

The biodiversity represented in the microbial world is only beginning to be explored, of course. This suggests that sequencing of the type that Venter performed will add data and will make paleomolecular reconstructions increasingly reliable over the coming decades. The same is true for many metazoan phyla. The diversity of genetic information within, for example, the beetles, is largely unexplored, meaning that we do not know how far back in time we will be able to resurrect proteins from extinct arthropods. Here, the mass extinction at the Permian–Triassic boundary appears to be the first problematic event in Earth history that might have removed sufficient genetic information to cause problems.

Unfortunately, this is not the case with mammals. Reconstructions in mammals can be well supported over the past 100 Ma, given the number of radiant mammal orders that continue to leave descendents in the modern world. Even here, the loss of large eutheria remains a problem. Those interested in, for example, the molecular paleontology of the dawn horse will be disappointed in the number of descendent lineages that survive. Here, the experimental paleomolecular biologist will probably be constrained forever by the history of the terrean biosphere.

3. Selecting Sites Considered to Be Important and Ignoring Ambiguity Elsewhere

If ambiguity cannot be resolved by additional sequencing, we might simply ignore the ambiguity at some sites by focusing on just a few where specific amino acids are believed to be critical to biological function, and where ambiguity is not observed. The strategy then involves ignoring the remaining ambiguity, in the hope that it does not influence the behavior of the protein that is the object of biological interpretation. A large number of the examples discussed below manage ambiguity in this way.

This strategy is controversial, and for good reasons. The behavior of a protein is generally not a linear function of the amino acids in its sequence. It is impossible, therefore, to say with certainty which sites are critical (indeed, that is often why the paleomolecular experiment is being done).

Thus, examples are known from protein engineering where the impact of an amino acid replacement at site *i* is different, depending on the amino acid occupying site *j*. This means that an amino acid replacement may have an impact on the behavior of a protein in some contexts that is different from its impact in others. Further, many examples are now known from protein engineering where the behavior of a protein at its active site is influenced by an amino acid replacement far from the active site.

Thus, ignoring ambiguity is recommended only if there is no alternative or if a relatively comprehensive examination of the sites in single mutagenesis experiments makes compelling the argument that ambiguity at these sites can be ignored. Confidence in this approach could be improved if the protein being examined has multiple domains that have been shown experimentally to function independently and the reconstruction ambiguity is not in the domain of interest.

4. Synthesizing Multiple Candidate Ancestral Proteins That Cover, or Sample, the Ambiguity

A relatively noncontroversial way to managing ambiguity involves the synthesis of all of the candidate ancestral sequences that are plausible given the model. Thus, if the ancestor has one site that is ambiguous, and its ambiguity arises from the failure of the analysis to choose decisively one of two amino acids, both sequences can be resurrected as candidate ancestral proteins. If the behavior to be interpreted is the same in both candidates, the ambiguity has no impact on the biological interpretation. The biological interpretation is said to be *robust* with respect to the ambiguity.

If the number of sites holding ambiguities is large, this strategy may require the synthesis of many candidate ancestral sequences. For example, if 10 sites are ambiguous with respect to a choice between two amino acids, a total of 1024 ($= 2^{10}$) different candidate ancestral sequences must be synthesized to cover all combinations of amino acids at the various sites. This can, of course, strain a laboratory budget.

An alternative is to sample among the candidate ancestral sequences. Here, a library is constructed that contains the candidate ancestral sequences, and a sample of these is studied. The library can be biased to reflect the fractional posterior probabilities in the ancestral sequence inferred, so that the sampling captures the Bayesian features of the analysis. If the behavior of all of the candidate ancestral sequences that are sampled is the same with respect to the phenotype that supports the biological

interpretation, it is possible to argue that the interpretation is robust with respect to the ambiguity. The extent to which the argument is persuasive will depend on the size of the sample, the extent of the ambiguity, and the taste of the scientist.

C. EXTENT TO WHICH AMBIGUITY DEFEATS THE PALEOGENETIC PARADIGM

If the hypersurface relating protein behavior to protein sequence were extremely rugged, and if every amino acid replacement caused a significant change in behavior, ambiguity would defeat the paleogenetic research approach in all but the most ideal cases. Fortunately, biochemical reality is different. For nearly all proteins, some amino acid replacements at some sites have a large impact on functional behaviors, replacements at other sites have a modest impact on those behaviors, and replacements at still other sites have even less impact on most behaviors.

This fact tends to ameliorate the extent to which ambiguity compromises work in experimental paleoscience. Ambiguity generally is found at sites that have suffered the most amino acid replacements. Multiple amino acid replacements often (but not always) reflect *neutral drift* at a site. Neutral drift implies that the choice of a residue at the site does not have a significant impact on fitness. This generally (but not always) means that replacement of an amino acid at that site does not have any impact on the behavior of a protein that can be detected by an *in vitro* experiment.

Stringing this logic together, we can expect (but not always) that biologically interpretable behavior will not differ greatly between ancestral sequences that differ only at ambiguous sites. To the extent that the premises are true, ambiguity in general will not limit our ability to draw inferences about the behavior of ancestral proteins by experimental analysis of ancestral sequences, even if our analysis does not capture all of the ambiguity in those sequences. This, in turn, means that we will generally be able to use those behaviors to generate interesting biological interpretations. In fact, this is the case, as is illustrated by approximately 20 examples of experimental paleogenetics to emerge over the past two decades.

IV. EXAMPLES

Further discussion of the details of evolutionary theories and tools to build models for the history of protein sequences can be found in the

literature. Below, we review the examples of experimental paleogenetics where these theories and tools have been applied. We present these in approximately the order in which they appeared in the literature. The presentation deviates from this order when it makes logical sense to group a series of studies together.

A. RIBONUCLEASES FROM MAMMALS: FROM ECOLOGY TO MEDICINE

The family of proteins related to bovine pancreatic ribonuclease A (RNase A) provided the first biomolecular system to be analyzed using experimental paleobiochemistry. These studies extended the long history during which RNase contributed to the development of tools in the biomolecular sciences. RNase was also the first protein to be observed by nuclear magnetic resonance methods (Saunders et al., 1957), one of the first to be reconstituted from its parts (Richards and Logue, 1962), one of the first to be analyzed by protein sequencing (Moore and Stein, 1973), the very first protein to be synthesized (Denkewalter et al., 1969; Hirschmann et al., 1969; Jenkins et al., 1969; Strachan et al., 1969; Veber et al., 1969), and the first enzyme for which a synthetic gene was prepared (Nambiar et al., 1984).

Members of the RNase family of proteins are typically composed of a signal peptide of about 25 amino acids and a mature peptide of about 130 amino acids. Most members of the RNase family have three catalytic residues (one lysine and two histidines, at positions 41, 12, and 119 in RNase A). These come together in the folded enzyme to form an active site. In addition, RNases generally have six or eight cysteines that form three or four disulfide bonds. Except for these conserved residues, the sequences of RNases have diverged substantially in vertebrates, with sequence identities as low as 20% when comparing oxen and frog homologs (for example).

Before paleobiochemical experiments began, RNase was known simply as a digestive enzyme. In the subsequent 20 years, developments on many fronts have shown that the digestive function in the RNase family is a relatively recent innovation and is important in only a few mammal orders (Benner, 1988). Behaviors ranging from immunosuppressivity and antitumor activity to duplex DNA binding and antiviral activity are now known in RNase family members. Today, paleobiochemistry is arguably the most important tool being used to sort out the rich functional diversity in this family of protein; it is unlikely that this functional understanding of this family could have been so effectively developed without paleobiochemical studies.

1. Resurrecting Ancestral Ribonucleases from Artiodactyls

In the early 1980s, the RNase A family was a practical choice to begin paleobiochemical work. Jaap Beintema and his co-workers had spent many years sequencing RNase homologs isolated from the pancreases of a variety of mammals (Beintema and Gruber, 1967, 1973; Gaastra et al., 1974, 1978; Groen et al., 1975; Welling et al., 1975, 1976; Emmens et al., 1976; Kuper and Beintema, 1976; Muskiet et al., 1976; Vandenberg et al., 1976; Vandijk et al., 1976; Beintema et al., 1979, 1984, 1985; Jekel et al., 1979; Lenstra and Beintema, 1979; Beintema and Martena, 1982; Breukelman et al., 2001). Done before the ''age of the genome,'' this work exploited classical Edman degradation of peptide fragments derived by selective cleavage of the protein. Such work required substantial amounts of protein, making convenient the large amount of RNase found in the digestive tracts of oxen and their immediate relatives.

As expected for enzymes found in the digestive tract, RNases were themselves robust. For example, the first step in the purification of RNase A involved the treatment of an extract from ox pancreas with 0.25 *M* sulfuric acid. This procedure precipitates most other proteins and removes the glycosyl groups from RNase, but otherwise leaves the protein intact. Thus, by 1980, about 50 different RNase sequences were known.

At that time, few other protein families were so well represented in the protein sequence database. Other families that had been well sequenced included cytochrome C, which had been developed as a paradigm for molecular evolution by Margoliash (1963, 1964), and hemoglobin, which was studied as a model for biomolecular adaptation (Riggs, 1959; Bonaventura et al., 1974).

Each of these alternative families was problematic as a system for developing paleobiochemistry as a field. The cytochromes are themselves substrates for other proteins, the cytochrome C oxidases. This suggested that studies on ancestral cytochromes would need to involve resurrected ancestral oxidases. As no U.S. federal agency was willing to fund paleobiochemistry in the 1980s, resurrecting ancestral sequences in one family was likely to be difficult; resurrecting two sets of ancestors from two families was considered to be impossible.

Hemoglobins remain a promising family for paleomolecular resurrections. To date, however, only one laboratory has explored them for this purpose (Benner and Schreiber, unpublished).

RNases proved to present several opportunities for biological interpretation and discovery. As digestive enzymes, pancreatic RNases lie at one interface between their host organisms and their changing environments, and are expected to evolve with the environment. Not all mammals, however, have large amounts of pancreatic RNase. In fact, RNase is abundant in the digestive systems primarily in ruminants (which include the oxen, antelopes, and other bovids, together with the sheep, the deer, the giraffe, okapi, and pronghorn) and certain other special groups of other herbivores (Barnard, 1969).

In 1969, Barnard proposed that pancreatic RNase was abundant primarily in ruminants because ruminant digestion created a special need for an enzyme that digested RNA. Ruminant digestive physiology is considerably different from human digestive physiology (for example). The ruminant foregut serves as a vat to hold fermenting microorganisms. The ox delivers fodder to these microorganisms, which produce digestive enzymes (including cellulases) that the ox cannot. The microorganisms digest the grass, converting its carbon into a variety of products, including low-molecular-weight fatty acids. The fatty acids then enter the circulation system of the ruminant, providing energy.

The ox then eats the microorganisms for further nourishment. According to the Barnard hypothesis, this digestive physiology creates a need for especially large amounts of intestinal RNase to digest microorganisms. The fermenting microorganisms are packed with ribosomes and ribosomal RNA, transfer RNA, and messenger RNA. Fermenting bacteria therefore deliver large amounts of RNA to the gastric region of the bovine stomach and the small intestine. Barnard estimated that between 10 and 20% of the nitrogen in the diet of a typical bovid enters the lower digestive tract in the form of RNA.

Barnard's hypothesis was certainly consistent with the high level of digestive enzymes in the ruminant generally. For example, ruminants have large amounts of lysozyme active against bacterial cell walls in their digestive tracts.

Was the Barnard hypothesis merely a "just so" story, based on correlations that did not require causality or functional necessity? The first experimental paleobiochemistry program set out to test this.

As discussed above, the available sequences were adequate to support the inference, with little ambiguity, of the sequence of the RNase represented (approximately) by the fossil ruminant *Pachyportax* (Figure 4) (Stackhouse et al., 1990). This was also the case for the more ancient

FIGURE 16. Fossils of *Eotragus*. (From Musée d'Histoire Naturelle; photograph courtesy of Steven Benner.)

Eotragus, which lived in the Miocene (Figure 16). The available RNase sequences also permitted the inference, with only modest ambiguity, of sequences for RNases in the first ruminant, approximated in the fossil record by the genus *Leptomeryx* (Figure 17). With slightly more ambiguity, the contemporary RNase sequences allowed the inference of the sequences of RNase in the first artiodactyl, the order of mammals having cloven

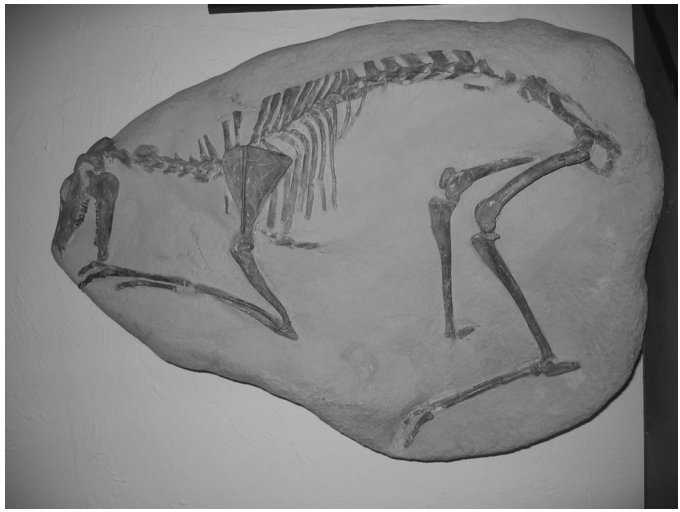


FIGURE 17. Fossils of *Leptomeryx*, a primitive ruminant from the North American Oligocene. (From the S. A. Benner collection; photograph courtesy of Steven Benner.)

hoofs, which includes the true ruminants as well as the camels, the pigs, and the hippos. This ancestor is approximately represented in the fossil record by the genus *Diacodexis.* A collaboration between the Center for Reproduction of Endangered Species at the San Diego Zoo and the Benner laboratory yielded several additional sequences that assisted in these inferences (Trabesinger-Ruef et al., 1996).

Once the ancestral sequences were reconstructed, the Benner group prepared by total synthesis a gene for RNase that was specially designed to support the resurrection of ancient proteins (Nambiar et al., 1984; Stackhouse et al., 1990). From this gene, approximately two dozen candidate ancestral genes for intermediates in the evolution of artiodactyls ribonucleases were synthesized, cloned, and expressed to resurrect the ancestral proteins for laboratory study (Stackhouse et al., 1990; Jermann et al., 1995).

To assess whether reconstructions yielded proteins that were plausible as intermediates in the evolution of the RNase family, the catalytic activities, substrate specificities, and thermal–proteolytic stabilities of the resurrected ancestral RNases were examined. Most of the resurrected proteins, and all of those corresponding to proteins expected in artiodactyls

TABLE 2
Kinetic Properties of Reconstructed Ancestral Ribonucleases[a]

RNase	Ancestor of:	k_{cat}/K_M UpA $\times 10^6$	k_{cat}/K_M as % of RNase A	poly(U) Relative to RNase A	poly(A)–poly(U) Relative to RNase A
RNase A		5.0	100	100	1.0
a[b]	Ox, buffalo, eland	6.1	122	106	1.4
b	Ox, buffalo, eland, nilgai	5.9	118	112	1.0
c	*b* and the gazelles	4.5	91	97	0.8
d	Bovids	3.9	78	86	0.9
e	Deer	3.6	73	77	1.0
f	Deer, pronghorn, giraffe	3.3	67	103	1.0
g	Pecora	4.6	94	87	1.0
h_1	Pecora and seminal RNase	5.5	111	106	5.2
h_2	Pecora and seminal RNase	6.5	130	106	5.2
i_1	Ruminata	4.5	90	96	5.0
i_2	Ruminata	5.2	104	80	4.3
j_1	Artiodactyla	3.7	74	73	4.6
j_2	Artiodactyla	3.3	66	51	2.7

[a]RNase names refer to nodes in the evolutionary tree shown in Figure 18. All assays were performed at 25°C.
[b]Reconstructed ancient sequences are designated by italic lowercase letters.

living after *Archaeomeryx*, behaved as expected for digestive enzymes. This was especially apparent from their kinetic properties (Table 2). Modern digestive RNases are catalytically active against small RNA substrates and single-stranded RNA (Blackburn and Moore, 1982). The RNase from *Pachyportax* was also, as were many of the earlier RNases. Thus, if one assumes that these catalytic properties are indicative of a digestive enzyme, these ancestral proteins were digestive enzymes as well.

This was also true quantitatively. Thus, the k_{cat}/K_M values for the putative ancestral RNases with the ribodinucleotide uridylyl 3′,5′-adenosine (UpA) as a substrate (Ipata and Felicioli, 1968) in many ancient artiodactyls proved not to differ by more than 25% from those of contemporary bovine digestive RNase (Table 2). With single-stranded poly(U) as substrate, the variance in catalytic activity was even smaller (18%).

TABLE 3
Thermal Transition Temperatures for Reconstructed Ancient Ribonucleases[a]

Enzyme	T_m (°C)	ΔT_m (°C)
RNase A[b]	59.3	0.0
RNase A[c]	59.7	+0.4
a[d]	60.6	+1.3
b	61.0	+1.7
c	60.7	+1.4
d	58.4	−0.9
e	61.1	+1.8
f	58.6	−0.7
g	59.1	−0.2
h_1	58.9	−0.5
h_2	59.3	0.0
i_1	58.2	−1.1
i_2	58.7	−0.6
j_1	56.5	−2.8
j_2	57.1	−2.2

[a]Thermal unfolding-proteolytic digestion temperatures (±0.5°C) were determined incubating the RNase ancestor in 100 m*M* NaOAc (pH 5.0) in the presence of trypsin.
[b]Expressed in *E. coli*.
[c]From Boehringer Mannheim.
[d]Reconstructed ancient sequences are designated by italic lowercase letters.

Like most digestive enzymes, modern digestive RNases are stable to thermal denaturation and cleavage by proteases. This suggested another metric for determining whether the ancestral proteins acted in the digestive tract. Using a method developed by Lang and Schmidt (1986), the sensitivity of the ancestral RNases to proteolysis as a function of temperature was measured (Table 3). Again, little change was observed in the thermal stability of the ancestral RNases back to the ancestral artiodactyls approximated by *Archaeomeryx* in the fossil record. The midpoints in the activity–temperature curves for these ancient proteins varied by only ±1.1°C compared with RNase A. This can be compared with typical experimental errors of ±0.5°C.

Had all of the ancestral RNases behaved like modern RNases, the resulting evolutionary narrative would have had little interest. The experiments in paleobiochemistry became interesting because the behavior of RNases resurrected from organisms *more* ancient than the last common

ancestor of the true ruminants (*Archaeomeryx* and earlier) did *not* behave like digestive enzymes using these metrics.

These more ancient resurrected ancestral RNases displayed a fivefold increase in catalytic activity against double-stranded RNA [poly(A)–poly(U)]. This is not a digestive substrate. Further, the ancestral RNases showed an increased ability to bind and melt double-stranded DNA. Bovine digestive RNase A has only low catalytic activity against duplex RNA under physiological conditions and does not bind and melt duplex DNA; these activities are presumably not needed for a digestive enzyme. At the same time, the catalytic activity of the candidate ancestral sequences against single-stranded RNA and short RNA fragments, the kinds of substrates that are expected in the digestive tract, was substantially lower (by a factor of 5) than in the modern proteins. Proposing that these phenotypes can be used as metrics, Jermann et al. (1995) concluded that RNases in artiodactyls that were ancestral to *Archaeomeryx* were *not* digestive enzymes.

A similar inference was drawn from stability studies. The more ancient ancestors displayed a modest but significant decrease in thermal–proteolytic stability using the assay of Lang and Schmidt (1986). Jermann et al. (1995) considered the possibility that the decrease in stability might reflect an incorrect reconstruction. A less stable enzyme, and a lower activity against single-stranded RNA, for example, might imply simply that the incorrect amino acid sequence was inferred for the ancestral protein. The fact that catalytic activity against double-stranded RNA, and the ability to melt duplex RNA, were *higher* in the ancestors argued against this possibility.

The issue was probed further by considering the ambiguity in the tree. The connectivity of deep branches in the artiodactyl evolutionary tree is not fully clarified by either the sequence data or the fossil record (Table 4) (Graur, 1993). This created a degree of ambiguity in the ancestral sequences. To manage this ambiguity, Jermann et al. synthesized a variety of alternative candidate ancestral RNase sequences. These effectively covered all of the ambiguity in the tree topology and the resulting ambiguity in the sequences. The survey showed that the measured phenotype (and the consequent biological interpretation) were robust with respect to the ambiguity.

Site 38 proved to be especially interesting. A variant of h_1 (Figure 18) that restores Asp at position 38 (as in RNase A) has a catalytic activity against duplex RNA similar to that of RNase A (Jermann et al., 1995; Opitz et al., 1998). Conversely, the variant of RNase A that introduces Gly alone at position 38 has catalytic activity against duplex RNA, essentially that of

TABLE 4
Sequence Changes in Reconstructed Ancient Ribonucleases[a]

Bovine RNase A		Ancestral Sequence												
		a	*b*	*c*	*d*	*e*	*f*	*g*	h_1	h_2	i_1	i_2	j_1	j_2
3	Thr	Thr	Thr	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser*	Thr*	Ser	Ser
6	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Glu	Glu	Lys	Lys
15	Ser	Ser	Ser	Ser	Ser	Pro	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser
16	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Gly*	Gly*	Gly*	Gly*	Gly	Gly
17	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Ser*	Thr*	Ser	Ser	Ser	Ser
19	Ala	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser
20	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ser	Ser	Ser	Ser	Ser	Ser
22	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Asn*	Asn*	Asn	Asn
31	Lys	Lys	Lys	Lys	Lys	Gln	Lys	Lys	Lys	Lys	Lys	Lys	Lys*	Lys*
32	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Arg	Arg	Arg	Arg
34	Asn	Asn	Asn	Asn	Asn	Asn	Asn	Asn*	Lys*	Lys*	Lys*	Lys*	Asn	Asn
35	Leu	Met	Met	Leu	Leu	Leu	Leu	Leu*	Met	Met	Met	Met	Met	Met
37	Lys	Lys	Gln	Gln	Gln	Gln	Gln	Gln	Gln	Gln	Gln	Gln	Gln	Gln
38	Asp	Asp	Asp	Asp	Asp	Asp	Asp	Asp	Gly	Gly	Gly	Gly	Gly	Gly
59	Ser	Ser	Ser	Ser	Ser	Phe	Ser	Ser	Ser	Ser	Ser	Ser	Ser	Ser
64	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Thr	Thr	Thr	Thr	Thr	Thr
70	Thr	Thr	Thr	Thr	Thr	Ser	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr
76	Tyr	Tyr	Tyr	Tyr	Tyr	Asn	Tyr	Asn	Asn	Asn	Asn	Asn	Asn	Asn
78	Thr	Thr	Thr	Thr	Thr	Ala	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr
80	Ser	Ser	Ser	Ser	Ser	His	Ser	Arg*	Arg*	Arg*	His	His	His	His
96	Ala	Ala	Ala	Ala	Ala	Val	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala
100	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Thr	Ser	Ser	Ser	Ser
102	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Ala	Val	Val	Val*	Val*	Val*	Glu*
103	Asn	Lys	Lys	Lys	Glu	Glu	Glu	Glu	Glu	Glu	Gln	Gln	Gln	Gln

[a]Reconstructed ancient sequences are designated by italic lowercase letters. The ancient sequences were adapted by applying maximum likelihood. Amino acids marked with asterists indicate positions where assignment depends on ambiguous parsimony reconstructions, or might be changed by plausible reorganization of the tree. In several of these cases, multiple sequences were reconstructed; subscripts indicate alternative sequence reconstructions for one node in the tree.

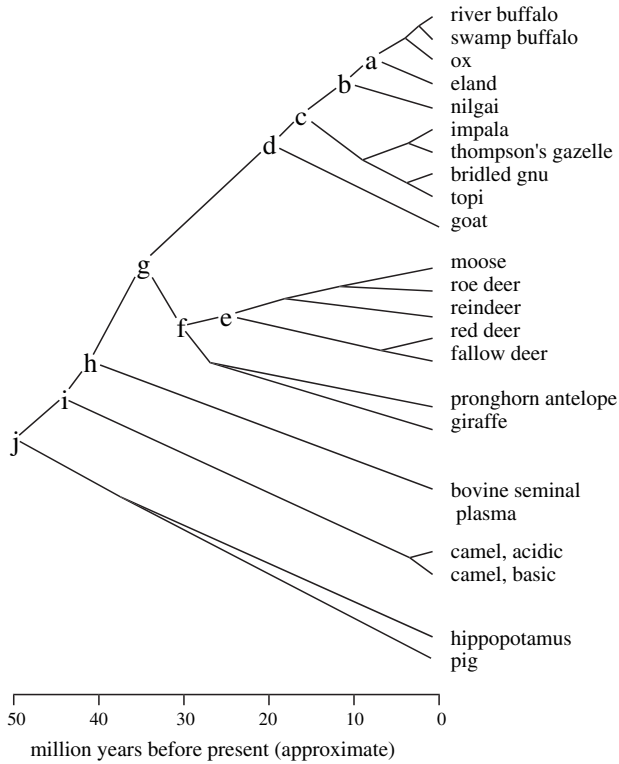


FIGURE 18. Evolutionary tree used in the analysis of ancestral RNases. Lowercase letters in the nodes in the graph designate putative intermediates in the evolution of the protein family (see Table 4). The time scale is approximate.

ancestor *h*. These results show that substitution at a single position, 38, accounts for essentially all of the increased catalytic activity against duplex RNA in ancestor *h*.

The reconstructed amino acids at position 38 are unambiguous before and after the *Archaeomeryx* sequence. Thus, it is highly probable that the changes in catalytic activity against duplex RNA in fact occurred in RNases as the ruminant RNases arose. In one interpretation, catalytic activity against duplex RNA was not necessary in the descendent RNases, and therefore was lost. This implies that the replacement of Gly

38 by Asp in the evolution of ancestor *g* from ancestor *h* was neutral. Jermann et al. (1995) could not, however, rule out an alternative model: that Asp 38 confers positive selective advantage on RNases found in the ruminants.

2. *Understanding the Origin of Ruminant Digestion*

The experimental paleobiochemical data within the pancreatic RNase family suggested a coherent evolutionary narrative consistent with the Barnard hypothesis. RNases with increased stability, decreased catalytic activity against duplex RNA, decreased ability to bind and melt duplex DNA, and increased activity against single-stranded RNA and small RNA substrates emerged near the time when *Archaeomeryx* lived. The properties that increased are essential for digestive function; the properties that decreased are not. *Archaeomeryx* was the first artiodactyl to be a true ruminant. This implies that a digestive RNase emerged when ruminant digestion emerged.

This converts the Barnard hypothesis (or, pejoratively, the Barnard ''just so'' story) into a broader narrative. This narrative became still more compelling when the molecular behavior is joined to the historical record as known from the fossil and geological records. These records suggested that the camels, deer, and bovid artiodactyl genera diverged about 40 Ma, together with ruminant digestion and the digestive RNases to support it, at the time of global climate change that began at the end of the Eocene.

This climate change eventually involved the lowering of the mean temperature of Earth by about 17°C, and the drying of large parts of the surface (Janis et al., 1998). This, in turn, was almost certainly causally related to the emergence of grasses as a predominant source of vegetable food in many ecosystems. Tropical rain forests receded, grasslands emerged, and the interactions between herbivores and their foliage changed. Grasses offer poor nutrition compared to many other flora, and ruminant physiology appears to have substantial adaptive value when eating grasses.

This, in turn, may help explain why ruminant artiodactyls were enormously successful in competition with the herbivorous perissodactyls (for example, horses, tapirs, and rhinoceroses) as the global climate change proceeded. Today, nearly 200 species of artiodactyls have displaced the approximately 250 species of perissodactyls that were found in the tropical

Eocene. Today, only three species groups of perissodactyls survive. This is the principal reason that resurrection of enzymes from the dawn horse will remain outside the reach of contemporary paleomolecular biologists.

3. Ribonuclease Homologs Involved in Unexpected Biological Activities

The paleobiochemical experiments with pancreatic RNases suggested that RNases having a digestive function emerged in artiodactyls from a nondigestive precursor about 40 Ma. This implies, in turn, that nondigestive cousins of digestive RNases might remain in the genomes of modern mammals, where they might continue to play a nondigestive role.

This suggestion, generated from the first experiments in paleogenetics, emerged at the same time as researchers were independently discovering nondigestive paralogs of digestive RNase A. These were termed *RIBAses* (ribonucleases with interesting biological activities) by D'Alessio et al. (1991). They include RNase homologs that display immunosuppressive (Soucek et al., 1986), cytostatic (Matousek, 1973), antitumor (Ardelt et al., 1991), endothelial cell stimulatory (Strydom et al., 1985), and lectinlike activities (Okabe et al., 1991). These proteins all appeared to be extracellular, based on their secretory signal peptides and the presence of disulfide bonds. Their existence suggested to some that perhaps a functional RNA existed outside cells (Benner, 1988).

These results suggested that the RNase A superfamily was extremely dynamic in vertebrates, with larger than typical amounts of gene duplication, paralog generation, and gene loss. In humans, for example, prior to the completion of the complete genome sequence, eight RNases were already known. These included the poorly named human pancreatic ribonuclease (RNase 1, which does not appear to be a protein specific for the pancreas), the equally poorly named eosinophil-derived neurotoxin (EDN or RNase 2, which does not appear to have a physiological role as a neurotoxin), the eosinophil-cationic protein (ECP or RNase 3, aptly named in the sense that the name captures about all that we know about the protein), RNase 4, angiogenin (also questionably named RNase 5), RNase 6 (sometimes known as k6), RNase 7 (Harder and Schroder, 2002; Zhang et al., 2003), and RNase 8 (Zhang et al., 2002).

The *in silico* analysis of the human genome showed that the human genes RNase 1 to 8 lie on chromosome 14q11.2 as a cluster of about 368 kb. From the centromere to the telomere, the genes are angiogenin (RNase 5),

RNase 4, RNase 6, RNase 1, ECP (RNase 3), an EDN pseudogene, EDN itself (RNase 2), RNase 7, and RNase 8, separated from each other by 6- to 90-kb intervals. The genome also identified two new human RNase homologs (RNases 9 and 10) in this cluster preceding angiogenin. In addition, three new open reading frames (ORFs) sharing a number of common features with other RNases were found. Beintema therefore proposed to name these RNases 11, 12, and 13. RNases 11 and 12 are located between RNase 9 and angiogenin. RNase 13 lies on the centromere side of RNase 7 and has a transcriptional direction opposite to that of RNases 7 and 8. The human genome reveals no other ORFs with significant similarity to these RNase genes. Therefore, it is likely that all human RNase A superfamily members have been identified.

As in humans, rat RNase genes are located on one chromosome (15p14) in a single cluster. The cluster in the rat genome contains the RNase family in the same syntenic order and transcriptional direction as in humans, with only a few exceptions. The RNase 1 family (RNase1h, RNase1g, and RNase1y), the eosinophil-associated RNase family (EAR) (R15-17, ECP, R-pseudogene, and Ear3), and the angiogenin family (Ang1 and Ang2) have undergone expansion in the rat (Zhao et al., 1998; Singhania et al., 1999; Dubois et al., 2002). Further, orthologs of human RNases 7 and 8 are not present in the rat genome. This permits us to propose a relatively coherent model for the order of gene creation in the time separating primates and rodents, and a list of the RNase homologs likely to have been present in the last common ancestor of primates and rodents.

The dynamic behavior of this group of genes is shown by the differences separating the rat and mouse groups. In mouse, two RNase gene clusters are found, on mouse chromosome 14qB-qC1 (bcluster AQ) and chromosome 10qB1 (bcluster BQ). Cluster A is syntenic to the human and rat clusters and is essentially identical to the rat cluster in gene content and order except for substantial expansions of the EAR and angiogenin gene subfamilies. Cluster B emerged in mouse after the mouse–rat divergence and contains only genes and pseudogenes that belong to the EAR and angiogenin subfamilies. It also includes a large number of pseudogenes.

This level of diversity presents many ‘‘why?’’ questions that might be addressed using molecular paleoscience. To date, two of these have been pursued, one in the Rosenberg laboratory, the second in the Benner laboratory.

4. Paleobiochemistry with Eosinophil RNase Homologs

In an effort to understand more about the function of these abundant RNase paralogs, Zhang and Rosenberg (2002) examined the eosinophil-derived neurotoxin (EDN) and eosinophil cationic protein (ECP) in primates. These proteins arose by gene duplication some 30 Ma in an African primate that was ancestral to humans and Old World monkeys.

Zhang and Rosenberg first asked the basic question: Why do eosinophils have two RNase paralogs? Eosinophils are associated with asthma, infective wheezing, and eczema (Onorato et al., 1996); their role in a nondiseased state remains enigmatic. Some textbooks say that eosinophils function to destroy larger parasites and modulate allergic inflammatory responses. Others suggest that eosinophils defend their host from outside agents, with allergic diseases arising as an undesired side effect.

Earlier work by Zhang, Rosenberg, and their associates had suggested that ECP and EDN might contribute to organismic defense in other ways. ECP kills bacteria *in vitro*, and EDN inactivates retroviruses (Rosenberg and Domachowske, 2001). *In silico* analysis of reconstructed ancestral sequences in primates suggested that the proteins had suffered rapid sequence change near the time of the duplication that generated the paralogs, a change that might account for their differing behaviors *in vitro* (Zhang et al., 1998). This suggests that in primate evolution, mutations in EDN and ECP may have adapted them for different, specialized roles during the episodes of rapid sequence evolution.

To obtain a more densely articulated tree for the protein family, Zhang and Rosenberg sequenced additional genes from various primates. They used these sequences to better reconstruct ancestral sequences for ancient EDN–ECPs. They estimated the posterior probabilities of these ancestral sequences using Bayesian inference. Then they resurrected these ancient proteins by cloning and expressing their genes (Zhang and Rosenberg, 2002).

Guiding the experimental work was the hypothesis that the antiretroviral activity of EDN might be related to the ability of the protein to cleave RNA. Studies of the ancestral proteins allowed Zhang and Rosenberg to retrace the origins of the antiretroviral and RNA cleaving activities of EDN. Both the ribonuclease and antiviral activities of the last common ancestor of ECP and EDN, which lived about 30 Ma, were low. Both activities increased in the EDN lineage after its emergence by duplication.

Zhang and Rosenberg showed that two replacements (at sites 64 and 132) in the sequence were together required to increase the ribonucleolytic activity of the protein; neither alone was sufficient. Zhang and Rosenberg then analyzed the three-dimensional crystal structure of EDN to offer possible explanations for the interconnection between sites suffering replacement and the changes in behavior that they created.

Zhang and Rosenberg concluded that in the EDN–ECP family, either of the two replacements at sites 64 and 132 individually had little impact on behavior. Each does, however, provide the context for the other to have an impact on behavior. This provides one example where a ''neutral'' (perhaps better, ''behaviorally inconsequential'') replacement might have set the stage for a second adaptive replacement.

This observation influences how protein engineering is done in general. Virtually all analyses of divergent evolution treat protein sequences as if they were linear strings of letters (Benner et al., 1998). With this treatment, each site is modeled to suffer replacement independent of all others, future replacement at a site is viewed as being independent of past replacement, and patterns of replacements are treated as being the same at each site. This has long been known to be an approximation, useful primarily for mathematical analysis (the ''spherical cow''). Understanding higher-order features of protein sequence divergence has offered *in silico* approaches to some of the most puzzling conundrums in biological chemistry, including how to predict the folded structure of proteins from sequence data (Benner et al., 1997a) and how to assign function to protein sequences (Benner et al., 1998). The results of Zhang and Rosenberg provide an experimental case where higher-order analysis is necessary to understand a biomolecular phenomenon.

Another interpretive strategy involving resurrected proteins was suggested from the results produced by Zhang and Rosenberg. This strategy identifies physiologically relevant *in vitro* behaviors for a protein where new biological function has emerged, as indicated by an episode of rapid (and therefore presumably adaptive) sequence evolution. The strategy examines the behavior of proteins resurrected from points in history before and after the episode of adaptive evolution. Those behaviors that are rapidly changing during the episode of adaptive sequence evolution, by hypothesis, confer a selective value on the protein in its new function, and therefore are relevant to the change in function, either directly or by close coupling to behaviors that are. The *in vitro* properties that are the same at the beginning and end of this episode are not relevant to the change in function.

While the number of amino acids changing is insufficient to make the case statistically compelling, the rate of change in the EDN lineage is strongly suggestive of adaptive evolution (Zhang et al., 1998). The antiviral and ribonucleolytic activities of the proteins before and after the adaptive episode in the EDN lineage are quite different. Benner (2002), interpreting the data of Zhang et al., suggested that these activities are important to the emerging physiological role for EDN. This adds support, perhaps only modest, for the notion that the antiviral activity of EDN became important in Old World primates about 30 Ma.

The timing of the emergence of the ECP–EDN pair in Old World primates might also contain information. The duplication occurred near the start of a global climatic deterioration that has continued until the present, with the Ice Ages in the past million years being the culmination (we hope) of this deterioration. These are the same changes as those that presumably drove the selection of ruminant digestion. If EDN, ECP, and eosinophils are part of a defensive system, it is appropriate to ask: What happened during the Oligocene that might have encouraged this type of system to be selected? Why might new defenses against retroviruses be needed at this time? If we are able to address these questions, we might better understand how to improve our immune defenses against viral infections, an area of biomedical research that is in need of rapid progress.

5. Paleobiochemistry with Ribonuclease Homologs in Bovine Seminal Fluid

New biomolecular function is believed to arise, at least in recent times, largely through recruitment of existing proteins having established roles to play new roles following gene duplication (Ohno, 1970; Benner and Ellington, 1990). Under one model, one copy of a gene continues to evolve divergently under constraints dictated by the ancestral function. The duplicate, meanwhile, is unencumbered by a functional role and is free to search protein *structure space*. It may eventually come to encode new behaviors required for a new physiological function, and thereby confer selective advantage.

This model contains a well-recognized paradox. Because duplicate genes are not under selective pressure, they should also accumulate mutations that render them incapable of encoding a protein useful for any function. Most duplicates should therefore become pseudogenes (Lynch and Conery, 2000), inexpressible genetic information (‘‘junk DNA’’)

(Li et al., 1981), in just a few million years (Jukes and Kimura, 1984; Marshall et al., 1994). This limits the evolutionary value of a functionally unconstrained gene duplicate as a tool for exploring protein structure space in search of new behaviors that might confer selectable physiological function.

One of the nondigestive RNase subfamilies offered an interesting system to use experimental paleobiochemistry to study how new function arises in proteins. This focused on the seminal RNase, paralogs found in ruminants that arose by duplication of the RNase A gene just as it was becoming a digestive protein. In ox, seminal RNase is 23 amino acids different from pancreatic RNase A, a protein made of 124 resdues. As suggested by its name, the paralog is expressed in the seminal plasma, where it constitutes some 2% of total protein (D'Alessio et al., 1972). Seminal RNase has evolved to become a dimer with composite active sites. It binds tightly to anionic glycolipids (Opitz, 1995), including seminolipid, a fusogenic sulfated galactolipid found in bovine spermatozoa (Vos et al., 1994). Further, seminal RNase has immunosuppressive and cytostatic activities that pancreatic RNase A lacked (Soucek et al., 1986; Benner and Allemann, 1989).

Laboratory reconstructions of ancient RNases (Jermann et al., 1995) suggested that each of these traits was not present in the most recent common ancestor of seminal and pancreatic RNase, but rather, arose in the seminal lineage after the divergence of these two protein families. To learn more about how this remarkable example of evolutionary recruitment occurred, RNase genes were collected from peccary (*Tayassu pecari*), Eld's deer (*Cervus eldi*), domestic sheep (*Ovis aries*), oryx (*Oryx leucoryx*), saiga (*Saiga tatarica*), yellow backed duiker (*Cephalophus sylvicultor*), lesser kudu (*Tragelaphus imberbis*), and Cape buffalo (*Syncerus caffer caffer*). These diverged approximately in that order within the mammal order Artiodactyla (Carroll, 1988). These complemented the known genes for various pancreatic RNases (Carsana et al., 1988) and seminal RNases from ox (*Bos taurus*) (Preuss et al., 1990), giraffe (*Giraffa camelopardalis*) (Breukelman et al., 1993), and hog deer.

Seminal RNase genes are distinguished from their pancreatic cousins by several marker substitutions introduced early after the gene duplication, including Pro 19, Cys 32, and Lys 62. By this standard, the genes from saiga, sheep, duiker, kudu, and the buffaloes were all assigned to the seminal RNase family. No evidence for a seminal-like gene could be found in peccary. Thus, these data are consistent with an analysis of previously

published genes that places the gene duplication separating pancreatic and seminal RNases about 35 million years before present (Beintema et al., 1988), and preceding the divergence of giraffe, sheep, saiga, duiker, kudu, Cape buffalo, and ox, in that order, consistent with mitochondrial sequence data (Allard et al., 1992; Hernandez Fernandez and Vrba, 2005) and global phylogenetic analyses of Ruminanta (Hassanin and Douzery, 2003).

Sequence analysis shows that the seminal RNase gene from hog deer almost certainly could not serve a physiological function as a folded stable protein. A five-base insertion in the beginning of the protein creates a frame shift. Lesions are also present in the seminal RNase gene from sheep, roe deer, and Cape buffalo.

To show that these seminal genes were indeed not expressed in semen, seminal plasmas from 15 artiodactyls were examined [ox, forest buffalo (*Syncerus caffer nanus*), Cape buffalo, kudu, sitatunga (*Tragelaphus spekei*), nyala (*Tragelaphus angasi*), eland (*Tragelaphus oryx*), Maxwell's duiker (*Cephalophus monticola maxwelli*), yellow backed duiker, suni (*Neotragus moschatus*), sable antelope (*Hippotragus niger*), impala (*Aepyceros melampus*), saiga (*Saiga tatarica*), sheep (*Ovis aries*), and Elds deer]. Catalytically active RNase was not detected in the seminal plasma in significant amounts in any artiodactyl genus diverging before the Cape buffalo, except in *Ovis*. Independent mutagenesis experiments showed that the proteins encoded by these genes, all carrying a Cys at position 32, should form dimers (Trautwein, 1991; Raillard, 1993; Jermann 1995; Opitz, 1995). By Western blotting, however, only small amounts of a monomeric, presumably pancreatic RNase, were detected in these seminal plasmas. In contrast, the seminal plasmas of forest buffalo, Cape buffalo, and ox all contained substantial amounts of Western blot–active RNase (Kleineidam et al., 1999). Only in the seminal plasma of ox, however, is seminal RNase expressed. Even though the gene is intact in water buffalo, no expressed protein could be found in its seminal plasma.

The seminal plasma from the *Ovis* genus (sheep and goat) was a notable exception. Sheep seminal plasma contained significant amounts of RNase protein and the corresponding ribonucleolytic activity. To learn whether RNases in the *Ovis* seminal plasma were derived from a seminal RNase gene, the RNase from goat seminal plasma was isolated, purified, and sequenced by tryptic cleavage and Edman degradation. Both Edman degradation (covering 80% of the sequence) and MALDI mass

spectroscopy showed that the sequence of the RNase isolated from goat seminal plasma is identical to the sequence of its pancreatic RNase (Beintema et al., 1988; Jermann, 1995). This shows that the RNase in *Ovis* seminal plasma is not expressed from a seminal RNase gene, but from the *Ovis* pancreatic gene. To confirm this conclusion, a fragment of the seminal RNase gene from sheep was sequenced and shown to be different in structure from the pancreatic gene.

Before a detailed paleogentic analysis, these results could be interpreted to be inconsistent with a model that the seminal RNase gene family gradually developed a new ''seminal'' function by stepwise point mutation and continuous selection under functional constraints in the seminal plasma following gene duplication. Rather, the duplicate RNase gene seems initially to have served no function at all. It therefore suffered damage, only to be repaired much later in evolution, after the divergence of kudu, but before the divergence of Cape buffalo, from the lineage leading to ox. Clades containing the saiga, duiker, and sheep are known in the early Miocene (23.8 to 16.4 Ma), and clades containing the kudu and Cape buffalo are known in the late Miocene (11.2 to 5.3 Ma). This view is supported by the fact that the gene seemed to have no function for 30 Ma after duplication until the emergence of the bovine lineage. In fact, it has been calculated that the half-life of a gene under no selection is 10 to 14 Ma (Lynch and Conery, 2000). Despite the incompleteness of the fossil record, one might conclude that the damaged gene was repaired extremely rapidly, in only a few million years. The paleogenetics study will suggest a different inference, however.

But what was this new function and its molecular phenotype? To address these questions, Sassi (2005) set out to reconstruct and resurrect the ancestral seminal proteins. The trees in Figure 19 shows the nodes where ancestral sequences were inferred, using a likelihood method. These nodes include the evolutionary period where the new biological function might be arising. Three different evolutionary models, one amino acid based and two codon based, were used to make the reconstructions. Two outgroups were also considered, those holding the pancreatic RNases and brain RNases, as the data did not unambiguously force the conclusion that one of these two RNAse subfamilies was the closest outgroup. Uncertainty in the topology of the tree holding the seminal RNases (the relative placement of the okapi sequence) was also considered; two topologies based on molecular and paleontological data were chosen for the analysis (Figure 19). In an effort to manage ambiguities, all

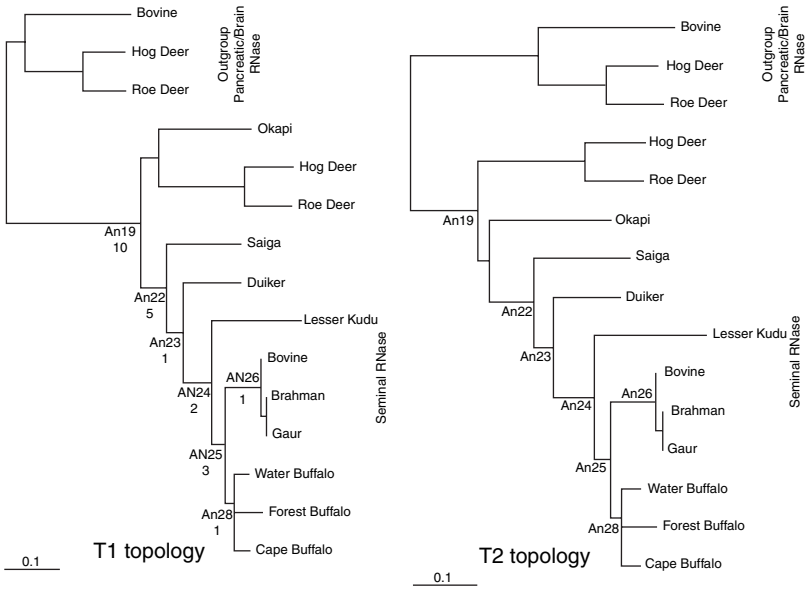


FIGURE 19. Tree showing the divergence of seminal ribonucleases from ruminants. The two topologies considered are shown. The numbers next to the nodes indicate the number of different candidate ancestral proteins prepared to ensure that the ambiguity in the ancestral sequences was covered.

candidate ancestral sequences were resurrected whenever the reconstructions disagreed.

In the first part of the paleogentic study, the distribution of ancestral replacements on the three-dimensional structure of seminal RNase was examined. The replacements followed a specific pattern of distribution. All of the active-site residues remained conserved after the gene duplication. Moreover, the RNA-binding site was also conserved. Most of the replacements were concentrated on the surface of the protein and away from the RNA-binding site. This replacement pattern is consistent with an evolutionary path where the enzymatic function of the protein was conserved; it is not consistent with an inference that the ancestral seminal RNase genes were pseudogenes. Furthermore, the lesions causing the pseudogene formation in the different lineages are different. These two observations taken together imply that the ancestral seminal RNases were enzymatically active and that independent inactivation events converted

active genes in the different lineages into pseudogenes in many of the modern artiodactyls. Consistent with this model, the resurrected ancestral seminal RNases were all enzymatically active (hydrolyzing a fluorescently labeled RNA substrate) (Kelemen et al., 1999).

What, then, are the properties of seminal RNase that were the targets for natural selection over the past 30 Ma? As noted above, there are many *in vitro* behaviors to choose from. For some (such as antiproliferative activity against cancer cells in culture), it is difficult to rationalize how such behaviors might be important for a protein that exists in seminal plasma. But the site of expression of a protein is changeable over short periods of evolutionary time, meaning that we cannot be certain where seminal RNase has been expressed over its history.

Since this protein is expressed in the seminal fluid and is immunosuppressive, it was hypothesized that the protein may confer a selective reproductive advantage to bulls when the female reproductive tract mounts an immune response against invading sperm. Indeed, it has been shown in reproductive biology that in many species sperm encounter a defensive immune response and that in many cases seminal plasma is capable of repressing this response (James and Hargreave, 1984; Schroder et al., 1990; Kelly and Critchley, 1997).

To use paleogenetics to generate an experimental test for this hypothesis, Sassi (2005) exploited the strategy to identify the physiologically relevant *in vitro* behaviors for a newly emerging function. As noted above, the strategy examines the behavior of proteins resurrected from points in history before and after the presumed episode of adaptive evolution. The *in vitro* behaviors that are changing rapidly during this episode are inferred to be those relevant to adaptive change. The *in vitro* behaviors that are the same at the beginning and end of this episode are not relevant to the change in function.

Episodes of *adaptive evolution* are frequently inferred from high-normalized nonsynonymous/synonymous ratios (significantly greater than unity), where amino acid replacements conferred new behaviors that conferred enhanced fitness on a protein subject to new functional demands. Thus, they characterize episodes where the derived sequence, at the end of the episode, has (in some sense) a physiological function different from that of the ancestral sequence at its beginning. Low ratios (<1, although these ratios approach zero in ''highly conserved'' proteins) characterize episodes where the ancestral and derived sequences at the beginning and end of the episode have the same physiological function.

The application of this tool in this gene family detected a phase of evolution during the emergence of bovine seminal ribonuclease after the ox diverged from the buffalo. A variety of models within the PAML program were used to determine d_N/d_S ratios for individual branches in the tree. The Alkaike information criterion (AIC) (Posada and Buckley, 2004) was then used to show that regardless of the model, the outgroup, or the tree, only the branch leading to the modern seminal RNase in ox, in its three forms (the gaur, Brahman, and ox) underwent adaptive evolution; for no other branch of the tree is this conclusion required. The d_N/d_S ratio was in the range 1.6 to 6, depending on the historical model, including the tree topology, choice of outgroup, and choice of codon substitution model.

This strongly suggests that the functional constrains on protein structure, and a correlated change in the physiological function of the protein, occurred in this episode. To identify which *in vitro* behaviors are also changing at this time, the genes encoding ancestral seminal RNase were synthesized by site-directed mutagenesis of a previously prepared RNase synthetic gene. The ancestral RNase candidates were expressed in *E. coli* and purified using a newly developed oligonucleotide affinity chromatography.

To address the biomolecular behaviors changing during this adaptive phase, Sassi (2005) then examined several biochemical and cell-based biomolecular behaviors. The k_{cat}/K_M ratio characterizing the enzyme's ability to catalyze the hydrolysis of a model fluorescently labeled RNA substrate (carboxyfluoresceinylhexyl-pdAUdAdAp-hexyltetramethylrhodamine, IDT) was similar to that of BS-RNase, implying that this behavior is not key to the newly emerging biological function in bovine seminal plasma. All of the candidate ancestral RNases could form dimers under oxidizing conditions, as does seminal RNA, implying that this behavior was not key to the newly emerging function. The rates of folding and other gross physical properties of the ancestors were also not greatly different in ancestral and modern seminal RNases.

In contrast, the immunosuppressivity of the seminal RNases, measured *in vitro* using a mixed lymphocyte reaction assay exploiting bovine leukocytes isolated from fresh peripheral bovine blood, increased noticeably in the descendents following the branch having adaptive evolution. This suggests that immunosuppression, as measured in this *in vitro* assay, is physiologically relevant for the new function of seminal RNase. This result was paralleled by results in mitogen induction assays, which is less "physiological." This suggests that the cell-based assays are

measuring a property that *is* important for the new biological function of seminal RNase.

Raines recently suggested that the cell-based activities of RNase might require a swapping of residues 1 to 20, mentioned above, to form composite active sites (Kim et al., 1995; Lee and Raines, 2005). Accordingly, the extent of the swap was measured using a divinylsulfone cross-linking reagent following the procedure of Ciglic et al. (1998). While the extent of swapping may be sensitive to the precise conditions under which the proteins were renatured, the extent of swapping measured *in vitro* also increases during the episode of adaptive evolution. This confirms the hypothesis of Raines and suggests a structural feature relevant to an adaptive change as well as a behavior.

It is important to note that these paleogenetic experiments suggest inferences about the structural changes and behavioral changes that may be important to changing physiological function without recourse to specific studies on the living animals. Further, these inferences are robust with respect to the ambiguities inherent in the reconstruction of historical states from derived sequences. This study presents an example where the evolutionary history of a gene and the physiological function of the protein were both unknown but the resurrection of the ancestral protein provided evidence for a hypothesis and hints on the evolutionary events shaping this gene's history.

6. Lessons Learned from Ribonuclease Resurrections

The ribonuclease family contains the best-developed example of the use of paleomolecular resurrections to understand protein function. It also illustrates most of the key issues that must be addressed when implementing this paradigm. This includes the management of ambiguities. In the cases reviewed here, additional sequences were obtained from additional organisms to increase the articulation of the evolutionary tree and thereby reduce ambiguity in the ancestral sequences inferred. When ambiguities remained, multiple candidate ancestral sequences were resurrected to determine that the behavior subject to biological interpretation was robust with respect to the ambiguity.

Paleogenetics experiments with the ribonuclease family also show the value of maximum likelihood tools in reconstructing ancestral sequences. The simplest parsimony tools, which minimize the number of changes in a tree, easily generate ambiguity by swaps around short branches. Ancestral

character states are less likely to be confused by incorrect detailed topology of a tree when they are constructed using maximum likelihood tools than by maximum parsimony tools.

More important, however, these examples show the potential of molecular paleoscience as a strategy to sort out the complexities of biological function in complex genome systems. Here, the potential of this strategy has only begun to be explored. In the long term we expect that paleomolecular resurrections will allow us to understand changing biomolecular function in an ecological and planetary systems context. Margulis and others have referred to this as *planetary biology* (Margulis and West, 1993; Margulis and Guerrero, 1995).

Last, these examples show the value of paleomolecular resurrections in converting ''just so'' stories into serious scientific narratives that connect phenomenology inferred by correlation into a comprehensive historical–molecular hypothesis that incorporates experimental data and suggests new experiments. Thus, they offer a key example of how paleobiology might enter the mainstream of molecular biology as the number of genome sequences becomes large and frustration with their lack of meaning becomes still more widespread.

B. LYSOZYMES: TESTING NEUTRALITY AND PARALLEL EVOLUTION

At approximately the time that the first ancestral ribonucleases were being resurrected in the Benner laboratory in Zurich, Allan Wilson and his colleagues at the University of California were considering the history of the evolution of lysozymes from bird eggs (Malcolm et al., 1990). Their work focused on three positions (sites 40, 55, and 9). A crystal structure showed that these lie just beneath the active site cleft inside the folded structure in a hinge between the two globular domains of lysozyme.

Malcolm et al. (1990) noticed that the pattern of divergence at these sites was peculiar when compared with the pattern of divergence at other sites in the lysozyme family. These sites show little variation in other bird lysozymes. In one branch leading to the lysozyme in the last common ancestor of the quail (*Callipepla californica*), northern bobwhite (*Colinus virginianus*), and guinea fowl (*Numida meleagris*) from the next-higher node in the tree, three amino acid replacements (T40S, I55V, and S91T) were inferred. The ancestor therefore had TIS at sites 40, 55, and 91, respectively, whereas the descendent had SVT at those sites.

No other internal sites suffered replacement along this branch. Further, these sites are conserved in a variety of other birds, including the chachalaca, turkey, pheasant, chicken partridge, and Old World quail.

The evolutionary tree is insufficiently articulated to infer the order in which these three amino acids are replaced. Thus, Malcolm et al. asked if these three changes occurred in a specific sequence. A total of six paths $(= 3 + 2 + 1)$ lead stepwise from the TIS to the SVT trio. Proteins representing all intermediates in each of these paths of amino acids (with reference to these three sites only) were prepared. Malcolm et al. then asked whether all proteins had similar thermal stability.

In this work, the thermal melting transition temperatures of the modern lysozymes were used to provide the upper and lower limits of "normal" stability. Transitions at temperatures higher or lower than these bounds were interpreted as being functionally significant. Some of the intermediates constructed in this work had transition temperatures that were outside these bounds. This led the authors to suggest that these intermediates were less likely to be neutral variants than were others. Further, evidence was obtained suggesting that the temperature of unfolding correlated with the total volume of the side chains of the residues at these three sites.

This experiment had only limited interpretive goals. Egg lysozyme is frequently viewed as functioning to protect the egg from microbial attack. It remains an open question why North American birds would need to have a different set of residues in the hinge region of their lysozyme, or lysozymes with a particular thermal stability, to perform a function distinctive with respect to North American microbial agents. The distinctiveness of the changes observed in the branch, and the likelihood that such distinctive changes do not represent neutral drift, suggest that this question might have an interesting answer.

C. ANCESTRAL TRANSPOSABLE ELEMENTS

A substantial fraction of the genomes of many organisms, including mammals, consists of interspersed repetitive DNA sequences. These are primarily degenerate copies of transposable elements, units of DNA that can migrate to different parts of the genome. Transposable elements include retrotransposons, which are derived from RNA molecules that are transposed to DNA via cDNA intermediates, and transposons, which move directly from DNA to DNA. These can be *short interspersed nuclear*

elements (SINES, about 300 nucleotides long) or *long interspersed nuclear elements* (LINES, 6000 to 8000 nucleotides long), both of which contain internal promoters for RNA polymerase III; transposable elements with long terminal repeats (which can contain a reverse transcriptase or its remnants); or DNA transposons that have an ORF (or its remnants) that encodes a transposase.

Because transposable elements do not serve any known function except perhaps as an evolutionary tool, it is expected that their functional elements will sustain mutations that render them inactive over extended periods of evolutionary time. By going backward in time starting from inactive transposable elements in contemporary organisms, it should be possible to resurrect active ancestral transposable elements that delivered those transposable elements to the modern genomes. This has now been done in several laboratories.

1. Long Interspersed Repetitive Elements of Type 1

Long interspersed repetitive elements of type 1 (LINE-1 or L1) are examples of retrotransposons that encode reverse transcriptase, lack long terminal repeats, and appeared to transpose via a polyadenylated RNA intermediate. The F-type subfamily of LINE-1 retroposons apparently began to be dispersed throughout the mouse genome about 6 Ma. The resulting paralogs are only about 75% sequence identical when compared pairwise. In this way, they differ from subfamily A of LINE-1 retrotransposons, which are over 95% identical in pairwise sequence comparisons.

The A and F subfamilies of LINEs also differ in terms of their activity. Members of the A subfamily appear to be fully active as transcriptional promoters. Members of the F subfamily, in contrast, appear to be both transcriptionally and transpositionally inactive.

Adey et al. (1994) hypothesized that these F-type L1's are ''evolutionarily extinct'' descendents of an ancestral LINE-1 that was functionally active. The inactivation was caused by accumulation of mutations in the transcription initiation region. To test this idea, they analyzed an alignment of 30 F sequences to generate a consensus sequence of the promoter that approximated the sequence of the ancestral LINE promoter. They then resurrected that sequence and demonstrated that it was indeed functional in that the resurrected promoter was able to drive transcription in promoter assays.

The reconstruction did not follow a single evolutionary model based on a coherently defined tree and therefore does not represent an example of precise resurrection. Rather, a consensus of the F-type sequences was obtained using the program PRETTY from the Wisconsin Genetics Computing Group. This was compared with a consensus of a subset of the sequences of elements that were believed to have diverged more recently. Next, positions that displayed CpG hypermutability within the younger subset of F-type sequences were converted back to CpG in the presumed ancestral sequence. At sites where no consensus was observed for all 30 sequences, the consensus nucleotide within the younger subset was placed at that site in the ancestral sequence.

The authors recognized that this combination of analytical tools did not infer an ancestral sequence in a formally coherent way, but rather, approximated the ancestral sequence. The sequence that resulted from this analysis differed at 11 positions from a previously reported consensus sequence that was based on a smaller set of data.

The resulting consensus sequence was then resurrected by chemical synthesis. The consensus sequence was placed in front of a chloramphenicol transferase reporter gene, as were eight F sequences from modern mammals that served as controls. Each construct was transferred into undifferentiated mouse F9 teratocarcinoma cells. There, the ability of the ancient and modern promoters to direct the expression of the reporter protein was determined.

The resurrected ancestral promoter generated a high level of reporter expression, a level approximately equal to the level of expression generated by the most active A-type promoter known. In contrast, seven of the eight modern F-type promoters generated no detectable expression of the reporter. Thus, these results supported the hypothesis of Adey et al. (1994). The currently inactive F-type transposable elements do appear to be descendents of a promoter that was active approximately 6 Ma.

2. Sleeping Beauty Transposon

An analogous experiment was done by Ivics et al. (1997), who used a ''majority rule'' consensus strategy to approximate another ancestral transposon sequence. As with the F LINE subfamily, members of the Tc1/mariner superfamily of transposons in fish appear to be transpositionally inactive, due to the accumulation of mutations following divergence of an active transposon. Ivics et al. analyzed a dozen sequences to infer a

consensus sequence of an ancestral transposon and called it *Sleeping Beauty.*

This sequence was then resurrected and studied. The consensus ancestral transposase was shown to bind to the inverted repeats of salmonid transposons in a substrate-specific manner, and to mediate precise cut-and-paste transposition in fish as well as in mouse and human cells. This result suggested that the modern, inactive, transposable elements are descendents of a more ancient transposable element that dispersed itself in fish genomes some time ago, in part by horizontal transmission between species.

3. Frog Prince

Another member of the Tc1/mariner superfamily from the northern leopard frog (*Rana pipiens*) was resurrected in much the same manner as Sleeping Beauty (SB). In fact, SB does not show host-dependent restrictions but does show some transposition efficiency variability, depending on the cell line derived from different species (Neidhardt et al., 1990). Consequently, Ivics and co-workers (Miskey et al., 2003) set out to resurrect another vertebrate transposon in an effort to have another genomics tool with different characteristics.

R. pipiens genome was estimated to contain 8000 copies of transposable element, closely related to Txr elements in *Xenopus laevis*. To clone a few uninterrupted transposase open reading frames (ORFs) from this collection, the authors designed a method to trap them. This method selected for the uninterrupted transposase ORFs. They used the cloned sequences to generate a consensus of the transposase gene, and along with the inverted repeats from *R. pipiens*, they obtained all the necessary components for the transposon system *Frog Prince*. Frog Prince (FP) transposons were shown to be phylogenetically closer to the Txr elements than to the Sleeping Beauty/Tdr1 transposons. To test its transposition activity and compare it to that of Sleeping Beauty (SB), the assay developed for SB was used on HeLa cells. Indeed, FP was active in these cells and was shown to cross-mobilize with *X. laevis* transposon but not with Sleeping Beauty. This indicates that the transposon families in *X. laevis* and *R. pipiens* have diverged recently, in contrast with SB transposons, where the divergence is more distant. Of course, phylogenetically it is not surprising since FP originated from amphibians and SB from fishes.

FB was tested in different cell lines and compared to SB activity. The transposon systems overlapped and differed in their levels of activity in the

cell lines tested, achieving the authors' goal of obtaining a new active transposon system that had some different characteristics from SB, increasing the number and application of genomics tools. In fact, FP was more active than SB in zebrafish, demonstrating the advantage of using a phylogenetically distant transposon if higher activity was the goal. This is possibly due to the presence of SB-like transposons (Tc1 transposons) and the corresponding inhibitory mechanisms in zebrafish. These mechanisms would be ineffective against the phylogenetically distant FB.

The study is worth repeating using ancestral sequences inferred using specific evolutionary trees. This would provide an example of whether consensus reconstructions differ from rigorous reconstructions. Using just the salmonid sequences used for the Sleeping Beauty resurrection or just the paralogous sequences used for Frog Prince would not be enough to get a reconstruction where the phylogenetic inference produces a different sequence than in the consensus approach. In fact, the SB salmonid sequences and certainly the FP sequences do not have enough phylogenetic variation for the inference to necessarily be any different from the consensus. On the other hand, if one were to include more sequences from more divergent organisms the phylogentic inference would most certainly be more useful than a consensus approach.

4. Biomedical Applications of Transposons

The active ancestral transposon Sleeping Beauty became an intense focus of research as other research groups succeeded in improving its activity by site-directed mutagenesis (Yant et al., 2004). The SB system became a tool in gene therapy, generating a large number of publications and a biotechnology company, Discovery Genomics, Inc., with a goal to develop a gene therapy delivery system based on this resurrected transposon (Hermanson et al., 2004; Ivics et al., 2004). This is yet another example of a novel application of paleobioscience where resurrected biomolecules are being used as biomedical tools to combat human disease.

Alternatively, transposons could be used as a tool to identify genes involved in cancer by monitoring their ability to disrupt open-reading frames or regulatory regions with a genome. SB has recently been used as such a tool to identify cancer-related genes (Collier et al., 2005; Dupuy et al., 2005; Weiser and Justice, 2005). Identifying cancer genes today often involves random mutagenesis through the use of radiation, chemical agents, or viruses. These approaches raise serious concern, as it is difficult

to separate the cancer-causing mutations from the benign mutations. On the other hand, Sleeping Beauty presents a way to tag cancer genes because it is a transposon with a known sequence and one that is very divergent from its mouse homolog. As an example, Collier et al. (2005) designed a way to identify oncogenes using SB (altered regulation by SB when integrated upstream of the gene) or tumor suppressor genes (disrupted by SB when integrated in the middle of the sequence). This resurrected transposon will enable researchers not only to identify new cancer genes but also to dissect the unknown pathways for cancer formation in different types of tumors, tissues, and developmental stages. Oncogenesis is indeed a very complicated process with many possible pathways involving a large number of genes. Paleobioscience presents a powerful tool to shed light on the biology of cancer.

D. CHYMASE–ANGIOTENSIN CONVERTING ENZYME: UNDERSTANDING PROTEASE SPECIFICITY

As with the ribonucleases, proteases in vertebrates present a baffling diversity of paralogs that have arisen from gene duplication throughout the past 500 million years. Many proteases have been studied to determine their substrate specificities, where a parallel diversity in their behavior is also observed.

The serine proteases offer one example of this. As with ribonucleases, the classical serine proteases, such as trypsin and chymotrypsin, are digestive enzymes isolated from the digestive tract of oxen. These proteases are paralogs of nondigestive proteases that are found in many tissues. For example, chymases form a clade of serine proteases homologous to trypsin and chymotrypsin. These are secreted from mast cells. Once secreted, the chymases help process peptide hormones, are involved in the inflammatory response, and may aid in the expulsion of parasites (Miller, 1996; Knight et al., 2000). True physiological function remains unclear, but pathologically these enzymes have been involved in vascular disease and might be attractive drug targets (Doggrell and Wanstall, 2004).

Proteases carrying the name *chymase* from different species differ in the details of their substrate specificity. For example, human chymase cleaves angiotensin I between the Phe8–His9 residues to give angiotensin II, but does not then further cleave the Tyr4–Ile5 bond in angiotensin II. In contrast, the chymase from rat does degrade the Tyr4–Ile5 bond

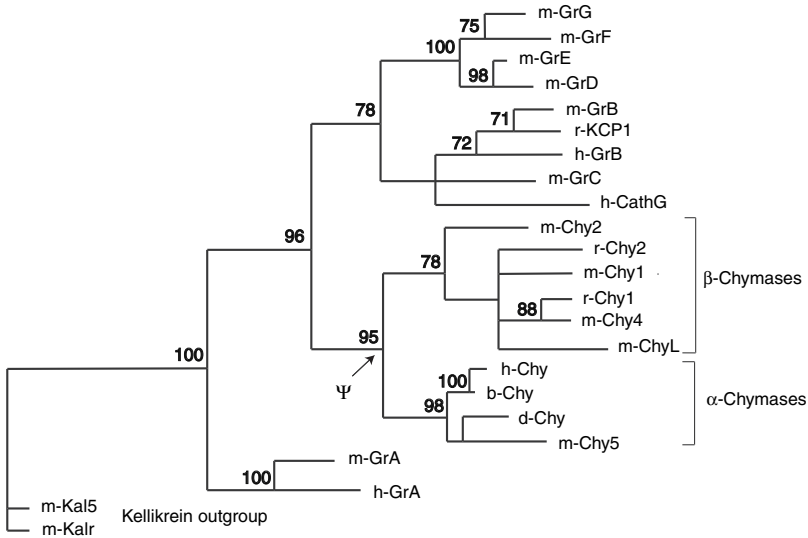


FIGURE 20. Tree showing relationship of a variety of serine proteases. Chy with the prefix letters h, b, d, r, and m indicate the human, baboon, dog, rat, and mouse chymases, respectively. Ψ indicates the ancestral chymase that was resurrected by Chandrasekharan et al. (1996). Bootstrap values are shown at the nodes.

in angiotensin II, leading to inactivation of the peptide hormone. This has many implications, not the least of which being that the pharmacology of hypertension management targeted at the angiotensin system is expected to differ in rats and humans.

The fact that these homologs in humans and rat display different substrate specificities raises the question: What was the specificity of the chymase in the last common ancestor of humans and rats? To answer this question, Chandrasekharan et al. (1996) applied parsimony analysis (using PAUP) to two dozen homologous serine proteases to reconstruct the last common ancestral chymase of modern chymases in humans and rats. In doing so, they used the kallikrein proteases as an outgroup (Figure 20). This ancestral protein was present in an organism such as *Eomaia* (Figure 21), which is the most primitive placental mammal fully attested to in the fossil record.

The resulting tree was supported at a 95% level by bootstrap analysis. There was, however, considerable ambiguity in the ancestral sequence

FIGURE 21. *Eomaia*, a primitive placental mammal, and *Sinodelphys*, a primitive marsupial, at the Jurassic–Cretaceous boundary, about 125 Ma, one of the best representative fossils at the time when mammalian protease systems were being developed. (From the S. A. Benner collection; photograph courtesy of Steven Benner.)

inferred using parsimony. The analysis did not identify a specific amino acid to occupy 15 sites in the ancestor. To generate the sequence to be resurrected, amino acids at seven of these sites were assigned arbitrarily. Amino acids at the remaining eight sites were chosen to ensure that the net positive charge of the ancestral chymase was +18. The candidate for the ancestral chymase differs from the modern chymase from 52 to 100 sites, depending on the modern descendent, corresponding to a difference of 23 to 34%.

A gene encoding the ancestral protease was then resurrected and expressed, and the ancestor was studied in the laboratory. The ancestor was shown to convert angiotensin I to angiotensin II efficiently with a turnover of about 700 per second. This kinetic performance was consistent with the hypothesis that the parsimony analysis generated a functionally active ancestor.

Relevant to the hypothesis, the ancestral chymase did not degrade angiotensin II further by cleavage of the Tyr4–Ile5 bond. In this respect, the behavior of the ancestor resembled the behavior of the more specific

modern human chymase, not the less specific rat chymase. This provided a case where protease specificity decreased over time. Chandrasekharan et al. (1996) noted that this contrasted with a common view of the evolution of protease specificity, where the protease begins with broad specificity that narrows over time as the role of the protease becomes narrower.

Is not clear exactly how to date the divergence of these enzymes. At the time the experiments were done, only a single chymase gene was known to be present in humans and baboons, and only a single chymase was known in dogs. In contrast, at least five alpha and beta chymases isozymes have been identified in mice and rats as of 1995. Correlation with the species tree suggests that the alpha and beta forms of chymases emerged long before mammals branched from therapsids. This implies that the nonrodent mammals lost some of the extra paralogs.

The narrative for the chymase family provides another interesting example where the biological information generated by an experiment in paleobiochemistry could not have been obtained in any other way. The narrative would, of course, be improved by resurrecting a sample of the alternative ancestral protease sequences to demonstrate that the inference is robust with respect to the reconstruction ambiguity. This would make it unnecessary to assume that the sites whose residues were assigned randomly have no impact on the interpreted phenotype.

Indeed, the abundance of genome sequences available today would make this system, as well as many protease systems, worth revisiting. It would be interesting to follow through paleogenetics the coevolution of these proteases as well as the angiotensin protein. The number of paralogs of different proteases in mammalian genomes, the diversity of protease types, and the complexity of the biological functions that they perform all suggest that molecular paleoscience should be a key part of any program to understand their biology.

E. RESURRECTION OF REGULATORY SYSTEMS: THE PAX SYSTEM

The difference in the morphology of different metazoans arises, in part, not from changes in the sequences of encoded proteins, but rather from changes in the regulation of expression of those proteins. Regulation of expression, in turn, often involves the specific binding of transcription factors to specific target DNA regulatory elements. Many of these factors are homologous. Thus, the divergent specificity of the DNA-binding

protein and binding partner is a historical process that can be examined using paleomolecular resurrections.

To explore one such history, Sun et al. (2002) examined the *Pax* system. *Pax* genes encode a well-conserved DNA-binding domain of 128 amino acids. In mammals, nine *Pax* genes (*Pax-1* to *Pax-9*) were known in 2002. These all play roles in the embryonic development of tissues and organs. Thus, *Pax* genes are implicated in human congenital defects (*Pax-2, 3, 6, 8, 9*), in the development of cancers (*Pax-3, 5, 7*), and in the development of the central nervous system (*Pax-2, 3, 5, 6, 7, 8*), the eye (*Pax-2, 6*), the pancreas (*Pax-4, 6*), and B-lymphocytes (*Pax-5*) (Engelkamp and van Heyningen, 1996; Dahl et al., 1997; Underhill, 2000).

Sequence analysis classifies *Pax* genes into five groups within two supergroups: *Pax-2*, *Pax-5*, *Pax-8*, *Pax-B*, *poxn/Pax-A*, and *Pax-6/ey* in supergroup I, and *Pax-1/Pax-9/poxm* and *Pax-3/Pax-7/gsb/gsbn* in supergroup II (Sun et al., 1997). *Pax* genes within each group often display similarities in their expression patterns; this may imply analogous roles in development (Chalepakis et al., 1993).

The process by which *Pax* genes and their binding sites duplicated and functionally diversified presents an example of the *transitional form conundrum* in divergent evolution. Briefly, this conundrum arises because incompletely specialized biomolecules that are intermediates in this evolution are expected to display cross-reactivity that might confuse their functional distinctiveness. Some of the amino acid changes between the duplicates are important in making these distinctions; others, however, reflect, instead, neutral evolution.

As was shown with ribonucleases, deliberate resurrection of ancestral forms can help distinguish amino acid replacements that are key to functional diversification from those that are neutral. More classical methods, including making hybrid structures that are widely used in molecular biology, do this less efficiently and might miss important functional residues. Two residues that come together in the folded structure of the protein, but are placed remotely in the primary structure, would be difficult to detect by classical molecular biology deletion analysis. In the *Pax* family, for example, hybrid constructs suggested that only three of the 30 amino acid differences between *Pax-5* and *Pax-6* paired domains are important for differences observed in their DNA-binding specificity (Czerny and Busslinger, 1995). These studies do not show, however, how this diversification actually occurred.

To address this issue in the *Pax* family, Sun et al. (2002) reconstructed a set of ancestral *Pax* sequences using the distance-based method developed by Zhang and Nei (1997). The ancestral proteins were then resurrected by expression of the gene in the *in vitro* reticulocyte-translation system. The proteins expressed were then assayed based on their ability to bind to various DNA sequences (as detected by gel mobility shifts) as well as a biological assay in *Drosophila*. For the DNA-binding assay, seven sequences were selected identified as *Pax-5* binding sites and are representative of the *Pax*-binding domain generally (Czerny et al., 1993).

Two of the ancestral sequences resurrected stand at the head of the two supergroups (Figure 22). Ancestor I (AnI) bound strongly to all test sequences except H2A2.2 (one of the seven binding sequences), which was bound less strongly. *Pax-2* and *Pax-A*, descendents of AnI, showed the same broad specificity as that of the ancestor, although overall they bound less strongly than ANI. In contrast, AN6, another descendent of AnI, showed a narrower binding specificity than that of its descendent, *Pax-6*.

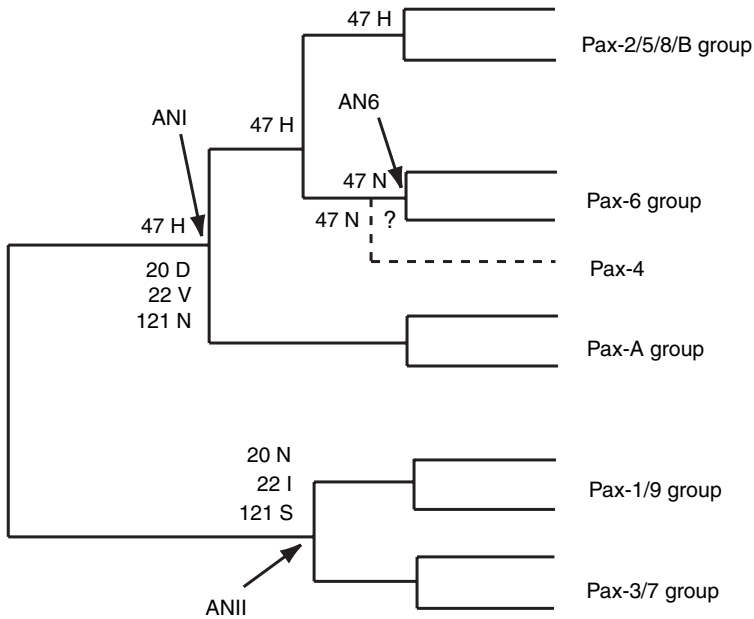


FIGURE 22. Tree relating the *Pax* genes. Note that *Pax-4* is not included in the analysis. The resurrected ancestral sequences are denoted by ANI, ANII, and AN6, at the nodes indicated.

These experiments suggest that the fundamental binding features of this supergroup were already established in the ancestor ANI.

ANII showed almost no binding with the test sequences. The mouse *Pax-1* paired domain (MPD1) showed modest binding with all of the test sequences (except for as H2A2.2), in contrast to the mouse *Pax-3* paired domain (MPD3), which showed little or no binding to the tested sequences.

Therefore, within the two supergroups, functional diversification during evolution appears to have involved changes from broad to narrow specificities in binding within the test sequences (ANI to AN6 and human *Pax-6*) and changes in binding affinities either across multiple binding sequences (e.g., ANI to mouse *Pax-2* and sea nettle *Pax-A*, and ANII to mouse *Pax-1*) or across a few sequences (ANII to mouse *Pax-3*). Further, changes in binding specificity can occur independent of the changes in binding affinity.

The next phase of the analysis is to sequence a wider diversity of *Pax* genes to better articulate the tree: to segment the branch leading to *Pax-4* so that it is not so long and so that its information can contribute to the resurrection.

Sun et al. (2002) then applied their analysis to identifying two amino acid changes that might dominate the differences in the DNA-binding properties of ANI and ANII. As there are seven amino acid differences between ANI and ANII, and 19 between AN6 and ANI, the authors used the difference in the evolutionary rates of these sites as a guide to select a more manageable subset of sites to test. They calculated the relative rates of amino acid substitution at the differing sites and selected the ones that show a relative low rate, as this suggests importance for the binding. Paleobioscience in this case combines ancestral reconstruction and site-specific evolutionary rates to select a functionally important set of sites. Testing the resurrected proteins and those with specific replacement in the sites selected is an efficient way to shed light on the functionally important residues. Of course, one has to be careful when comparing site-specific rates and deciding which are important, as deciding the cutoff for ''fast'' and ''slow'' could be challenging. However, the synergy between the reconstruction and the relative rates could make this problem manageable. The authors selected three different sites between ANI and ANII to investigate (sites 121, 22, and 20) and three others to investigate the difference between AN6 and ANI (sites 44, 47, and 66).

No single mutation changed the binding properties of ANI significantly, as shown in the binding patterns of ANI-NVN (D20N), ANI-DVS (N121S),

and ANI-DIN (V22I). Changes at positions 20 (N20D, ANII-DIS) and 121 (S121N, ANIININ) in ANII greatly increased the binding strength of ANII to the test sequences, however. The binding properties of ANI and ANII with combined mutations in positions 20, 22, and 121 were then examined. The hybrid ANI-NVS (D20N and N121S) had substantially decreased binding to test sequences relative to ANI, whereas the ANII-DIN (N20D and S121N), unlike ANII, bound efficiently to several sequences. When the binding strengths of ANI-NVS and ANII-DIN to test sequence 5S2A were compared using serially diluted concentrations of 5S2A, the binding of ANI-NVS was visibly weaker than that of ANII-DIN at every concentration tested. When the ratios of intensity of the shifted band versus the free band of 5S2A are calculated for ANI-NVS and ANII-DIN, the ratio for ANII-DIN ranges from 2- to 11-fold higher than that for ANI-NVS (five independent replicate assays). This range is rather larger than that expected from experimental error, suggesting that different preparations of the proteins differ in the specific activity of their total binding affinity. Nevertheless, ANII-DIN clearly had a higher affinity to the test sequences than that of ANI-NVS. DNA-binding properties of ANI and ANII are strongly influenced by the amino acid occupying sites 20 and 121, although replacements at all seven sites that separate the ANI and ANII sequences have some effect.

Next, they repeated the foregoing approach to evaluate the importance of the three sites selected (sites 44, 47, and 66) between AN6 and ANI. They tested R44Q, H47N, and G66R in ANI and the reciprocal changes in AN6. This series of experiment showed that only the amino acid change in site 47 is sufficient to cause a nearly complete specificity swap between ANI and AN6, although the other two sites (44 and 66) have minor effects. Next, Sun et al. (2002) investigated the importance of site 47 in an assay *in vivo* in *Drosophila*. Their rationale for this assay is that a difference in affinity, caused by a mutation in site 47, would result in an ectopic eye phenotype in *Drosophila*. This is based on previous studies showing that ectopic expression of *Pax-6* homologs and *eyeless* (homolog to the vertebrate *Pax-6* in *Drosophila*) result in the formation of supernumerary eyes in the fly (Brand and Perrimon, 1993; Halder et al., 1995). In addition, another study (Czerny et al., 1999) showed that the ectopic eye phenotype is also observed when the second *Pax-6* homolog, *twin of eyeless* (toy), is expressed ectopically. The two *Pax-6* paralogs in *Drosophila*, *eyeless* and *toy*, were also shown to produce proteins with different DNA-binding properties.

In the *in vivo* assay, the size and frequency of the ectopic eyes were evaluated as well the pigment concentration in thoraces with ectopic eyes. These measurements were used to compare overexpression of the wild-type *eyeless* transgene (EU transgene), *eyeless*-N47H (DP6M3 transgene with a mutation at site 47), *eyeless/Pax-2* (M2 transgene replacing the paired box domain with mouse *Pax-2*), and *eyeless/Pax-2*/H47N (M2M3 transgene with a mutation at site 47 in the introduced mouse *Pax-2*). As a result, the *in vivo* experiments showed that a change toward the *Pax-6* specific N at position 47 of the paired domain leads to larger and more ectopic eyes (or both). Smaller or fewer ectopic eyes were observed when the change was toward H47 present in *Pax-2, 5, 8.* An interesting result is that a complete replacement of the *eyeless* paired domain for the *Pax-2* paired domain did not abolish ectopic eye induction entirely but led instead to the induction of fewer and smaller ectopic eyes. This result strongly suggests that specificity can be conveyed by a single amino acid change but that the interaction of paired domains and their *in vivo* binding sites may be more flexible than expected.

Sun et al. (2002) analyzed these changes in light of the crystal structure of the protein to make further sense of their paleoscience result regarding functional residues (Xu et al., 1995, 1999). As with all experimental resurrections, certain simplifying assumptions were made. Thus, some *Pax* proteins contain a homeodomain, which may interact with the paired domain during DNA binding (Underhill et al., 1995; Fortin et al., 1998). By not considering this interaction in detail, this experiment represents a simplified approach to recapitulate the evolution of DNA-binding properties of regulatory proteins.

That this approximation is serviceable is shown by the insights into the evolution of *Pax* domains that these studies produced. The ancestors of supergroups I and II have very different binding properties from the panel of test sequences used in this study, and two amino acid substitutions have dominant effects on swapping their binding properties. Only one amino acid was enough to do the same for An6 and ANI. Because there is no reliable root to the phylogenetic tree of *Pax* genes, we cannot predict the sequence of the common ancestor of all *Pax* genes and its binding properties to complete the entire picture of early *Pax* evolution. However, it is intriguing to speculate that gene duplication of a common ancestor gave rise to the two ancestors of supergroups I and II, and these two ancestor genes mutated at positions 20 and 121 and acquired different DNA-binding properties to initiate differentiation of the two supergroups. Within

supergroup I, a similar result was observed as site 47. Although the paired domains of the *Pax-2, 5, 8* and *Pax-6* groups differ by 19 amino acids, their distinct DNA-binding properties are determined almost completely by a single amino acid change. Thus, a small number of amino acid changes can account in large part for the divergence in binding properties among the known paired domains.

As with ribonuclease, Sun et al. (2002) proposed that this evolutionary approach is an efficient strategy to select candidate sites responsible for the functional divergence between genes. In this example, they used this approach to identify candidate amino acid changes responsible for the differences in binding properties between different groups of paired domains. The candidate changes were then tested, individually or in combination, by *in vitro* binding and *in vivo* functional assays.

F. VISUAL PIGMENTS

Vertebrates see light using visual pigments that upon absorption of light, trigger a biochemical cascade using standard G-protein pathways. The pigment itself consists of a protein (an opsin) that binds to a chromophore, generally 11-*cis* retinal, via a protonated imine linkage. A simple model for the protein–chromophore pairing is the protonated imine formed via reaction of a simple amine with retinal. This imine absorbs light maximally at 440 nm (blue-violet) in organic solvents. This absorption maximum is sensitive to the environment, however.

The opsin protein that binds the retinal provides the environment, which it can change by placing different amino acids around the chromophore. Thus, within one homologous family of visual pigments, the absorption maximum can be tuned within the range 360 to 600 nm, that is, from the ultraviolet to red. For example, the four opsins in humans absorb at 414 (violet), 497 (blue-green), 530 (green-yellow), and 560 (yellow) nanometers.

Vertebrate opsins are classified into five subfamilies, known as RH1 (rod opsin, also known as rhodopsin), RH2 (RH1-like, or green, cone opsin), SWS1 (short wavelength–sensitive type 1, or ultraviolet-blue, cone opsin), SWS2 (short wavelength–sensitive type 2, or blue, cone opsin), and M/LWS (middle to long wavelength–sensitive, or red-green, cone opsin). Some species (such as goldfish and chicken) have representatives of each. Humans, with the four opsins mentioned above, lost the SWS2 and RH2 genes but gained an additional member of the M/LWS family.

As with the digestive enzymes, visual pigments stand between an organism and its environment. Thus, the wavelength of absorbance is presumably driven by adaptive pressures. The coelecanth, for example, a fish that lives in the deep sea where ultraviolet light is absent, does not have a visual pigment that responds to ultraviolet light. For scotopic vision (seeing in dim light), for example, pigments need to have a very high quantum efficiency, generate an extremely low level of noise in darkness, and absorb with an absorption maximum at about 500 nm (Menon et al., 2001).

The demands on vision have changed frequently throughout the history of vertebrates. It is therefore expected that an evolutionary narrative will help expand our understanding of the evolving function of vision. Not surprising in this light, three fascinating paleogenetics studies have been done in this area.

1. Rhodopsins from Archaeosaurs: An Ancestor of Modern Alligators and Birds

Chang et al. (2002) began work with visual pigments by seeking to understand how vision might have evolved in the ancestors of modern birds. Their focus was the visual pigment of the primitive archosaur, an ancestor that gave rise to the modern alligator as well as to birds such as the pigeon, chicken, and zebra finch. The sequences of rhodopsins from all of these species was known when their work began, and this clade was rooted by the sequence from the green anole (Figure 23).

A multiple sequence alignment for these rhodopsin genes, together with genes from two dozen other vertebrates, was built using Clustal W. The alignment was then adjusted by hand to ensure that amino acids believed to be structurally important were aligned and to remove gaps within codons. The amino acids at the ends of the rhodopsin genes (encoding the first 21 amino acids and the 25 amino acids following a palmitoylation site) were difficult to align and could not generate an ancestral sequence with a manageable level of ambiguity. These segments were therefore excluded from the alignment, with the exclusion justified by the observation that these segments are believed not to be important in determining the features of photon absorption or activation of transducin. Bovine rhodopsin provided the sequence for these regions of the putative archosaur protein.

A phylogenetic tree was chosen to reflect the accepted relation between the taxa providing the rhodopsin genes, as supported by cladistic and

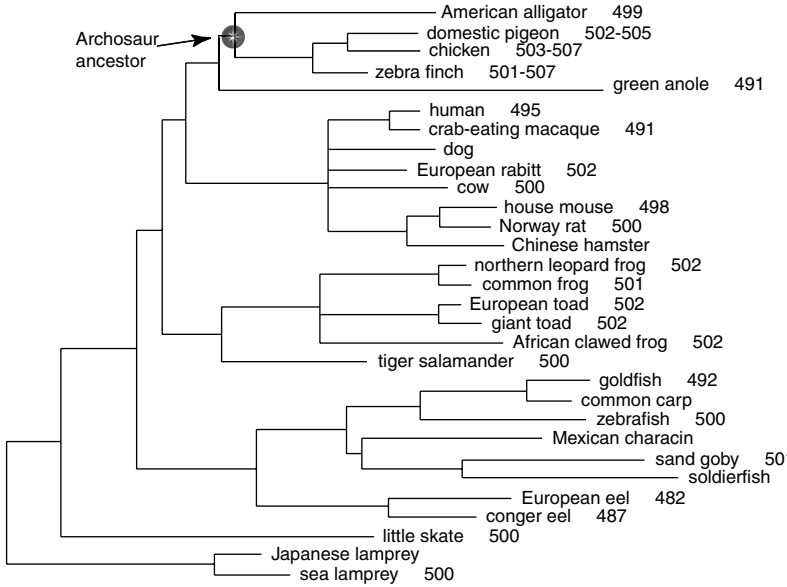


FIGURE 23. Visual pigments from vertebrates analyzed by Chang et al. (2002).

paleontological data. Amino acids in the ancestral archosaur sequence were then computed using the maximum likelihood methods as implemented in PAML (Yang, 1997). This implementation calculated marginal probabilities at each site using an empirical Bayesian approach. Pairwise likelihood ratio tests were also used to select the models with the best fit for the data (Navidi et al., 1991). A gene for the ancestral archosaur rhodopsin was prepared by chemical synthesis and expressed in COS cells. The protein was regenerated with 11-*cis* retinal, purified, and assayed for function in the laboratory.

Especially interesting in this work was its exploration of different models for reconstructing ancestral rhodopsins. Phylogenetic reconstructions using model-based methods may be sensitive to the assumptions underlying the particular model used in the analysis (Cao et al., 1994; Huelsenbeck, 1997) and may also be problematic when reconstructing ancestral states (Chang and Donoghue, 2000).

Chang et al. (2002) used several different models for nucleotide-based, amino acid–based, and codon-based reconstruction strategies, and

FIGURE 24. Fossil of *Confuciusornis sanctus* from the Jurassic–Cretaceous boundary, about 125 Ma. This represents the earliest bird known with a beak, representing a taxa that lived after the archosaur whose visual pigment was resurrected by Chang and her co-workers. (From the S. A. Benner collection; photograph courtesy of Steven Benner.)

compared the results using likelihood ratio tests [see Chang et al. (2002) for details of the comparison]. In particular, for all models tested, eliminating the gamma distribution, which accounts for different rates of change at different sites, resulted in a significantly worse fit to the data.

For the three best-fitting models from the three strategies, reconstructions of the ancestral archosaur rhodopsin were in agreement at all but three sites (213, 217, and 218). At these sites, two of the three models agreed with each other, permitting a ''majority rule'' principle to be applied in inferring the ancestral residue at these sites. Given that these sites lie in a helix facing the lipid bilayer, the behavior of the pigment was not expected to be influenced by this ambiguity. To demonstrate this, Chang et al. synthesized alternative candidate ancestral rhodopsin variants that contained the various amino acids at the three sites.

The ancestral rhodopsin had an absorption maximum in the visible at 508 nm. This maximum is at a longer wavelength than the rhodopsin from most mammals and fish; it lies within the higher end of the range of absorption maxima in reptiles and birds. Chang et al. then demonstrated that the ancestral rhodopsin was able to activate the G-protein transducin.

Rhodopsin, which is found in rods, is essential for vision at low light intensities. The paleogenetic evidence for an active rhodopsin in the archosaur might therefore also be interpreted as evidence that the archaosaur could see in dim light, and therefore may have been nocturnal.

2. *History of Short Wavelength–Sensitive Type 1 Visual Pigments*

The detection of light in the violet and near ultraviolet is key to the survival strategy of many vertebrates, who may detect light in these regions while foraging, selecting a mate, and communicating. The extent to which ultraviolet vision is distributed across the vertebrate tree (even though it is lacking in humans) suggests that perhaps the ancestral vertebrate also had ultraviolet vision. Inferences from such an analysis are notoriously insecure, of course, and even more so given the ability of opsin evolution to generate different absorbance spectra.

To understand better the evolution of these pigments, Shi and Yokoyama (2003) inferred the sequences of various ancestral SWS1 pigments throughout the tree interconnecting a variety of mammals (including rodents, artiodactyls, and primates), birds, reptiles, and fish (Figure 25). This tree topology was based on the amino acid sequences of SWS1 pigments as well as some DNA–DNA hybridization data, and represented sequences as old as 400 million years.

The sequences of ancestral SWS1 pigments were then inferred at seven nodes in the tree using likelihood-based Bayesian methods implemented in PAML (Yang et al., 1995; Yang, 1997) and resurrected. Various hybrid pigments were also designed and constructed by swapping different SWS1 cDNAs.

As with the experiments by Chang et al. (2002), the N- and C-terminal peptide segments were examined to show that they did not influence the absorption of the pigment. Thus, ambiguity in the ancestral sequences within these regions was therefore ignored. Potentially problematic ambiguity in the ancestral sequences was encountered between 1 and 11 sites (depending on the ancestor). This was revealed by posterior

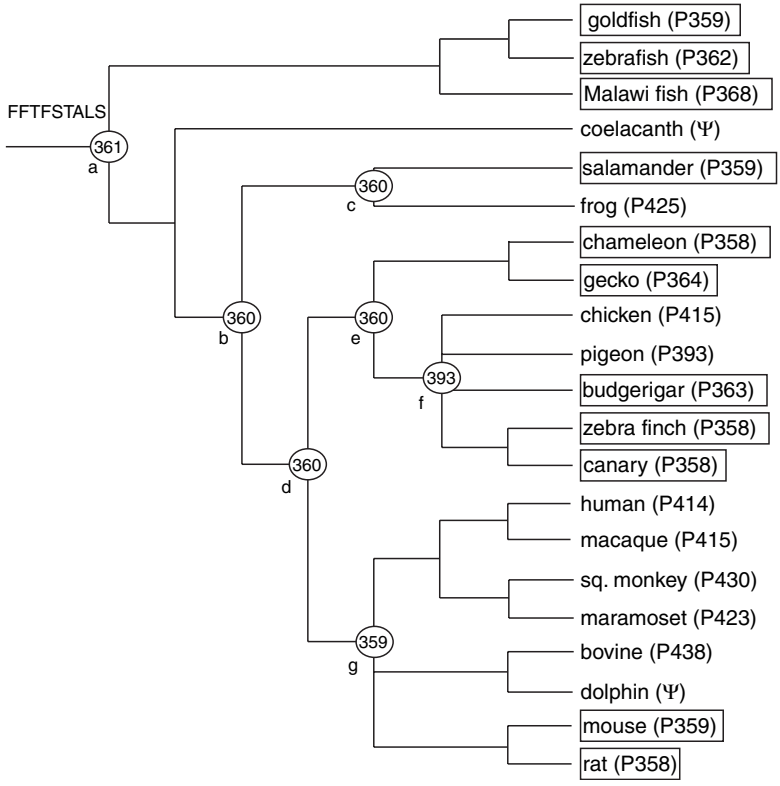


FIGURE 25. Evolutionary tree of 21 vertebrate SWS1 pigments. The tree was inferred from both the sequence data and DNA–DNA hybridization data. The seven ancestral sequences that were resurrected (circled) were inferred by a likelihood-based Bayesian method from Yang et al. (1995). FFTFSTALS refers to the amino acids at the critical sites 46, 49, 52, 86, 90, 93, 114, 116, and 118 for the ancestral pigment. The ultraviolet pigments are boxed. The numbers after P and those at nodes *a* to *g* refer to λ_{max} values. The numbers next to nodal arrows indicate the total numbers of amino acid changes introduced in constructing ancestral pigments.

probabilities of less than 0.9, and differing posterior probabilities depending on whether the JTT or Dayhoff model for amino acid replacement was chosen. This represents another illustration of the impact of the model when inferring ancestral amino acids. This ambiguity was managed through the resurrection of a variety of hybrid pigments. In addition, for the most ancient pigment, if the probability of a residue at a

site was less than 0.9, it was replaced individually by the second-best residue to emerge from the inference.

These studies showed that nearly all of the ambiguity had no impact on the absorption maximum of the candidate ancestral sequences. This type of analysis assumes, of course, that the effects of amino acid replacements at individual sites are independent of each other. Although such an assumption is always open for dispute (see above), it can be evaluated by sequencing more opsin genes to better articulate the tree. This will undoubtedly be done in the future.

Shi and Yokoyama (2003) also considered whether alternative tree topologies might compromise biological inferences based on the absorption maxima of candidate ancestral proteins. The three avian sequence groups are represented in Figure 25 as branching as a "starburst" from ancestor *f*. Alternative tree topologies at this node influenced the oldest opsin sequence (represented by *a* in Figure 25) at three of the 20 variant sites, exchanging the order of likelihood of the two most likely residues. For ancestor *f* (which has the high absorbance maximum), the probability of the preferred amino acid increased. This included an increase in the probability of Ala at site 118, a site where ambiguity did have an impact on the absorbance maximum. For the remainder of the sites, the changes in the reconstructions created by changing the tree topology did not have an impact on the phenotype of the ancestors that generates the biological interpretation.

Two biological interpretations were drawn. First, most vertebrate ancestors could see in the ultraviolet. As the numbers on the trees show, most resurrected ancestral pigments were sensitive to ultraviolet light (absorbance maxima at 360 ± 1 nm). In the emergence of the last common ancestor of the avian sequences, however, the absorbance maximum evolved from 360 nm to 393 nm. Further, in a variety of modern vertebrates, evolution to longer wavelength occurred.

The most curious aspect of the resurrected history of SWS1 is that after evolving from an ultraviolet-sensitive to a violet-sensitive pigment in the last common ancestor of the birds, it regained ultraviolet sensitivity (homoplasy) in some of the lineages leading to modern birds, including in the canary, the budgerigar, and the zebra finch. This behavior provides an example showing that the behaviors of ancestral proteins are not necessarily the behaviors expected by averaging the behaviors of the descendents.

The SWS1 pigment narrative also provides another example illustrating how resurrections lead to the discovery of specific amino acids that have

particular impact. Thus, just as the ribonuclease resurrection identified site 38 as being important to substrate specificity, comparison of ancestral pigments *e* and *f* led to the discovery of previously identified sites for tuning the spectrum of SWS1 that were not among the eight known sites (46, 49, 52, 86, 90, 93, 114, and 118). Three of these sites are suffering replacement along this branch (49, 86, and 118), but replacements at these sites proved to be insufficient to account for evolution of the absorbance spectrum. To account for the rest of the shift in the absorbance maximum, a quadruple change including an L116V replacement was required. Thus, site 116 was discovered as a new site involved in the tuning of the spectrum of SWS1.

To explain the ultraviolet sensitivity in zebra finch, a cysteine at position 90 was the focus of classical studies that focus on leaf–leaf comparisons (Wilkie et al., 2000; Yokoyama et al., 2000). When S90C was introduced into a candidate ancestral pigment *e*, the absorbance of the variant pigment dropped to 360 nm. Thus, the S90C change by itself is sufficient to produce ultraviolet sensitivity. However, in the lineage leading from ancestor *e* to the last common ancestor of finch and canary, S86C replacement also occurred. Individually, this also shifts the absorbance maximum down to 360 nm. Changing S86C and S90C together does not result in an additive shift; these two changes *also* create a pigment having an absorbance at 360 nm. Thus, both the S86C and S90C replacements, either separately or jointly, decrease the absorbance by about 30 nm, providing a textbook example of nonadditivity of phenotype-sequence relationships.

The SWS1 case illustrates the use of paleobiochemistry to manage a family that has seen a relatively large amount of sequence divergence, homoplasy, and functional adaptation. Approximately 43 amino acid replacements have occurred in the history at the nine sites now believed to be important in tuning the absorption wavelength of the pigment. Given 21 sequences and a reasonable effort to study hybrid and alternative ancestral candidates, the associated ambiguity can be managed and interesting information can be extracted.

Shi and Yokoyama then extended the analysis to ask planetary biology questions. Clearly, the lifestyles and the need for ultraviolet vision changed dramatically during the divergent evolution of vertebrates. Indeed, lifestyle, including exposure to ultraviolet light, is different within vertebrate classes and individual orders. In cases such as the coelacanth and dolphin, where ultraviolet light is assumed to be absent in deep ocean water, the SWS1 gene appears to be nonfunctional.

Shi and Yokoyama also noted that ultraviolet light can damage the retina. Therefore, yellow pigments in the lenses or corneas of many species, including humans, filter out ultraviolet light to prevent it from reaching the retina. Therefore, especially in animals exposed to bright sunlight, the switch from ultraviolet vision to violet vision might be driven adaptively. Other factors might include the use of ultraviolet vision in migratory birds, the detection of rodent tracks, or the selection of food by larval fishes. The planetary biology of visual pigments has only begun to be explored. As it is explored using paleobiochemistry, more and more of these cases will advance from ''just so'' stories to serious scientific narratives that join biochemistry to the cellular phenotype, and from there to the ecosystem and the changing planetary environment.

3. Green Opsin from Fish

The question of ecological function in the visual pigments has recently been examined further by Chinen et al. (2005a). These authors analyzed the difference between visual pigments in goldfish and the zebrafish. Zebrafish have four paralogous green (RH2) opsin genes, designated RH2-1, RH2-2, RH2-3, and RH2-4. These have different absorption maxima when reconstituted with 11-*cis* retinal at 467, 476, 488, and 505 nm. Goldfish have two pigments that diverge somewhere in the time when the four zebrafish pigments were diverging (Figure 26). Both are diurnal freshwater species belonging to the same family, the Cyprinidae.

To understand the diversity of paralogs and their different numbers in these two fish, Chinen et al. (2005a) reconstructed the amino acid sequences of ancestral zebrafish RH2 opsins by likelihood-based Bayesian statistics. They then resurrected the ancestral pigments and measured their photophysical properties.

The pigment ancestral to the four zebrafish RH2 pigments (termed A1) (see Figure 26) and the pigment ancestral to RH2-3 and RH2-4 (A3) both absorbed maximally at 506 nm. In contrast, the pigment ancestral to RH2-1 and RH2-2 (A2) absorbed maximally at 474 nm. This indicates that the RH2-3 and RH2-4 modern pigments display the ancestral photophysics, whereas the RH2-1 and RH2-2 pigments have the derived phenotype, with the derived phenotype emerging along branch A (Figure 26).

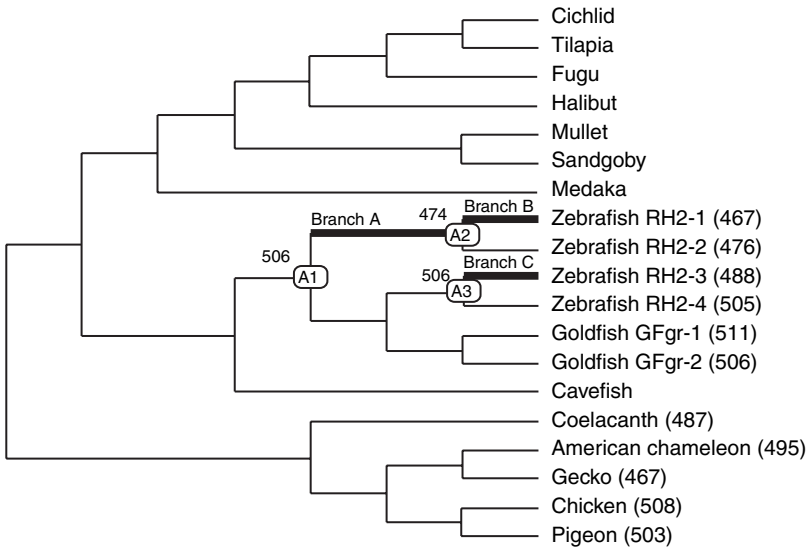


FIGURE 26. Phylogenetic tree of the vertebrate RH2 opsins: *In vitro* absorption maximum values (in nm) are indicated in parentheses. The values of the ancestral opsins at nodes A1 (ancestor 1), A2 (ancestor 2), and A3 (ancestor 3), constructed using the JTT model are also indicated. The branches A, B, and C denote the branches A1–A2, A2–zebrafish RH2-1, and A3–zebrafish RH2-3, respectively, and are emphasized with thick lines. The absorbance shifts that occurred along the branches A, B, and C are indicated. (Adapted from Chinen et al., 2005a.)

Hybrid pigments were then constructed to show that a large contribution of the spectral tuning (about 15 nm) arose by a substitution of a glutamate by a glutamine at site 122. The remaining spectral differences appeared to arise from complex interactive effects of a number of amino acid replacements, each of which has only a minor impact (1 to 3 nm). Thus, the four zebrafish RH2 pigments cover nearly an entire range of absorbance found in vertebrate RH2 pigments.

4. Blue Opsins

Chinen et al. (2005b) then continued these studies by examining the blue opsins. In a separate study, the sequence of the ancestral SWS2 pigment of

the two species was inferred by applying likelihood-based Bayesian statistics. The ancestral protein was then resurrected by site-directed mutagenesis on a cloned gene. The reconstituted ancestral photopigment had a λ_{max} of 430 nm, indicating that zebrafish and goldfish achieved short-wavelength (14-nm) and long-wavelength (13-nm) spectral shifts, respectively, from the ancestor. Unexpectedly, the S94A mutation resulted in only a 3-nm spectral shift when introduced into the goldfish SWS2 pigment. Nearly half of the long-wavelength shift toward the goldfish pigment was achieved instead by T116L (6 nm). The S295C mutation toward zebrafish SWS2 contributed to creating a ridge of absorbance around 400 nm and broadening its spectral sensitivity in the short-wavelength direction. These results indicate that the evolutionary engineering approach is very effective in deciphering the process of functional divergence of visual pigments. Among the amino acid differences between the two pigments, only one (alanine in zebrafish and serine in goldfish at residue 94) was previously known to cause a difference in absorption spectrum (14-nm λ_{max} shift in newt SWS2).

Zebrafish and goldfish are both diurnal freshwater fish species belonging to the Cyprinidae family. Their ecological surroundings differ considerably with respect to their visual needs, however. Zebrafish are surface swimmers in conditions of broad and shortwave-dominated background spectra. Goldfish are generalized swimmers whose light environment extends to a depth of elevated short-wavelength absorbance with turbidity. The peak absorption spectrum (λ_{max}) of the zebrafish blue (SWS2) visual pigment is consistently shifted to a short wavelength (416 nm) compared with that of the goldfish SWS2 (443 nm).

5. Planetary Biology of the Opsins

In due course we will be able to tie the molecular and biophysical history of these visual pigments to historical changes in the environment of the organisms that carried them. The formation of paralogous visual pigments in zebrafish (*Danio rerio*) occurred well after the divergence of the medaka (*Oryzias latipes*) from the *Otophysi*, which includes the Mexican cavefish (*Astyanax mexicanus*), the goldfish (*Carassius auratus*), and the zebrafish (*Danio rerio*). Further, they appear to have diverged after the divergence of the Otophysi to

separate the Cypriniphysi (which includes the zebrafish and goldfish) from the Characiphysi (which includes the Mexican cavefish). Rather, the paralogs arose via a duplication that occurred near the time of divergence of the goldfish (the family Cyprininae) from the zebrafish (the family Rasborinae).

Limits have been put on the dates of divergence of various families (Kruiswijk et al., 2002). For example, the family Rasborinae (which includes the zebrafish) is estimated to have diverged 50 Ma from the lineage leading to the family Cyprininae (Ohno et al., 1967; Cavender, 1991; Stroband et al., 1995; Dixon et al., 1996). These, in turn, are events that occurred in a time when the fossil and geological records are strong.

These dates are obtained primarily using molecular clocks. Many of these fish are known from the lakes in the Rift Valley of central Africa, however, the same area where primate evolution was to generate *Homo sapiens*. The area is associated with multiple lava flows and ash beds, including a lava flow dated 5 Ma that blocked Lake Tana (in Ethiopia), which is home to *Barbus intermedius*, believed to branch from the lineage leading to carp and goldfish at 30 Ma. A coherent narrative will combine the geological and paleontological records to explain what was happening in the environment of these fish that caused them to change their lifestyle, which in turn made the changes in the photophysical properties of the visual pigments adaptive.

G. AT WHAT TEMPERATURE DID EARLY BACTERIA LIVE?

By the end of the last decade, the most ancient paleomolecular resurrections had traveled back in time only about 200 to 300 million years. This has left untouched many of the most widely discussed questions about the nature of early life on Earth. One of these relates to the role of thermophily in the early history of terrean life.

The issue has been confused somewhat by contradicting analytical strategies. Thus, various authors, observing that thermophilic organisms are placed at deep branches of a tree, suggested that the last common ancestor of all organisms should be a thermophile. Others inferred the G + C content of an ancestral ribosomal RNA gene, and noted that this was inconsistent with the ancestor being a hyperthermophile (Galtier et al., 1999). Various models for environments deep in the Archaean

suggest that the Earth was hot (Knauth, 2005) and covered with snow (Runnegar, 2000). Other models suggest that early bacteria may have been thermophiles, or possibly extreme thermophiles. Arguments based on indirect evidence, such as the lengths of branches of various trees (Woese, 1987), the G + C content of reconstructed ancestral ribosomal RNA (Galtier et al., 1999), and the distribution of thermophily in contemporary taxa (Hugenholtz et al., 1998), have generated contradictory inferences.

1. Elongation Factors

Gaucher et al. (2003) conjectured that an experiment in paleogenetics might shed light on this question. If the sequences of ancestral proteins from bacteria that lived in the Archaean could be resurrected and their properties over a range of temperatures could be studied, we might be able to obtain direct evidence for the temperature(s) at which the ancestral bacteria lived.

Elongation factor Tu (from Bacteria) and elongation factor 1A (from Archaea and Eukarya) proved to be suitable for such a study. EFs are G-proteins that present charged aminoacyl-tRNAs to the ribosome during translation. Because of their relatively slow rates of sequence divergence, most character states of ancient EF sequences can be reconstructed robustly for proteins from bacteria deep in the eubacterial tree. Further, the optimal thermal stabilities of EFs correlate with the optimal growth temperature of the host organism. Thus, EFs from mesophiles, thermophiles, and hyperthermophiles, defined as organisms that grow at 20 to 40, 40 to 80, and $>80^{\circ}$C, respectively, and represented by species of *Escherichia*, *Thermus*, and *Thermotoga*, have temperature optima in their respective ranges (Arai et al., 1972; Nock et al., 1995; Sanangelantoni et al., 1996). This is consistent with a previous study based on a large set of proteins in which a correlation coefficient of 0.91 was calculated between environmental temperatures of the host organisms and protein melting temperatures (Gromiha et al., 1999).

To infer the sequences of EFs deep within the bacterial lineage, Gaucher et al. collected amino acid sequences of 50 EF-Tu's from various bacterial lineages. Because saturation at silent sites in the DNA sequence had occurred, amino acid sequences were used to build a multiple sequence alignment and candidate trees. The differences in the rates of amino acid

replacement at different sites in the sequence was captured using a gamma distribution (Gaucher et al., 2001).

Ambiguity was encountered immediately in constructing a tree to interrelate these proteins. Thus, two trees were used. The first was constructed from EF-Tu sequences alone using a combination of phylogenetic tools. The second was constructed from the literature, which contains various views of bacterial phylogeny. These trees were largely congruent. Where they differed, a tree was extracted that captured those differences (Hugenholtz et al., 1998) (Figure 27).

Candidate ancestral sequences were then reconstructed at the most basal node of the bacterial domain in each tree, using archaeal and eukaryotic EF sequences as outgroups. Marginal reconstructions, opposed to joint reconstructions, were calculated to compare probabilities of multiple character states at a single interior node and selecting the character with the highest posterior probability (Yang et al., 1995). The *most probabilistic ancestral sequence* (MPAS) was then reconstructed by accepting at each site the amino acid with the highest posterior probability.

The most probabilistic ancestral sequences for the two trees were found to be surprisingly similar to sequences from modern *Aquifex*; only four to six amino acid replacements out of about 400 residues were inferred to have occurred from the most recent common ancestor of bacteria to the modern *Aquifex*. The placement of the branch leading to Aquificaceae at the base of the tree appeared, however, to be due to long-branch attraction (Brochier and Philippe, 2002; Cavalier-Smith, 2002). To test this, the 23 sites that displayed no variation within the outgroup subfamily, the Aquificaceae subfamily, and the subfamily containing all other bacteria, but not conserved between the subfamilies, were removed from the analysis. The resulting analysis no longer places Aquificaceae at the base of the bacterial lineage. This is consistent with the working of long-branch attraction. To eliminate bias due to this artifact, ancestral sequences were recalculated from a data set that excluded *Aquifex* sequences.

Figure 27 shows the two topologies used to reconstruct ancestral sequences at the node representing the hypothetical organism lying at the stem of the bacterial tree. The number of sequences in the outgroup, 3 to 20, did not affect the amino acid reconstructions at these nodes, ML-stem (maximum likelihood stem bacteria) and Alt-stem (alternative stem bacteria). The ancestral sequence at the node representing the most recent common ancestor of only mesophilic bacterial lineages was reconstructed

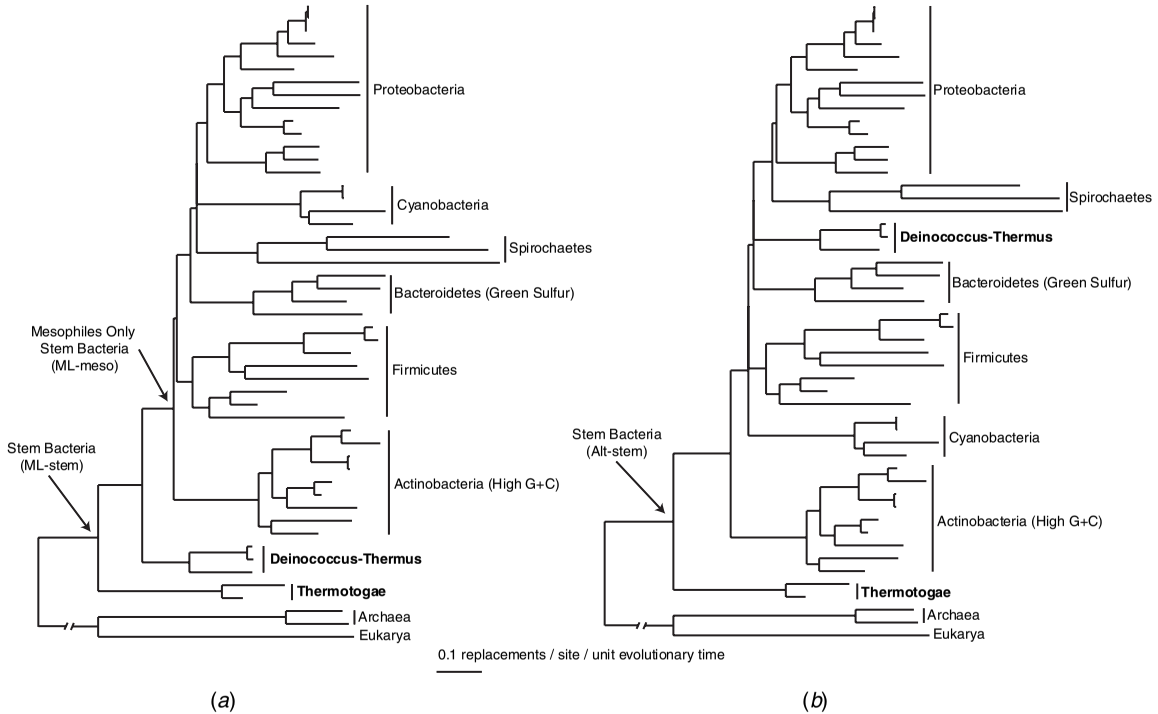


FIGURE 27. The two unrooted universal trees used to reconstruct ancestral bacterial sequences: (*a*) maximum likelihood topology used to reconstruct the stem bacteria (ML-stem), or most recent common ancestor of bacteria, and the ancestral sequence for mesophilic lineages only (ML-meso); (*b*) alternative topology used to reconstruct the stem bacteria (Alt-stem). Archaea and Eukarya serve as outgroups for Bacteria, and thus provide a point at the base of the bacterial subtree from which ancient sequences can be inferred. Thermophilic lineages are highlighted in bold.

TABLE 5
Percent Sequence Identity Between Ancestral and Modern Proteins[a]

	Alt-Stem	ML-Meso	T.m.	T.a.	E.c.	G.s.
ML-stem	93	87	85	84	79	84
Alt-stem		89	84	80	80	83
ML-meso			74	78	82	84

[a]T.m., *Thermotoga maritima*; T.a., *Thermus aquaticus*; E.c., *Escherichia coli*; G.s., *Geobacillus stearothermophilus*.

and named the ML-meso (maximum likelihood mesophiles only). This node captures one feature of models that have concluded that the last common ancestor of Bacteria was mesophilic (Brochier and Philippe, 2002). In all, these reconstructed ancestral sequences did not appear to be influenced by long-branch attraction or nonhomogeneous modes of molecular evolution, such as changes in the mutability of individual sites in different branches of the bacterial subtree (Gaucher et al., 2001). Table 5 shows the sequence identity relating the reconstructed putative ancestral sequences and their descendants. ML-stem and Alt-stem are most similar to the sequences of EFs from *Thermoanaerobacter tengcongensis* (a thermophile) and *Thermotoga maritima* (a hyperthermophile), respectively, and differ from each other by 28 amino acids. ML-meso is most similar to the sequence of EF from *Neisseria meningitides* (a mesophile).

If we assume that similarity in sequence implies similarity in thermostability, it might have been predicted that the stem bacterium was thermophilic or hyperthermophilic, while the ancestral node constructed without considering thermophiles was a mesophile. To test these predictions based on this (undersubstantiated) assumption, genes encoding the ancestral sequences were synthesized, expressed in an *E. coli* host, and purified. The thermostabilities of these ancestral EFs, and three representative EFs from contemporary organisms, were then assessed by measuring the ability of each to bind GDP across a range of temperatures.

Each resurrected protein behaved similarly. Both ML-stem and Alt-stem bound GDP with a temperature profile similar to that of the thermophilic EF from modern *Thermus aquaticus*, with optimal binding at about 65°C. Although the sequence similarity was higher between Alt-stem and the modern hyperthermophilic *T. maritima*, the temperature profile of Alt-stem was not similar to that from *maritima*, which is maximally active up to at 85°C. The observation that the amino acid sequences of ML-stem and

Alt-stem shared only 93% identity, but display the same thermostability profiles, suggest that inferences of this ancestral property are robust with respect to both varying topologies and ancestral character state predictions. This suggests, based on these given evolutionary models, that the paleoenvironment of ancient bacterium was approximately 65°C.

Inferences were then drawn from a resurrected elongation factor whose sequence was reconstructed from the last common ancestral sequences of contemporary organisms that for the most part, grow optimally at mesophilic temperatures. The temperature profile of the ancestral protein, which displayed a maximum at 55°C, suggests that the ancestor of modern mesophiles lived at a higher temperature than any of its descendants (Figure 28). This result shows that the behavior of an ancestor need not be an average of the behaviors of its descendants.

The observation that a tree-based ancestral sequence reconstructions can give results different from consensus sequence reconstructions may be general (Gaschen et al., 2002). It underscores a fact, well known in protein chemistry, that physical behavior in a protein is not a linear sum, or even a simple function, of the behavior of its parts. This, in turn, implies that an experiment in paleobiochemistry can yield information beyond that yielded by analysis of descendent proteins alone.

The resurrection is made still more complicated by the incomplete fossil record of microorganisms (Figure 29). As the sequence, geological, and paleontological records improve, we may expect the overall historical model to improve as well.

2. Isopropylmalate and Isocitrate Dehydrogenases

A similar approach was exploited by the Yamagishi group (Miyazaki et al., 2001; Iwabata et al., 2005) using 3-isopropylmalate and isocitrate dehydrogenases as systems to address questions regarding the origins of archae. Although the authors' methodology raises concern, a summary of their conclusions is valuable. These two classes of enzyme proteins appear to be related by common ancestry. In their first study in 2001, ancestral residues were inferred using the Protpars program in the PHYLIP package. These were introduced into an enzyme of a strain of extreme thermophiles of the genus *Sulfolobus*. Of the seven proteins engineered, five displayed thermal stability higher than that of the modern enzyme. This was interpreted as consistent with the hypothesis that the universal ancestor was a hyperthermophile.

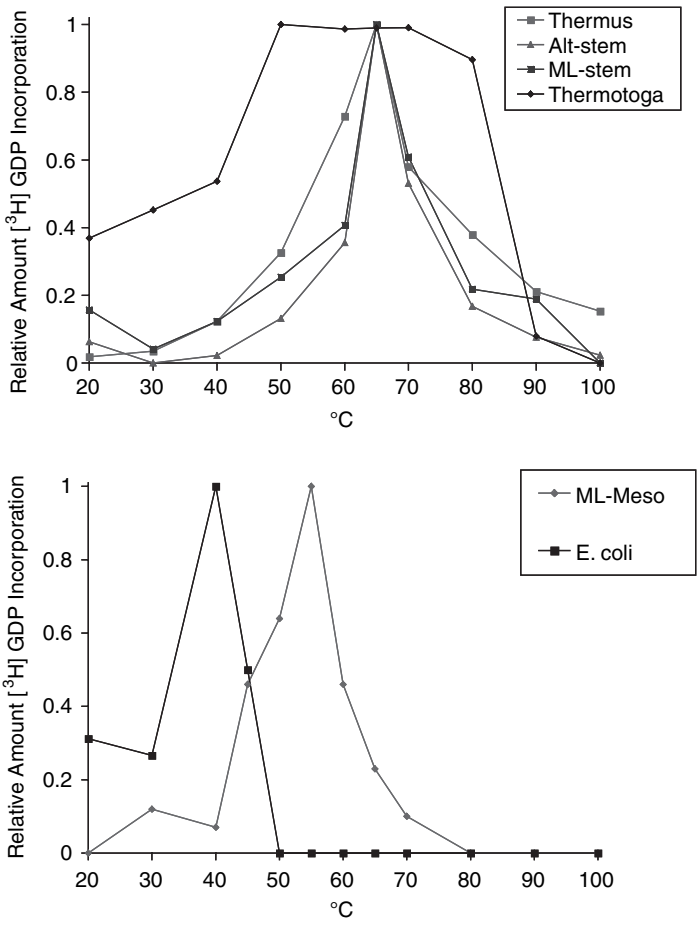
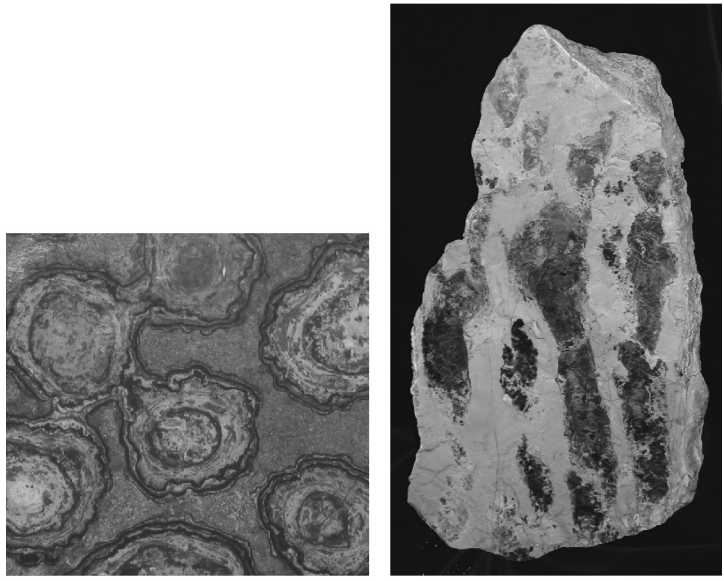


FIGURE 28. GDP-binding assay to test thermostability of ancestral and modern EF proteins. The amount of tritium-labeled GDP bound at zero degrees was subtracted from all other temperature values for a given protein. Shown is the relative amount of GDP bound compared to the amount bound at the optimal temperature for each protein.

The study was recently extended (Iwabata et al., 2005). Here, a total of 18 sequences of isocitrate dehydrogenases and isopropylmalate dehydrogenases were aligned with Clustal X, and well-aligned regions were selected using the program Gblock. Composite trees were constructed using a neighbor-joining heuristic, and a most probable tree was selected.



(*a*) (*b*)

FIGURE 29. Evidence of stromatolites in the geologic record: (*a*) Salt Lake of Uyuni, South America, about 2.7 Ga; (*b*) Dutch Creek dolomite from the Proterozoic Wyloo Group in the Ashburton basin, Western Australia. Age is about 2 Ga. The morphology of the structures and the apparent place of deposition suggest that these organisms lived after the eubacteria whose protein was resurrected by Gaucher et al. (2003) [(*a*) From the E. A. Gaucher collection; photograph courtesy of Eric Gaucher; (*b*) from the S. A. Benner collection; photograph courtesy of Steven Benner.]

The PAML program was used to construct ancestral sequences using a gamma distribution and the WAG amino acid substitution matrix. A parallel analysis permitted these inferences to be compared to those inferred using maximum parsimony. One or two ancestral residues were then introduced into the isocitrate dehydrogenase from *Caldococcus noboribetus*, and the thermostability of the resulting protein assayed to discover variants that had altered thermal stabilities. In the first pass, four sets of mutations representing ancestral states were introduced individually and assayed for their thermostability (variants Y309I/I310L, I321L, A325P/G326S, and A336F). The half-inactivation temperatures were 88.4, 88.9, 91.3, and 74°C, compared to 87.5°C for the wild type. A construct consisting of mutations Y309I/I310L, I321L, and A325P/G326S

was then prepared. This variant had a half-inactivation temperature of 90°C. Curiously, the authors did not report a variant consisting of all ancestral states inferred.

With that said, the results were interpreted as evidence for the hyperthermophilic hypothesis for the *last universal common ancestor* (LUCA). However, it is possible that the last common ancestor of Archaea (tested here) did not inhabit an environment similar to the environment that hosted LUCA. It was also noted, however, that the results were not additive. In some cases, an individual change that increased thermostability relative to the wild-type enzyme did not further enhance the thermal stability of an enzyme already stabilized by one of the putative ancestral states.

3. Conclusions from "Deep Time" Paleogenetics Studies

These studies push the experimental paleogenetics research strategy back in time 2 to 3 billion years, to the most primitive ancestors from which descent can be traced. Accordingly, the ambiguity encountered is substantial, and available sequence data are not sufficient to manage it convincingly. Here, the ambiguities do not depend primarily on the details of the model used to infer ancestral states. Rather, they seem to arise from the uncertainty of the phylogenetic tree joining the protein family members.

The fact that reconstructions can be made at all is therefore noteworthy. Further, if the large-scale sequencing of random bacterial genomes undertaken by Venter et al. (2004) continues, there is good reason to hope that the reconstructions become better. Indeed, the temperature history of eubacteria is already beginning to be defined (Gaucher, personal communication) by studies throughout that kingdom.

The preliminary results (Gaucher et al., 2003) suggest that the temperature environment of more recent bacterial is still higher than the descendents presently living as mesophiles. The work of Lowe, Knauth, and others extracting information from the geological record about the temperature history of Earth will soon be tied to this molecular record using experimental paleomolecular biology.

H. ALCOHOL DEHYDROGENASE: CHANGING ECOSYSTEM IN THE CRETACEOUS

The ultimate goal of molecular paleoscience is to connect the molecular records for all proteins from all organisms in the modern biosphere with the

geological, paleontological, and cosmological records to create a broadly based, coherent narrative for life on Earth. Because much of natural selection is driven by species–species interactions, developing this narrative will require tools that broadly connect genomes from different species as well as interconnect events within a single species. It remains an open question, of course, how much of the record has been lost through extinction, erosion, and poor fossil preservation.

The first paradigms of this broad interconnection are now beginning to emerge. As might be imagined, they feature paleomolecular resurrections. One recently published study concerns the interaction between yeast, fruits, and other forms of life near the age of the dinosaurs. It appears that this was the time when the first yeast developed the metabolic strategy to make and consume ethanol in fleshy fruits.

Modern yeasts living in modern fleshy fruits rapidly convert sugars into bulk ethanol via pyruvate (Figure 30). Pyruvate then loses carbon dioxide to give acetaldehyde, which is reduced by alcohol dehydrogenase 1 (Adh1) to give ethanol, which accumulates. Yeast later consumes the accumulated ethanol, exploiting Adh2 and Adh1 homologs differing by 24 (of 348) amino acids.

Generating ethanol from glucose in the presence of dioxygen, only to then reoxidize the ethanol, is energetically expensive (Figure 30). For each molecule of ethanol converted to acetyl-CoA, a molecule of ATP is used. This ATP would not be ''wasted'' if the pyruvate that is made initially from glucose were delivered directly to the citric acid cycle.

This implies that yeast has a reason, transcending simple energetic efficiency, for rapidly converting available sugar in fruit to give bulk ethanol in the presence of dioxygen. One ''just so'' story to explain this inefficiency holds that yeast, which is relatively resistant to ethanol toxicity, may accumulate ethanol to defend resources in the fruit from competing microorganisms (Boulton et al., 1996). Although the ecology of wine yeasts is certainly more complex than this simple hypothesis implies (Fleet and Heard, 1993), fleshy fruits offer a large reservoir of carbohydrate, and this resource must have value to competing organisms as well as to yeast. For example, humans have exploited the preservative value of ethanol since prehistory (McGovern, 2004).

The timing of Adh expression in *Saccharomyces cerevisiae* and the properties of the proteins expressed are both consistent with this story. The yeast genome encodes two major alcohol dehydrogenases (Adh's) that interconvert ethanol and acetaldehyde (Figure 30) (Wills, 1976). The first

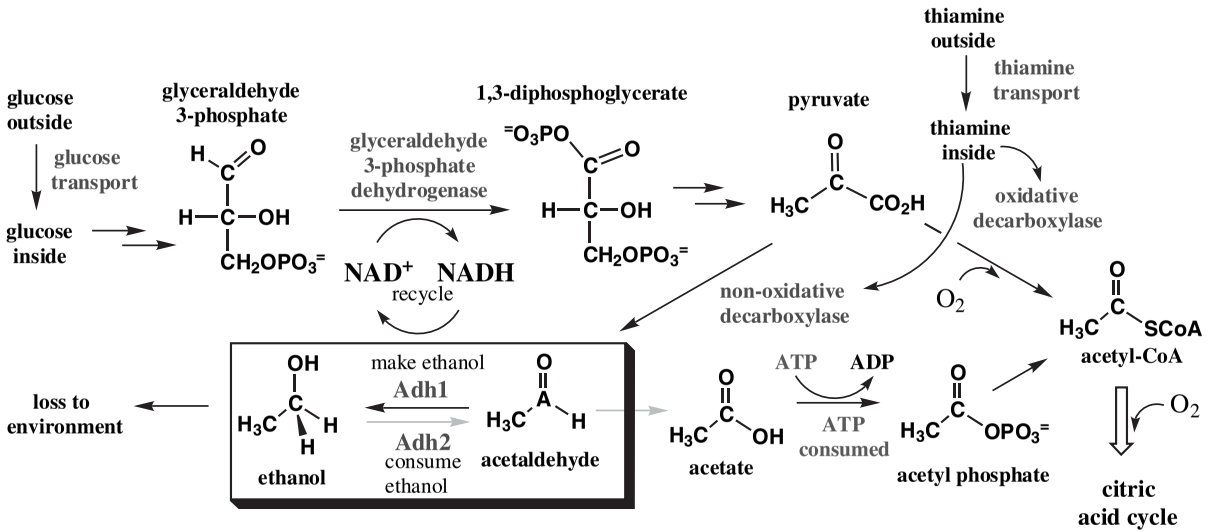


FIGURE 30. Pathway by which yeast makes, accumulates, and then consumes, ethanol. Enzymes in red are associated with gene duplications that, according to the TREx clock, arose nearly contemporaneously. The make–accumulate–consume pathway is boxed. Note that the shunting of the carbon atoms from pyruvate into (and then out of, arrows in blue) ethanol is energy-expensive, consuming a molecule of ATP (green) for every molecule of ethanol generated. This ATP is not consumed if pyruvate is oxidatively decarboxylated directly to give acetyl-CoA to enter the citric acid cycle directly (dashed arrow to the right). If dioxygen is available, the recycling of NADH does not need the acetaldehyde-to-ethanol reduction.

(Adh1) is expressed at high levels constitutively. Its kinetic properties optimize it as a catalyst to make ethanol from acetaldehyde (Fersht, 1977; Ellington and Benner, 1987). In particular, the Michaelis constant (K_M) for ethanol in Adh1 is high (17,000 to 20,000 μM), consistent with ethanol being a product of the reaction. After the sugar concentration drops, the second dehydrogenase (Adh2) is derepressed. This paralog oxidizes ethanol to acetaldehyde with kinetic parameters suited for this role. The K_M value for ethanol for Adh2 is low (600 to 800 μM), consistent with ethanol being its substrate.

Adh1 and Adh2 are homologs (Ellington and Benner, 1987) differing by 24 of 348 amino acids. Their common ancestor, ADH_A, had an unknown role. If ADH_A existed in a yeast that made but did not accumulate ethanol, its physiological role would presumably have been the same as the role of lactate dehydrogenase in mammals during anaerobic glycolysis: to recycle NADH generated by the oxidation of glyceraldehyde-3-phosphate (Figure 30) (Stryer, 1995). Lactate in human muscle is removed by the bloodstream; ethanol would be lost by the yeast to the environment. If so, ADH_A should be optimized for ethanol synthesis, as is modern Adh1. The kinetic behaviors of ADH_A should resemble those of modern Adh1 more than Adh2, with a high K_M value for ethanol.

To add paleobiochemical data to convert this ''just so'' story into a more compelling scientific narrative, a collection of Adh's from yeasts related to *S. cerevisiae* was cloned, sequenced, and added to the existing sequences in the database (Thomson et al., 2005). A maximum likelihood evolutionary tree was then constructed using PAUP*4.0 (Figure 31) (Swofford, 1998). Maximum likelihood sequences for ADH_A were then reconstructed using both codon and amino acid models in PAML (Ynag, 1997). When the posterior probability that a particular amino acid occupied a particular site was greater than 80%, that amino acid was assigned at that site in ADH_A.

When the posterior probability was less than 80% and/or the most probabilistic ancestral state estimated using the codon and amino acid models were not in agreement, the site was considered ambiguous, and alternative ancestral genes were constructed. For example, the posterior probabilities of two amino acids (methionine and arginine) were nearly equal at site 168 in ADH_A, three amino acids (lysine, arginine, and threonine) were plausibly present at site 211, and two (aspartic acid and asparagine) were plausible for site 236. To handle these ambiguities, all 12 (all $2 \times 2 \times 3$ combinations) candidate ADH_A's were resurrected by constructing genes that encoded them, transforming these genes into a

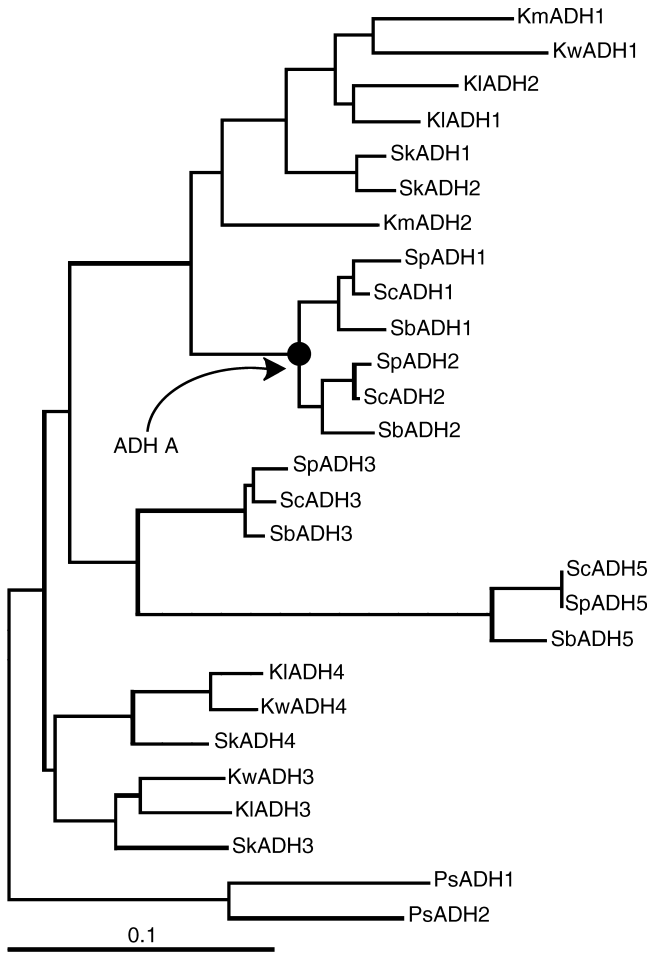


FIGURE 31. Maximum likelihood trees interrelating sequences determined in this work with sequences in the publicly available database. Shown is the tree with the best ML score using the following parameters estimated from the data: Substitutions $A \leftrightarrow C$, $A \leftrightarrow T$, $C \leftrightarrow G$, $G \leftrightarrow T = 1.00$, $A \leftrightarrow G = 2.92$, and $C \leftrightarrow T = 5.89$, empirical base frequencies, and proportion of invariable sites and the shape parameter of the gamma distribution set to 0.33 and 1.31, respectively.

strain of *S. cerevisiae* (Thomson, 2002) from which both Adh1 and Adh2 had been deleted, and expressing them from the Adh1 promoter. All of the ancestral sequences could rescue the double-deletion phenotype.

Table 6 collects kinetic data from the candidate ancestral ADH_A's. To assess the quality of the data, Haldane values [$= K_{eq} = V_f K_{iq} K_p / V_r K_{ia} K_b$ (Segel, 1975), where V_f and V_r are forward and reverse maximal velocities, K_{ia} and K_{iq} are disassociation constants for NAD^+ and NADH, and K_b and K_p are Michaelis constants for ethanol and acetaldehyde, respectively]

TABLE 6
Kinetic Properties of Adh1, Adh2, and Candidate Ancestral ADH_A's

Sample[a]	K_M/NADH (μM)	K_M/EtOH (μM)	K_M/NAD^+ (μM)	K_M/acet. (μM)
Adh1	20,060	218	1,492	164
MKD	17,280	511	1,019	144
MKN	13,750	814	1,067	1106
MRD	11,590	734	1,265	287
MRN	10,960	554	1,163	894
MTD	10,740	467	959	190
MTN	N/A	N/A	N/A	N/A
RKD	8,497	449	1,066	142
RKN	7,238	407	1,085	735
RRD	7,784	400	1,074	203
RRN	8,403	172	1,156	1,142
RTD	6,639	254	1,083	316
RTN	7,757	564	1,158	477
Adh1[b]	24,000	240	3,400	140
Adh1[c]	17,000	170	1,100	110
Adh2[b]	2,700	140	45	28
Adh2[c]	810	110	90	50
Adh3[c]	12,000	240	440	70
Adh1[c] (*S. pombe*)	14,000	160	1,600	100
Adh1(M270L)[c]	19,000	630	1,000	80
KlP20369[d]	27,000	2,800	1,200	110
KlX64397[d]	23,000	2,200	1,700	180
KlX62766[d]	2,570	310	100	20
KlX62767[d]	1,560	200	3,100	30

[a]The three letters designate the amino acids at positions 168, 211, and 236; thus, MKD = Met168, Lys211, Asp236. The remaining residues were the same as in Adh1, except for the following changes (using sequence numbering of Adh1 from *S. cerevisiae*): Asn15, Pro30, Thr58, Ala74, Glu147, Leu213, Ile232, Cys259, Val265, Leu270, Ser277, Asn324. N/A, not applicable.
[b]From Ellington, 1987.
[c]From Kellis, 2004.
[d]From Bozzi, 1997; Kl, *Kluyveromyces lactis*.

were calculated from the experimental data. These reproduced the literature equilibrium constant for the reaction to within a factor of 2. One variant, termed MTN, had very low catalytic activity in both directions. This suggested that this particular candidate ancestor was not present in the ancient yeast.

Significant to the hypothesis, the kinetic properties of the remaining candidate ancestral ADH_As resembled those of Adh1 more than Adh2 (Table 6). From this it was inferred that the ancestral yeast did not have an Adh specialized for the consumption of ethanol, such as that of modern Adh2, but rather had an Adh specialized for making ethanol, like modern Adh1. This, in turn, suggests that the ancestral yeast prior to the time of the duplication did not consume ethanol. This implies that the ancestral yeast also did not make and accumulate ethanol under aerobic conditions for future consumption, and that the make–accumulate–consume strategy emerged after Adh1 and Adh2 diverged. These interpretations are robust with respect to the ambiguities in the reconstructions.

Several details are noteworthy. For modern Adh1, K_M values reported for ethanol range from 17,000 to 24,000 μM, from 170 to 240 μM for NAD^+, from 1100 to 3400 μM for acetaldehyde, and from 110 to 140 μM for NADH (Ganzhorn et al., 1987). These comparisons, together with the Haldane analysis, provide a view of the experimental error in the kinetic parameters reported here. Thus, the interpretations are based on differences well outside experimental error.

Further, when paralogs are generated by duplication, the duplicate acquiring the new functional role is often believed to evolve more rapidly than the one retaining the primitive role (Kellis et al., 2004). If this were generally true, one might identify the functionally innovative duplicate by a bioinformatics analysis. Although this may be true for many genes, chemical principles do not obligate this outcome, and it is not manifest with these Adh paralogs. Here, the rate of evolution is not markedly faster in the lineage leading to Adh2 (having the derived function) than in the lineage leading to Adh1 (having the primitive function). Thus, a paleobiochemistry experiment was necessary to assign the primitive behavior.

Further, the Haldane ratio relates various kinetic parameters (k_{cat}, K_M, K_{diss}) that can change via a changing amino acid sequence to the overall equilibrium constant, which the enzyme (being a catalyst) cannot change. Thus, if a lower K_M value for ethanol is selected, other terms in the Haldane ratio must change to maintain the ratio. This is

observed in data for the ancestral proteins prepared here and the natural enzymes.

The assignment of a primitive function to ADH_A raises a broader historical question: Did the Adh1/Adh2 duplication, and the accumulate–consume strategy that it presumably enabled, become fixed in response to a particular selective pressure? Connecting molecular change to organismic fitness is always difficult (Kreitman and Akashi, 1995), but is necessary if reductionist biology is to move through systems biology to a planetary biology that answers ''why?'' as well as ''how?'' questions (Benner et al., 2002).

The emergence of a make–accumulate–consume strategy may have been driven by the domestication of yeast by humans selecting for yeasts that accumulate ethanol. Alternatively, the strategy might have been driven by the emergence of fleshy fruits that offered a resource worth defending using ethanol accumulation. They might distinguished by estimating a date when the Adh1/2 duplication occurred. Even with large errors in the estimate, a distinction should be possible, as human domestication occurred in the past million years, while fleshy fruits arose in the Cretaceous, after the first angiosperms appeared in the fossil record 125 Ma (Sun, 2002), but before the extinction of the dinosaurs 65 Ma (Collinson and Hooker, 1991; Fernandez-Espinar et al., 2003).

The topology of the evolutionary tree in Figure 31 suggests that the Adh1/2 duplication occurred before the divergence of the *sensu strictu* species of *Saccharomyces* (Fernandez-Espinar et al., 2003), but after the divergence of *Saccharomyces* and *Kluyveromyces*. The date of divergence of *Saccharomyces* and *Kluyveromyces* is unknown, but might be estimated to have occurred 80 ± 15 Ma (Berbee and Taylor, 1993). This date is consistent with a transition redundant exchange (TREx) clock (Benner, 2003), which exploits the fractional identity (f_2) of silent sites in conserved twofold redundant codon systems to estimate the time since the divergence of two genes. Between pairs of presumed orthologs from *Saccharomyces* and *Kluyveromyces*, f_2 is typically 0.82, not much lower than the f_2 value (0.85) separating Adh1 and Adh2 (Benner et al., 2002), but much lower than paralog pairs within the *Saccharomyces* genome that appear to have arisen by more recent duplication (about 0.98) (Lynch and Conery, 2000).

Interestingly, Adh1 and Adh2 are not the only pair of paralogs where $0.80 < f_2 < 0.86$ (Benner et al., 2002). Analysis of about 350 pairs of paralogs contained in the yeast genome (considering pairs that shared at least 100 silent sites and diverged by less than 120 accepted point

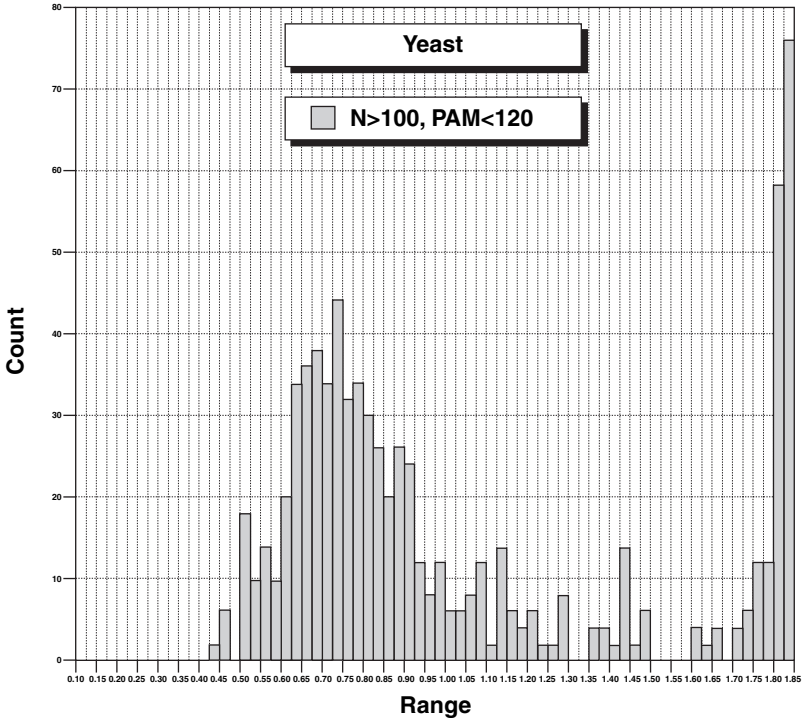


FIGURE 32. Histogram showing all of the pairs of paralogs in the *Saccharomyces cerevisiae* genome, dated using the TREx tool.

replacements per 100 sites) identified 15 pairs having $0.80 < f_2 < 0.86$ (Figure 32). These represent eight duplications that occurred near the time of the Adh1 and Adh2 duplication, if f_2 values are assumed to support a clock.

These duplications are not randomly distributed within the yeast genome. Rather, six of the eight duplications involve proteins that participate in the conversion of glucose to ethanol (Table 7). Further, the enzymes arising from the duplicates are those that appear, from expression analysis, to control flux from hexose to ethanol (Schaaff et al., 1989; Pretorius, 2000). These include proteins that import glucose, pyruvate decarboxylases that generates the acetaldehyde from pyruvate, the transporter that imports thiamine for these decarboxylases, and the Adh's

TABLE 7
Duplication in the *Saccharomyces cerevisiae* Genome Where $0.80 < f_2 < 0.86$

SGD Name	gi Number	Trivial Name	Annotation and Comments
Inosine-5′-Monophosphate Dehydrogenase Family (3 paralogs, 3 pairs, 2 duplications)[a]			
$f_2 = 0.803$[b]	Pair associated with Wolfe duplication blocks 1 and 44		
YAR073W	gi | 456156	IMD1	Nonfunctional homolog, near telomer, not expressed
YLR432W	gi | 665971	IMD3	Inosine-5′-monophosphate dehydrogenase
$f_2 = 0.825$[b]			
YLR432W	gi | 665971	IMD3	Inosine-5′-monophosphate dehydrogenase
YHR216W	gi | 458916	IMD2	Inosine-5′-monophosphate dehydrogenase
Subfamily pair: YHR216W:YAR073W, $f_2 = 0.93$ (proposed recent duplication creating a pseudogene)			
Sugar Transporter Family A (4 paralogs, 4 pairs, 3 duplications)[c]			
$f_2 = 0.805$[b]	Pair not associated with any duplication block		
YJR158W	gi | 1015917	HXT16	Sugar transporter repressed by high glucose levels
YNR072W	gi | 1302608	HXT17	Sugar transporter repressed by high glucose levels
$f_2 = 0.806$[b]	Pair not associated with any duplication block		
YDL245C	gi | 1431418	HXT15	Sugar transporter induced by low glucose levels, repressed by high glucose levels
YNR072W	gi | 1302608	HXT17	Sugar transporter repressed by high glucose levels
$f_2 = 0.809$[b]	Pair not associated with any duplication block		
YJR158W	gi | 1015917	HXT16	Sugar transporter repressed by high glucose levels
YEL069C	gi | 603249	HXT13	Sugar transporter induced by low glucose levels, repressed by high glucose levels
$f_2 = 0.810$[b]	Pair not associated with any duplication block		
YEL069C	gi | 603249	HXT13	Sugar transporter induced by low glucose levels, repressed by high glucose levels
YDL245C	gi | 1431418	HXT15	Sugar transporter
Subfamily pair: YEL069C:YNR072W, $f_2 = 0.932$ (proposed recent duplication)			
Subfamily pair: YJR158W:YDL245C, $f_2 = 1.000$ (proposed very recent duplication)			
Chaperone Family A (2 paralogs, 1 pair, 1 duplication)[a]			
$f_2 = 0.81$	Pair associated with Wolfe duplication block 48		
YMR186W	gi | 854456	HSC82	Cytoplasmic chaperone induced two- to threefold by heat shock
YPL240C	gi | 1370495	HSP82	Cytoplasmic chaperone, pheromone signaling, Hsf1p regulation

TABLE 7
(*Continued*)

SGD Name	gi Number	Trivial Name	Annotation and Comments
Phosphatase/Thiamine Transport Family A (2 paralogs, 1 pair, 1 duplication)[c]			
$f_2 = 0.818$	Pair not associated with any duplication block		
YBR092C	gi | 536363	PHO3	Acid phosphatase implicated in thiamine transport
YBR093C	gi | 536365	PHO5	Acid phosphatase one of three repressible phosphatases
Pyruvate Decarboxylase Family A (2 paralogs, 1 pair, 1 duplication)[c]			
$f_2 = 0.835$	Pair not associated with any duplication block		
YLR044C	gi | 1360375	PDC1	Pyruvate decarboxylase, major isoform
YLR134W	gi | 1360549	PDC5	Pyruvate decarboxylase, minor isoform
By ortholog analysis, *S. bayanus* (gi | 515236) diverged from *cerevisiae* after the $f_2 = 0.835$ duplication, *Kluyveromyces* diverged before			
Glyceraldehyde-3-Phosphate Dehydrogenase Family (3 paralogs, 3 pairs, 2 duplications)[c]			
$f_2 = 0.845$[b]	Pair not associated with any duplication block		
YJL052W	gi | 1008189	TDH1	Glyceraldehyde-3-phosphate dehydrogenase
YGR192C	gi | 1323341	TDH3	Glyceraldehyde-3-phosphate dehydrogenase
$f_2 = 0.845$[b]	Pair not associated with any duplication block		
YJL052W	gi | 1008189	TDH1	Glyceraldehyde-3-phosphate dehydrogenase
YJR009C	gi | 1015636	TDH2	Glyceraldehyde-3-phosphate dehydrogenase
Subfamily pair: YJR009C: YGR192C, $f_2 = 0.991$ (proposed very recent duplication)			
Alcohol Dehydrogenase Family (2 paralogs, 1 pair, 1 duplication)[c]			
$f_2 = 0.848$	Pair not associated with any duplication block		
YMR303C	gi | 798945	ADH2	Alcohol dehydrogenase, glucose-repressible
YOL086C	gi | 1419926	ADH1	Alcohol dehydrogenase, constitutive
Spermine Transporter Family (2 paralogs, 1 pair, 1 duplication)[a]			
$f_2 = 0.86$	Pair associated with Wolfe duplication block 34		
YGR138C	gi | 1323230	TPO2	Spermine transporter activity
YPR156C	gi | 849164	TPO3	Spermine transporter activity
Sugar Transporter Family B (3 paralogs, 3 pairs, 2 duplications)[c]			
$f_2 = 0.847$[b]	Pair not associated with any duplication block		
YDR343C	gi | 1230670	HXT6	Sugar transporter, high-affinity, high basal levels
YDR345C	gi | 1230672	HXT3	Sugar transporter, low-affinity glucose transporter

(*Continued*)

TABLE 7
(*Continued*)

SGD Name	gi Number	Trivial Name	Annotation and Comments
$f_2 = 0.854$[b]	Pair not associated with any duplication block		
YDR342C	gi | 1230669	HXT7	Sugar transporter, high-affinity, high basal levels
YDR345C	gi | 1230672	HXT3	Sugar transporter, low affinity
Subfamily pair: YDR342C:YDR343C, $f_2 = 0.994$ (proposed very recent duplication)			

[a]Not associated with fermentation. These are associated with duplication blocks within the yeast genome, where the high value of f_2 (typically equilibrated in block paralog pairs) may reflect either variance, or selective pressure to conserve silent sites in individual codons.
[b]These pairs represent a family generated with a single duplication with $0.80 < f_2 < 0.86$ and subsequent duplication(s) in the derived lineages. Paralog pairs are considered only if they have at least 100 aligned silent sites and are not separated by more than 120 accepted point mutations per 100 aligned amino acid sites (PAM units). f_2 = fraction of nucleotides conserved at twofold redundant codon sites only, and only at sites where the amino acid is identical.
[c]Associated with the pathway to make–accumulate–consume ethanol. Genes involved in the fermentation pathway that are not rate limiting generally do not have duplicates in the yeast genome (e.g., hexokinase, glucose-6-phosphate isomerase, phosphofructokinase, aldolase, triose phosphate isomerase, and phosphoglycerate kinase are all present in one isoform). Enolase has two paralogs (ENO1 and ENO2), where $f_2 = 0.946$. These are distantly related to a homolog known as ERR1, with the silent sites equilibrated. Phosphoglycerate mutase has three paralogs (GM1, GM2 and GM3), with silent sites that are essentially equilibrated.

(the red proteins in Figure 30). If the f_2 clock (within its expected variance) is assumed to date paralogs in yeast, this cluster suggests that several genes other than Adh duplicated as part of the emergence of the new make–accumulate–consume strategy, near the time when fleshy fruit arose.

The six gene duplications proposed to have enabled the emergence of a make–accumulate–consume strategy (in the $0.80 < f_2 < 0.86$ window) are *not* associated with one of the documented blocks of genes were duplicated in ancient fungi, possibly as part of a whole genome duplication (WGD) (Wolfe and Shields, 2001; Kellis et al., 2004). Two duplications in genes that are *not* associated with fermentation that fall in the $0.80 < f_2 < 0.86$ window *are* part of a duplication block (see Table 7). The silent sites for most gene pairs associated with blocks are nearly equilibrated (with the exception of ribosomal proteins) and therefore suggest that most blocks arose by duplications more ancient than duplications in the $0.80 < f_2 < 0.86$ window. Therefore, the hypothesis that a set of six time-correlated duplications (Figure 31 and Table 7) generated the

FIGURE 33. (Right) Cretaceous fruit from Patagonia showing insect damage. (Left) Similar example in extant fruit. (From Jorge F. Genise, Museo Paleontológico E. Feruglio, Trelew, Argentina; www.ub.es/dpep/meganeura/53ichnology.htm.)

make–accumulate–consume strategy in yeast near the time when fermentable fruit emerged is not inconsistent with the WGD hypothesis.

The ecology of fermenting fruit is complex. In rotting fruits, *cerevisiae* become dominant after fermentation begins, while osmotic stress and pH, as well as ethanol, appear to inhibit the growth of competing organisms (Pretorius, 2000). Nevertheless, the emergence of bulk ethanol may not be unrelated to other changes in the ecosystem at the end of the Cretaceous (Figure 33), which include the extinction of the dinosaurs and the emergence of mammals and fruit flies (Baudin et al., 1993; Ashburner, 1998; Barrett and Willis, 2001). Thus, this paleogenetics experiment is an interesting step to connect the chemical behavior of individual enzymes operating as part of a multienzyme system, via metabolism and physiology, to the ecosystem and the fitness of organisms within it.

I. RESURRECTING THE ANCESTRAL STEROID RECEPTOR AND THE ORIGIN OF ESTROGEN SIGNALING

Another narrative to be supported by paleomolecular resurrection focused on receptors for steroid hormones by Thornton et al. (2003). Steroid receptors are widely distributed throughout the chordates. In the

pre-genomic age, steroid receptors were not known in invertebrates, either by analysis of fully sequenced invertebrate genomes or by classical biochemical studies. Because of this, steroid receptors were thought to have arisen early in the divergence of chordates, perhaps some 400 to 500 Ma, presumably via the duplication of a more ancient receptor gene (Escriva et al., 1997; Giguere, 2002; Baker, 2003).

Vertebrate-type steroids appear to be involved in the reproductive endocrinology of certain mollusks, however (Di Cosmo et al., 2001, 2002). Further, arthropods and nematodes, where the most intensive studies had been done to suggest that steroid receptors are absent, both use ecdysone as a molting hormone (Peterson and Eernisse, 2001), making it conceivable that steroid receptors were lost in the lineage leading to ecdysozoa. Further, analysis of receptor sequences, the ancestral sequence of steroid receptor (AncSR1), and structure mapping studies indicated that the ancient progenitor of this protein class was most similar to extant estrogen receptors (Thornton, 2001).

Guided by these hints, Thornton and co-workers (2003) used degenerate PCR and rapid amplification of cDNA ends to isolate putative estrogen receptor sequence from the mollusk *Aplysia californica*. The protein sequence of the *Aplysia* receptor's DNA-binding domain (DBD) was found to be more similar to that of the vertebrate estrogen receptors than to those of other nuclear receptors. This suggested that *Aplysia* might have a true estrogen receptor.

To characterize the molecular function of the *Aplysia* estrogen receptor, Thornton separately analyzed the activity of its DNA binding domain (DBD) and its ligand-binding domain (LBD) by expressing them in fusion constructs in a cell culture system (Green and Chambon, 1987). The *Aplysia* DBD was fused with a constitutive activation domain (AD) and the construct was cotransfected with an estrogen response element (ERE) luciferase reporter gene into CHO-K1 cells. The *Aplysia* estrogen receptor–DBD fusion protein activated luciferase expression approximately 10-fold above control levels, which is slightly more than that produced by the analogous construct using the human estrogen receptor DBD.

The identification of a putative estrogen receptor from the mollusk suggested that steroid receptors might be more ancient in metazoans than had been thought. Barring lateral transfer, the presence of an estrogen receptor in both mollusks and chordates suggested that steroid receptors arose prior to the divergence of bilaterally symmetric animals. According

to this hypothesis, this gene was lost in the lineage leading to arthropods and nematodes.

The existence of a receptor family in chordates and mollusks does not, of course, require that the ligand specificity be the same throughout the family history. To use a paleobiochemical experiment to determine this, Thornton et al. analyzed 74 steroid and related receptors, including the putative *Aplysia* estrogen receptor, using maximum parsimony and Bayesian Markov chain Monte Carlo (BMCMC) techniques. Both methods suggested that the *Aplysia* sequence is an ortholog of the vertebrate estrogen receptors, with a BMCMC posterior probability of 100%, a bootstrap proportion of 90%, and a decay index of 6. Although BMCMC probabilities can overestimate statistical confidence (Hillis and Bull, 1993; Suzuki et al., 2002), the result suggests that the steroid receptors originated about 600 to 1200 Ma as the major metazoan phyla were just beginning to diverge (Benton and Ayala, 2003).

To characterize the functionality of the LBD, a fusion of the *Aplysia* ER-LBD with a Gal4-DBD was cotransfected with a luciferase reporter driven by an upstream activator sequence, the response element for Gal4-DBD. The *Aplysia* ER-LBD activates transcription constitutively, and none of a set of vertebrate steroid hormones, including estrogens, androgens, progestins, and corticoids, further activated or repressed this activation, even at micromolar doses.

Thornton et al. (2003) then resurrected the conserved functional domains of the ancestral steroid receptor (AncSR1) from which all extant steroid receptors evolved. A maximum likelihood joint reconstruction algorithm in PAML (Yang, 1997) was used to obtain the protein sequences of the ancestral AncSR1 DNA binding domain and ligand binding domains using a sequence matrix (Gonnet et al., 1992), the Jones model for amino acid replacement, a gamma distribution of evolutionary rates across sites with for rate categories, and a tree determined by maximum parsimony methods. Although support for the AncSR1 node is not strong in a parsimony context, it has 100% posterior probability, and the best tree with this node has a likelihood 4.7 million times greater than the best tree without this node.

The matrix included broadly sampled representatives of both ligand-regulated and ligand-independent receptors. The maximum likelihood sequence inferred for the AncSR1-DBD had a mean probability of 81% per site; the AncSR1-LBD had a mean probability of 62%.

The amino acid chosen for each position in the ancestral (AncSR1-DBD and AncSR1-LBD) sequences was the one amino acid preferred at each site.

The appropriate gene was synthesized and subcloned into fusion constructs. When the constructs containing AncSR1 domains were expressed in CHOK1 cells, their activity was consistent with the prediction that the ancestral receptor would function like an estrogen receptor (Thornton, 2001). For example, the AncSR1-DBD fusion increased transcription from an estrogen response element about fourfold, which is slightly less than the human ER-DBD but significantly greater than that seen in controls. Other extant steroid receptors do not activate effectively on EREs (Zilliacus et al., 1994).

The AncSR1-LBD activated transcription in a dose-dependent fashion in the presence of estradiol, estrone, and estriol. The magnitude of hormone-induced activation was smaller than that effected by the human ER-LBD with estradiol. Similar results were obtained with a radioligand binding assay. Here, the concentration to the midpoint was in the range 10 to 100 n*M*; the corresponding range in human ER-LBD construct was in the range 1 to 3 n*M*.

To assess specificity, other steroids, including androgens, progestins, and corticoids, were examined. These gave still smaller maximal activation of the AncSR1, ranging from 1 to 45% of that observed with estradiol. Further, dose–response analysis shows that the ancestor was 30,000-fold less sensitive to these ligands than to estradiol. This suggests a high level of estrogen specificity.

It is unlikely that the specificity of estrogen activation is an artifact. Of the 26 sites in the ligand-binding pocket, 22 were identical to a human estrogen receptor. Thornton et al. (2003) argued that because random mutation impairs function more frequently than enhancing it in estrogen receptors, error in the reconstruction is more likely to reduce the efficiency of steroid binding and activation than to create it de novo. This is analogous to the argument used in resurrecting ancestral digestive ribonucleases.

These findings provide empirical support for the hypothesis that the ancient ancestral SR functioned a billion years ago as an estrogen receptor. This result is also consistent with the hypothesis that this ancient steroid receptor gene was lost along the lineage leading to ecdysozoa. Thornton et al. suggested that in the lineage leading to the *Aplysia* ER, ligand regulation was lost.

This narrative indicates how paleomolecular resurrections can open up biological understanding in a way that is difficult to do in any other way. First, the narrative suggests a previously unrecognized degree of functional and genomic lability in steroid receptors. The existence of a primitive steroid receptor also suggests that many other non-ecdysozoan metazoans,

including echinoderms, annelids, and platyhelminthes, will also have steroid receptors. Given limited experimental evidence for a reproductive role of steroid hormones in cephalopod and gastropod mollusks (Di Cosmo et al., 2001, 2002; Oberdorster and McClellan-Green, 2002), Thornton et al. suggested that the loss of estrogen-dependent activation in the *Aplysia* system is recent and unique to the Opisthobranchs.

Thornton et al. did not take the next step—to place their work within a planetary context. It is possible to do so in many ways. For example, the biosynthesis of estrogen requires molecular oxygen. This is not only because of the conversion of squalene to squalene oxide, an early step in lanosterol biosynthesis, or in the conversion of lanosterol to cholesterol, all very basic to eukaryotic biology. Estradiol also requires aromatase, the enzyme discussed in Section I.B. This enzyme also requires dioxygen and is part of a cytochrome P450 family of proteins that underwent explosive divergence near the time of the origins of the metazoans, relatively late compared with the dioxygen-utilizing enzymes that make squalene oxide and oxidatively decarboxylate lanosterol.

It would be helpful to revisit the problem of the steroid receptors as more metazoan genomes become available. This would include sampling a larger number of candidate ancestral estrogen receptors to demonstrate the robustness of the inferences with respect to the ambiguity inherent in this analysis. The emergence of the metazoan endocrine system is a fascinating theme in the emergence of multicellularity in animals. Molecular paleoscience will undoubtedly play a key role in the development of our understanding of this important event in the history of life on Earth, and generate a deeper understanding of developmental biology in the process.

J. ANCESTRAL CORAL FLUORESCENT PROTEINS

Some experiments in paleobiochemistry are exceptionally elegant in their execution. For example, proteins homologous to the green fluorescent protein (GFP) from *Aequorea victoria* exploit two autocatalytic consecutive reactions to complete the synthesis of the chromophores that generate the green and cyan emissions. Red fluorescent proteins and purple-blue chromoproteins require a third reaction to prepare the emitter. By this criterion, the red and purple-blue chromoproteins are more ''advanced'' (Uglade et al., 2004).

Nevertheless, the natural world contains several examples of red/green color diversification within this superfamily (Shagin et al., 2004). This may

reflect convergent evolution of molecular complexity. To examine this issue, Ugalde et al. (2004) studied the historical event that gave rise to the color diversity exhibited by the great star coral *Montastraea cavernosa.* This coral has several genes coding for fluorescent proteins that generate cyan, shortwave green, longwave green, and red emissions (Kelmanson and Matz, 2003).

The sequences were inferred and synthesized for the common ancestor of all *M. cavernosa* colors (All ancestor), the common ancestor of red proteins (Red ancestor), and two intermediate nodes corresponding to the possible common ancestors of red and longwave green proteins (Red/Green ancestor and pre-Red ancestor) (Figure 34). Three models of evolution based on different types of sequence information—amino acids, codons, and nucleotides (Yang, 1997)—were used. The reconstructions of all four ancestral sequences were largely robust under these models, with average posterior probabilities at a site ranging from 0.96 to

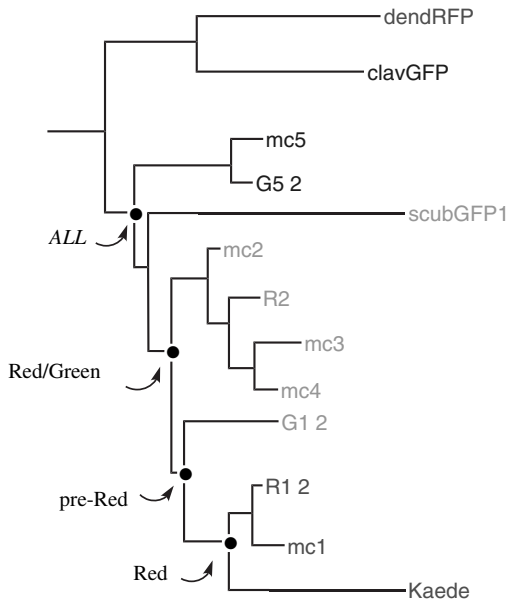


FIGURE 34. Phylogeny of GFP-like proteins from the great star coral *M. cavernosa* and closely related coral species. The red and cyan proteins from soft corals (dendRFP and clavGFP) represent an outgroup. The highlighted nodes (All, Red/Green, pre-Red, and Red) were resurrected in the study reviewed here.

0.99. Still, the models were in disagreement at between four and eight sites (out of a total of 217).

Ugalde et al. (2004) therefore designed codons corresponding to these sites to be degenerate. Bacteria were transformed with plasmids carrying the genes encoding the ancestral proteins. This permitted a library to sample alternative predictions. For each type of the ancestral gene, the protein products were found to display identical fluorescent phenotypes regardless of the amino acid at the ambiguous site.

Green color at any of the ancestral nodes would indicate that red proteins from other parts of the phylogeny were the result of convergent evolution. In contrast, if red fluorescent arose only once, all ancestors would emit red fluorescence. The All ancestor turned out to emit in the shortwave green. The two possible common ancestors of red and green proteins (Red/Green and pre-Red) showed an intermediate longwave green/red emission. Although most of the protein remained longwave green, a small fraction was able to complete the third autocatalytic step in the chromophore synthesis, resulting in a small amount of red emission.

Clones of the Red ancestor showed an ''imperfect red'' phenotype. Although the red emission dominated, the conversion of green to red was still less efficient than in modern reds, resulting in a prominent peak of green fluorescence.

These results suggested that because the All ancestor was green and not red, the red emission color within the superfamily of GFP-like proteins has originated more than once. This establishes the convergent evolution of this complex molecular system. The red emitting system evidently evolved from the green emitting system through a stepwise accumulation of amino acid replacements.

The most elegant part of this report was undoubtedly the presentation of the results. Here, a phylogenetic tree was traced onto the growth medium in a petri dish, with the ancestral and modern sequences appropriately placed. When observed under ultraviolet light, the dish gives a visual image of the evolution of fluorescent colors.

The planetary biology of this system was not discussed by Ugalde et al. (2004). It remains a subject for discussion why these organisms fluoresce. Clearly, however, the emission of light is a way of interacting with an environment, which is changing. Although the visual pigments of the organisms intended to receive the light are not, to our knowledge, a parallel evolutionary history must exist with these. As the post-genomic age develops, there is little doubt that these will be found and understood.

K. ISOCITRATE DEHYDROGENASE

Evolutionary analysis indicates that eubacterial NADP-dependent isocitrate dehydrogenases (IDH) evolved from an NAD-dependent precursor about 3.5 billion years ago. Various authors have suggested that selection in favor of utilizing NADP was a result of an expansion of an ecological niche during growth on acetate, where isocitrate dehydrogenase provides 90% of the NADPH necessary for biosynthesis.

Dean and his co-workers (Zhu et al., 2005) used experimental paleobiochemistry to explore the divergence of cofactor specificity in this system starting from the NADP-dependent IDH (NADP-IDH) (Hurley et al., 1989) and the NAD-dependent isopropylmalate dehydrogenase (NAD-IMDH) (Imada et al., 1991). This example illustrated the limits of an analysis when overall amino acid composition was low. The degree of divergence here is very high; only 17 to 24% of the sites are identical across the system. Therefore, the sequence alignments obtained using Clustal W (Thompson et al., 1994) misaligned residues that were independently known to be critical to substrate binding and catalysis. Therefore, the multiple sequence alignment modified using the crystallographic structures of *E. coli* NADP-dependent IDH (NADP-IDH) (Hurley et al., 1989) and *T. thermophilus* NAD-dependent isopropylmalate dehydrogenase (NAD-IMDH) (Imada et al., 1991) as guides.

Three amino acids responsible for differing coenzyme specificities were identified from x-ray crystallographic structures of NADP bound to *E. coli* isocitrate dehydrogenase and the distantly related *T. thermophilus* NAD-dependent isopropylmalate dehydrogenase bound to NAD (an enzyme that is homologous to IDH and uses NAD) (Hurley et al., 1991; Hurley and Dean, 1994; Yasutake et al., 2003). In a previous study (Dean and Golding, 1997) the authors determined, using a maximum likelihood ancestral reconstruction, that the six conserved residues binding NADP in IDH were introduced very early in evolution. The complementation of the structural and ancestral reconstruction studies revealed that the three residues responsible for binding the cofactor (NADP) were ancestral. The other residues were not conserved in the NAD-binding enzymes and are therefore not involved in binding NAD. Based on these results, site-directed mutagenesis targeting these sites in the binding pocket has been used to invert the coenzyme specificity of *E. coli* IDH from NADP to NAD. The specificity was confirmed with kinetic experiments.

Having determined the ancestral residues responsible for cofactor specificity, the authors set out to test the hypothesis that NADP use is an

adaptation to growth on acetate (Dean and Golding, 1997). When bacteria grow on highly reduced (energy-rich) compounds such as glucose, NADPH (the reduced form of NADP) for biosynthesis is obtained by diverting energy-rich carbon from glycolysis into the oxidative branch of the pentose phosphate pathway (Neidhardt et al., 1990). During growth on acetate, which is a highly oxidized (energy-poor) compound, there is no energy-rich carbon to divert into the oxidative branch, and therefore alternative sources of NADPH are required. A competition study of genetically identical bacterial strains, except for the gene coding for NADP-dependent IDH, was undertaken. One strain had the wild-type NADP-IDH and the second had the IDH with the ancestral NAD binding residues obtained from the first part of the study. Chemostat competition experiments were conducted on either glucose or acetate as the only limiting factor. The ancestral NAD binding IDH was strongly selected against when grown on acetate, yet it was favored over the NADP binding wild-type when grown on glucose. These results supported the hypothesis.

To further establish selection on acetate, not some other factor, such as altered kinetics or less efficient regulation, the authors devised a test to confirm their finding. If the selection at the IDH gene is truly caused by cofactor specificity, then if other sources of NADPH were removed, selection for the IDH-NADP gene should increase. Conversely if this effect is independent of coenzyme use, no such increase should be observed. Indeed, competition studies on strains with deletion background in genes contributing NADPH (maeB, pntAB, and udhA encoding a soluble transhydrogenase, UdhA) showed that selection on acetate against the ancestral NAD-IDH increases as the sources of NADPH decline. These results further support the hypothesis.

Genomic comparisons also supported the plaeoscience results. It was found that isocitrate lyase (ICL), an essential enzyme for growth on acetate, was always found in all of the 46 genomes that also encode NADP-dependent IDH. Further, no ICL is found in the 12 genomes encoding NAD-dependent IDH. Members of both groups represent highly diverse phylogenies, including archaea, bacilli, and α- and γ-proteobacteria, and with variable metabolic lifestyles and habitats. The genomic studies in isolation would have been merely a hint, but when added to the paleobiochemical experimental results, they become part of a sophisticated scientific narrative that not only shows evolutionary adaptation but also clarifies the structural and mechanistic basis for such selection.

V. GLOBAL LESSONS

The results of these 20 examples provide the first glimpse of the power of a field that is just beginning. These examples also provide rejoinders to some of the objections that are frequently raised by those who criticize the paradigm. First, the resurrected sequences provide more information than consensus models, single-residue swapping, or simply comparing ancestral sequences. In many cases, the properties of the ancestral proteins are not simply an average of the properties of the descendent proteins (Axe, 2000). This is a consequence of a very fundamental principle of organic chemistry; the whole is not a linear sum of the parts in most molecules, and certainly not in proteins. The implication of this generalization is that selective reconstruction might not be sufficient to draw accurate inferences. The more complete the resurrection, the more likely the result is to give a reliable result.

These examples show a wide range of tactics for managing ambiguities. With the exception of those who chose to use consensus sequences to approximate an ancestral sequence, the degree to which ambiguity was managed reflected in large part the extent to which the data allowed the ambiguity to be managed. Sequence data are adequate to ensure that if the target for resurrection lived in the past 100 million years, the ambiguity could be sampled comprehensively. In more ancient resurrections, limitations from the data set meant that the sampling was less comprehensive.

One outcome of these studies was how infrequently it was found that ambiguities compromised the biology. In general, when the ambiguity was sampled comprehensively, a few of the proteins appeared to be defective to an extent that that particular sequence was deemed to be unlikely to be a true ancestor. Otherwise, it was difficult to find an example where ambiguity caused a biological interpretation to be uncertain.

This observation is, of course, good news for those working in the area. As suggested in Section II, it may reflect the fact that ambiguity is generally at sites that have suffered a large amount of neutral amino acid replacement, and that neutral drift generally means that replacement of an amino acid at that site does not have any impact on a functionally significant behavior of a protein. Further work is required to explore the generality of this observation.

It is clear that functionally significant *in vitro* behaviors for a protein where new biological function has emerged can also be identified by

paleomolecular resurrections. Several examples now exist of a strategy that examines the behavior of proteins resurrected from points in history before and after the episode of adaptive evolution, as indicated by a high rate of sequence change. Those behaviors that are rapidly changing during the episode of adaptive sequence evolution, by hypothesis, confer selective value on the protein in its new function, and therefore are relevant to the change in function, either directly or by close coupling to behaviors that are. The *in vitro* properties that are the same at the beginning and end of this episode are not relevant to the change in function.

By the end of the current century, all information that we have about life on Earth will be captured within a network of narratives that describe specific biomolecular systems, interactions between these systems, and models (formal and heuristic) that organize these interactions at various levels (cell, organism, and ecosystem). This network will, of course, include molecular structures from natural products chemistry, genomics, metabolomics, and crystallography, as well as mathematical models from systems biology. But it will also describe biomolecular systems as the products of Darwinian evolution, including historical accidents, drift, and adaptive selection. It will therefore include narratives describing the *history* of these systems, set within a model for the history of the global biosphere built from the geological, paleontological, and genomic records.

None of these narratives will be individually compelling or in any sense proven. Nor could they be. Human knowledge is (at its core) heuristic and intuitive, even that captured by mathematical formalisms. The strength of the global theory must therefore come from the density of its interconnections between its narratives, its ability to ''explain'' (in a human sense) behaviors of these systems, and the range of information, from molecular to geological, that it coherently unifies. Thus, the global biotherapy will be analogous to structure theory from chemistry, which combines heuristic and formal models, all lacking ''proof'' but nevertheless offering understandable explanations that when combined with human intuition, confer predictive and manipulative power.

The impact of incorporating natural history into medicine in the coming century will be as broad as the impact created by incorporating chemistry into medicine over the past century. Human disease is in many cases the consequence of imperfect adaptation to a changing biosphere, is aggavated by changing environments in modern civilization, and evolves in patients both spontaneously and is response to medical treatment. The global theory will capture this historical dynamic and allow medicine to exploit the

truism that *any* system, natural or human-made, is better understood when we understand *both* its structure *and* its history. Under this new paradigm, the global theory will allow biomedical researchers to use natural history to answer the ''why?'' questions that reductionist models leave largely unanswered.

Paleogenetics experiments will be key to the development of this vision. The fundamental objection that many molecular biologists have to historical explanations for modern biology come from the view that historical models are inherently untestable, and therefore intrinsically unscientific. Similarly, much evolutionary bioinformatics examines collections of protein sequences, not specific cases, because most statisticians and computer scientists view a theory based on case studies as not being entirely ''correct.'' Further, many scientists are not prepared to concede that human knowledge is fundamentally heuristic. This fact, well understood by those who study logic, epistemology, and the history of science, is viewed by many scientists as an affront.

Therefore, the key task in engineering the maturation of the biomedical paradigm from one that focuses on ''hard'' science (chemistry and mathematics) to one that also includes ''soft'' natural history is to identify experimental tools to test historical models and generalize case studies. Experimental paleogenetics is one tool now available to do both.

REFERENCES

Adachi, J., and Hasegawa, M. (1996). MOLPHY, version 2.3: programs for molecular phylogenetics based on maximum likelihood, *Comput. Sci. Monogr.*, 28.

Adey, N. B., Tollefsbol, T. O., Sparks, A. B., Edgell, M. H., and Hutchison, C. A., III (1994). Molecular resurrection of an extinct ancestral promoter for mouse L1, *Proc. Natl. Acad. Sci. U.S.A.*, 91: 1569–1573.

Allard, M. W., Miyamoto, M. M., Jarecki, L., Kraus, F., and Tennant, M. R. (1992). DNA systematics and evolution of the artiodactyl family Bovidae, *Proc. Natl. Acad. Sci. U.S.A.*, 89: 3972–3976.

Arai, K. I., Kaziro, Y., and Kawakita, M. (1972). Studies on polypeptide elongation factors from *Escherichia coli*, *J. Biol. Chem.*, 247: 7029–7039.

Ardelt, W., Mikulski, S. M., and Shogen, K. (1991). Amino acid sequence of an anti-tumor protein from *Rana pipiens* oocytes and early embryos: homology to pancreatic ribonucleases, *J. Biol. Chem.*, 266: 245–251.

Arnason, U., Gullberg, A., and Janke, A. (1998). Molecular timing of primate divergences as estimated by two nonprimate calibration points, *J. Mol. Evol.*, 47: 718–727.

Ashburner, M. (1998). Speculations on the subject of alcohol dehydrogenase and its properties in *Drosophila* and other flies, *Bioessays*, 20: 949–954.

Axe, D. D. (2000). Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors, *J. Mol. Biol.*, 301: 585–595.

Baker, M. E. (2003). Evolution of adrenal and sex steroid action in vertebrates: a ligand based mechanism for complexity, *Bioessays*, 25: 396–400.

Barnard, E. A. (1969). Biological function of pancreatic ribonuclease, *Nature*, 221: 340–344.

Barrett, P. M., and Willis, K. J. (2001). Did dinosaurs invent flowers? Dinosaur angiosperm coevolution revisited, *Biol. Rev.*, 76: 411–447.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*, *Nucleic Acids Res.*, 21: 3329–3330.

Beintema, J. J., and Gruber, M. (1967). Amino acid sequence in rat pancreatic ribonuclease, *Biochim. Biophys. Acta.*, 147: 612–614.

Beintema, J. J., and Gruber, M. (1973). Rat pancreatic ribonuclease. 2. Amino acid sequence, *Biochim. Biophys. Acta.*, 310: 161–173.

Beintema, J. J., and Martena, B. (1982). Primary structure of porcupine (*Hystrix cristata*) pancreatic ribonuclease: close relationship between African porcupine (an Old World hystricomorph) and New World caviomorphs, *Mammalia*, 46: 253–257.

Beintema, J. J., Gaastra, W., and Munniksma, J. (1979). Primary structure of pronghorn pancreatic ribonuclease: close relationship between giraffe and pronghorn, *J. Mol. Evol.*, 13: 305–316.

Beintema, J. J., Wietzes, P., Weickmann, J. L., and Glitz, D. G. (1984). The amino acid sequence of human pancreatic ribonuclease, *Anal. Biochem.*, 136: 48–64.

Beintema, J. J., Broos, J., Meulenberg, J., and Schuller, C. (1985). The amino acid sequence of snapping turtle (*Chelydra serpentina*) ribonuclease, *Eur. J. Biochem.*, 153: 305–312.

Beintema, J. J., Schuller, C., Irie, M., and Carsana, A. (1988). Molecular evolution of the ribonuclease superfamily, *Prog. Biophys. Mol. Biol.*, 51: 165–192.

Benner, S. A. (1988). Extracellular "communicator RNA," *FEBS Lett.*, 233: 225–228.

Benner, S. A. (2002). The past as the key to the present: resurrection of ancient proteins from eosinophils, *Proc. Natl. Acad. Sci. U.S.A.*, 99: 4760–4761.

Benner, S. A. (2003). Interpretive proteomics: finding biological meaning in genome and proteome databases, *Adv. Enzyme Regul.*, 43: 271–359.

Benner, S. A., and Allemann, R. K. (1989). The return of pancreatic ribonucleases, *Trends Biochem. Sci.*, 14: 396–397.

Benner, S. A., and Ellington, A. D. (1990). Evolution and structural theory: the frontier between chemistry and biology, in *Bioorganic Chemistry Frontiers* (Dugas, H., Ed.), Springer-Verlag, Berlin, pp. 1–54.

Benner, S. A., and Gerloff, D. (1991). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases, *Adv. Enzyme Regul.*, 31: 121–181.

Benner, S. A., Cannarozzi, G., Gerloff, D., Turcotte, M., and Chelvanayagam, G. (1997a). Bona fide predictions of protein secondary structure using transparent analyses of multiple sequence alignments, *Chem. Rev.*, 97: 2725–2843.

Benner, S. A., Haugg, M., Jermann, T. M., Opitz, J. G., Raillard-Yoon, S.-A., Soucek, J., Stackhouse, J., Trabesinger-Ruef, N., Trautwein-Fritz, K., and Zankel, T. R. (1997b). Evolutionary reconstructions in the ribonuclease family, in *Ribonucleases* (D'Alessio, G. and Riordan, J. F., Eds.), Academic Press, New York, pp. 214–244.

Benner, S. A., Trabesinger, N., and Schreiber, D. (1998). Post-genomic science: converting primary structure into physiological function, *Adv. Enzyme Regul.*, 38: 155–180.

Benner, S. A., Caraco, M. D., Thomson, J. M., and Gaucher, E. A. (2002). Planetary biology: paleontological, geological, and molecular histories of life, *Science*, 296: 864–868.

Benton, M. J., and Ayala, F. J. (2003). Dating the tree of life, *Science*, 300: 1698–1700.

Berbee, M. L., and Taylor, J. W. (1993). Dating the evolutionary radiations of the true fungi, *Can. J. Bot.*, 71: 1114–1127.

Bielawski, J. P., and Yang, Z. H. (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution, *J. Mol. Evol.*, 59: 121–132.

Blackburn, P., and Moore, S. (1982). *Pancreatic Ribonucleases: The Enzymes*, Academic Press, New York, pp. 317–433.

Bonaventura, J., Bonaventura, C., and Sullivan, B. (1974). Urea tolerance as a molecular adaptation of elasmobranch hemoglobins, *Science*, 186: 57–59.

Boulton, B., Singleton, V. L., Bisson, L. F., and Kunkee, R. E. (1996). Yeast and biochemistry of ethanol fermentation, in *Principles and Practices of Winemaking*, Chapman & Hall, New York, pp. 139–172.

Bozzi, A., Saliola, M., Falcone, C., Bossa, F., and Martini, F. (1997). Structural and biochemical studies of alcohol dehydrogenase isozymes from *Kluyveromyces lactis*, *Biochim. Biophys. Acta*, 1339: 133–142.

Brand, A. H., and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes, *Development*, 118: 401–415.

Breukelman, H. J., Beintema, J. J., Confalone, E., Costanzo, C., Sasso, M. P., Carsana, A., Palmieri, M., and Furia, A. (1993). Sequences related to the ox pancreatic ribonuclease coding region in the genomic DNA of mammalian species, *J. Mol. Evol.*, 37: 29–35.

Breukelman, H. J., Jekel, P. A., Dubois, J. Y. F., Mulder, P., Warmels, H. W., and Beintema, J. J. (2001). Secretory ribonucleases in the primitive ruminant chevrotain (*Tragulus javanicus*), *Eur. J. Biochem.*, 268: 3890–3897.

Brochier, C., and Philippe, H. (2002). Phylogeny: a non-hyperthermophilic ancestor for bacteria, *Nature*, 417: 244–244.

Buck, M., and Rosen, M. K. (2001). Structural biology: flipping a switch, *Science*, 291: 2329–2330.

Cai, W., Pei, J., and Grishin, N. V. (2004). Reconstruction of ancestral protein sequences and its applications, *BMC Evol. Biol.*, 4: 33.

Cao, Y., Adachi, J., Yano, T. A., and Hasegawa, M. (1994). Phylogenetic place of guinea pigs: no support of the rodent polyphyly hypothesis from maximum likelihood analyses of multiple protein sequences, *Mol. Biol. Evol.*, 11: 593–604.

Carroll, R. L. (1988). *Vertebrate Paleontology and Evolution*, W. H. Freeman, New York.

Carsana, A., Confalone, E., Palmieri, M., Libonati, M., and Furia, A. (1988). Structure of the bovine pancreatic ribonuclease gene: the unique intervening sequence in the 5′ untranslated region contains a promoter-like element, *Nucleic Acids Res.*, 16: 5491–5502.

Cavalier-Smith, T. (2002). The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification, *Int. J. Syst. Evol. Microbiol.*, 52: 7–76.

Cavender, T. M. (1991). The fossil record of the Cyprinidae, in *Cyprinid Fishes: Systematics, Biology and Exploitation* (Winfield, I. J., and Nelson, J. S., Eds.), Chapman & Hall, London.

Chalepakis, G., Stoykova, A., Wijnholds, J., Tremblay, P., and Gruss, P. (1993). Pax: gene regulators in the developing nervous system, *J. Neurobiol.*, 24: 1367–1384.

Chandrasekharan, U. M., Sanker, S., Glynias, M. J., Karnik, S. S., and Husain, A. (1996). Angiotensin II forming activity in a reconstructed ancestral chymase, *Science*, 271: 502–505.

Chang, B. S. W., and Donoghue, M. J. (2000). Recreating ancestral proteins, *Trends Ecol. Evol.*, 15: 109–114.

Chang, B. S. W., Jonsson, K., Kazmi, M. A., Donoghue, M. J., and Sakmar, T. P. (2002). Recreating a functional ancestral archosaur visual pigment, *Mol. Biol. Evol.*, 19: 1483–1489.

Chinen, A., Matsumoto, Y., and Kawamura, S. (2005a). Reconstitution of ancestral green visual pigments of zebrafish and molecular mechanism of their spectral differentiation, *Mol. Biol. Evol.*, 22: 1001–1010.

Chinen, A., Matsumoto, Y., and Kawamura, S. (2005b). Spectral differentiation of blue opsins between phylogenetically close but ecologically distant goldfish and zebrafish, *J. Biol. Chem.*, 280: 9460–9466.

Ciglic M. I., Jackson, P. J., Raillard, S. A., Haugg, M., Jermann, T. M., Opitz, J. G., Trabesinger-Ruef, N., and Benner, S. A. (1998). Origin of dimeric structure in the ribonuclease superfamily, *Biochemistry*, 37: 4008–4022.

Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J., and Largaespada, D. A. (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse, *Nature*, 436: 272–276.

Collinson, M. E., and Hooker, J. J. (1991). Fossil evidence of interactions between plants and plant-eating mammals, *Philos. Trans. R. Soc. London Ser. B*, 333: 197–208.

Corbin, C. J., Mapes, S. M., Marcos, J., Shackleton, C. H., Morrow, D., Safe, S., Wise, T., Ford, J. J., and Conley, A. J. (2004). Paralogues of porcine aromatase cytochrome P450: a novel hydroxylase activity is associated with the survival of a duplicated gene, *Endocrinology*, 145: 2157–2164.

Cunningham, C. W. (1999). Some limitations of ancestral character-state reconstruction when testing evolutionary hypotheses, *Syst. Biol.*, 48: 665–674.

Czerny, T., and Busslinger, M. (1995). DNA binding and transactivation properties of *Pax-6*: three amino acids in the paired domain are responsible for the different sequence recognition of *Pax-6* and BSAP (*Pax-5*), *Mol. Cell. Biol.*, 15: 2858–2871.

Czerny, T., Schaffner, G., and Busslinger, M. (1993). DNA sequence recognition by *Pax* proteins: bipartite structure of the paired domain and its binding site, *Genes Dev.*, 7: 2048–2061.

Czerny, T., Halder, G., Kloter, U., Souabni, A., Gehring, W. J., and Busslinger, M. (1999). *Twin of eyeless*, a second *Pax-6* gene of *Drosophila*, acts upstream of *eyeless* in the control of eye development, *Mol. Cell. Biol.*, 3: 297–307.

Dahl, E., Koseki, H., and Balling, R. (1997). *Pax* genes and organogenesis, *Bioessays*, 19: 755–765.

D'Alessio, G., Floridi, A., De Prisco, R., Pignero, A., and Leone, E. (1972). Bull semen ribonucleases. 1. Purification and physico-chemical properties of the major component, *Eur. J. Biochem.*, 26: 153–161.

D'Alessio, G., Di Donato, A., Parente, A., and Piccoli, R. (1991). Seminal RNase: a unique member of the ribonuclease superfamily, *Trends Biochem. Sci.*, 16: 104–106.

Dean, A. M., and Golding, G. B. (1997). Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase, *Proc. Natl. Acad. Sci. U.S.A.*, 94: 3104–3109.

Denkewalter, R. G., Veber, D. F., Holly, F. W., and Hirschma.R. (1969). Studies on total synthesis of an enzyme. 1. Objective and strategy, *J. Am. Chem. Soc.*, 91: 502–503.

Di Cosmo, A., Di Cristo, C., and Paolucci, M. (2001). Sex steroid hormone fluctuations and morphological changes of the reproductive system of the female of *Octopus vulgaris* throughout the annual cycle, *J. Exp. Zool.*, 289: 33–47.

Di Cosmo, A., Di Cristo, C., and Paolucci, M. (2002). A estradiol-17 β receptor in the reproductive system of the female of *Octopus vulgaris*: characterization and immunolocalization, *Mol. Reprod. Dev.*, 61: 367–375.

Dietmann, S., and Holm, L. (2001). Identification of homology in protein structure classification, *Nat. Struct. Biol.*, 8: 953–957.

Dixon, B., Nagelkerke, L. A. J., Sibbing, F. A., Egberts, E., and Stet, R. J. M. (1996). Evolution of MHC class II beta chain-encoding genes in the Lake Tana barbel species flock (*Barbus intermedius* complex), *Immunogenetics*, 44: 419–431.

Doggrell, S. A., and Wanstall, J. C. (2004). Vascular chymase: pathophysiological role and therapeutic potential of inhibition, *Cardiovasc. Res.*, 61: 653–662.

Dubois, J. Y. F., Jekel, P. A., Mulder, P., Bussink, A. P., Catzeflis, F. M., Carsana, A., and Beintema, J. J. (2002). Pancreatic type ribonuclease 1 gene duplications in rat species, *J. Mol. Evol.*, 55: 522–533.

Dupuy, A. J., Akagi, K., Largaespada, D. A., Copeland, N. G., and Jenkins, N. A. (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system, *Nature*, 436: 221–226.

Ellington, A. D., and Benner, S. A. (1987). Free energy differences between enzyme bound states, *J. Theor. Biol.*, 127: 491–506.

Emmens, M., Welling, G. W., and Beintema, J. J. (1976). Amino acid sequence of pike-whale (lesser-rorqual) pancreatic ribonuclease, *Biochem. J.*, 157: 317–323.

Engelkamp, D., and van Heyningen, V. (1996). Transcription factors in disease, *Curr. Opin. Genet. Dev.*, 6: 334–342.

Escriva, H., Safi, R., Hanni, C., Langlois, M.-C., Saumitou-Laprade, P., Stehelin, D., Capron, A., Pierce, R., and Laudet, V. (1997). Ligand binding was acquired during evolution of nuclear receptors, *Proc. Natl. Acad. Sci. U.S.A.*, 94: 6803–6808.

Faculty of 1000 (2004). Evaluations for Gaucher, E. A., et al., *BMC Biol*., Aug. 17, 2004, 2(1): 19; Shelley Copley, Faculty of 1000, Oct. 5, 2004, http://www.f1000biology.com/article/15315709/evaluation.

Feder, M. E., and Mitchell-Olds, T. (2003). Evolutionary and ecological functional genomics, *Nat. Rev. Genet.*, 4: 651–657.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.*, 17: 368–376.

Felsenstein, J. (1989). PHYLIP: phylogeny inference package (version 3.2), *Cladistics*, 5: 164–166.

Fernandez-Espinar, M. T., Barrio, E., and Querol, A. (2003). Analysis of the genetic variability in the species of the Saccharomyces *sensu stricto* complex, *Yeast*, 20: 1213–1226.

Fersht, A. R. (1977), *Enzyme Structure and Mechanism*, W.H. Freeman, New York.

Fleet, G. H., and Heard, G. M. (1993). Yeast growth during fermentation, in *Wine Microbiology and Biotechnology*, Harwood Academic, Chur, Switzerland, pp. 27–54.

Foote, M., Hunter, J. P., Janis, C. M., and Sepkoski, J. J. (1999). Evolutionary and preservational constraints on origins of biologic groups: divergence times of eutherian mammals, *Science*, 283: 1310–1314.

Fortin, A. S., Underhill, D. A., and Gros, P. (1998). Helix 2 of the paired domain plays a key role in the regulation of DNA binding by the *Pax-3* homeodomain, *Nucleic Acids Res.*, 26: 4574–4581.

Franzen, J. L. (1997). *Fossiler Paarhufer Embryo. Nat. Mus.*, 127: 61–62.

Gaastra, W., Groen, G., Welling, G. W., and Beintema, J. J. (1974). Primary structure of giraffe pancreatic ribonuclease, *FEBS Lett.*, 41: 227–232.

Gaastra, W., Welling, G. W., and Beintema, J. J. (1978). Amino acid sequence of kangaroo pancreatic ribonuclease, *Eur. J. Biochem.*, 86: 209–217.

Galtier, N., and Gouy, M. (1998). Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis, *Mol. Biol. Evol.*, 15: 871–879.

Galtier, N., Tourasse, N., and Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms, *Science*, 283: 220–221.

Ganzhorn, A. J., Green, D. W., Hershey, A. D., Gould, R. M., and Plapp, B. V. (1987). Kinetic characterization of yeast alcohol dehydrogenases: amino acid residue 294 and substrate specificity, *J. Biol. Chem.*, 262: 3754–3761.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., and Korber, B. (2002). AIDS: diversity considerations in HIV-1 vaccine selection, *Science*, 296: 2354–2360.

Gaucher, E. A., Miyamoto, M. M., and Benner, S. A. (2001). Function structure analysis of proteins using covarion based evolutionary approaches: elongation factors, *Proc. Natl. Acad. Sci. U.S.A.*, 98: 548–552.

Gaucher, E. A., Gu, X., Miyamoto, M. M., and Benner, S. A. (2002). Predicting functional divergence in protein evolution by site specific rate shifts, *Trends Biochem. Sci.*, 27: 315–321.

Gaucher, E. A., Thomson, J. M., Burgan, M. F., and Benner, S. A. (2003). Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins, *Nature*, 425: 285–288.

Gaucher, E., Graddy, L., Li, T., Simmen, R., Simmen, F., Schreiber, D., Liberles, D., Janis, C., and Benner, S. (2004). The planetary biology of cytochrome P450 aromatases, *BMC Biol.*, 2: 19.

Gerloff, D. L., Cohen, F. E., Korostensky, C., Turcotte, M., Gonnet, G. H., and Benner, S. A. (1997). A predicted consensus structure for the N-terminal fragment of the heat shock protein HSP90 family, *Protein Struct. Funct. Genet.*, 27: 450–458.

Gerloff, D. L., Cannarozzi, G. M., Joachimiak, M., Cohen, F. E., Schreiber, D., and Benner, S. A. (1999). Evolutionary, mechanistic, and predictive analyses of the hydroxymethyldihydropterin pyrophosphokinase family of proteins, *Biochem. Biophys. Res. Commun.*, 254: 70–76.

Giguere, V. (2002). To ERR in the estrogen pathway, *Trends Endocrinol. Metab.*, 13: 220–225.

Glenner, H., Hansen, A. J., Sorensen, M. V., Ronquist, F., Huelsenbeck, J. P., and Willerslev, E. (2004). Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach, *Curr. Biol.*, 14: 1644–1649.

Gonnet, G. H., and Benner, S. A. (1991). *Computational Biochemistry Research at ETH*, Technical Report 154, Departement Informatik, ETH, Zurich.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein sequence database, *Science*, 256: 1443–1445.

Gould, S. J. (1980), *The Panda's Thumb: More Reflections in Natural History*, W. W. Norton, New York.

Graur, D. (1993). Towards a molecular resolution of the ordinal phylogeny of the eutherian mammals, *FEBS Lett.*, 325: 152–159.

Green, S., and Chambon, P. (1987). Estradiol induction of a glucocorticoid responsive gene by a chimeric receptor, *Nature*, 325: 75–78.

Groen, G., Welling, G. W., and Beintema, J. J. (1975). Amino acid sequence of gnu pancreatic ribonuclease, *FEBS Lett.*, 60: 300–304.

Gromiha, M. M., Oobatake, M., and Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins, *Biophys. Chem.*, 82: 51–67.

Halder, G., Callaerts, P., and Gehring, W. J. (1995). Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*, *Science*, 267: 1788–1792.

Harder, J., and Schroder, J. M. (2002). RNase 7, a novel innate immune defense antimicrobial protein of healthy human skin, *J. Biol. Chem.*, 277: 46779–46784.

Hassanin, A., and Douzery, E. J. (2003). Molecular and morphological phylogenies of ruminantia and the alternative position of the moschidae, *Syst. Biol.*, 52: 206–228.

Hermanson, S., Davidson, A. E., Sivasubbu, S., Balciunas, D., and Ekker, S. C. (2004). Sleeping Beauty transposon for efficient gene delivery, in *Zebrafish, 2nd ed., Genetics Genomics and Informatics* (Dietrich, W. H., Westerfield, W. M., and Zon, L. I., Eds.), Elsevier Academic Press, San Diego, pp. 349ff.

Hernandez Fernandez, M., and Vrba, E. S. (2005). A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants, *Biol. Rev. Camb. Philos. Soc.*, 80: 269–302.

Hesse, M. (2002). *Alkaloids: Nature's Curse or Blessing?* Wiley-VCH, Weinheim, Germany.

Hey, J. (1999). The neutralist, the fly and the selectionist, *Trends Ecol. Evol.*, 14: 35–38.

Hillis, D. M., and Bull, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis, *Syst. Biol.*, 42: 182–192.

Hirschmann, R., Nutt, R. F., Veber, D. F., Vitali, R. A., Varga, S. L., Jacob, T. A., Holly, F. W., and Denkewalter, R. G. (1969). Studies on total synthesis of an enzyme. 5. Preparation of enzymatically active material, *J. Am. Chem. Soc.*, 91: 507–508.

Huelsenbeck, J. P. (1997). Is the Felsenstein zone a fly trap? *Syst. Biol.*, 46: 69–74.

Huelsenbeck, J. P., and Bollback, J. P. (2001). Empirical and hierarchical Bayesian estimation of ancestral states, *Syst. Biol.*, 50: 351–366.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology, *Science*, 294: 2310–2314.

Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo, *Mol. Biol. Evol.*, 21: 1123–1133.

Hugenholtz, P., Goebel, B. M., and Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity, *J. Bacteriol.*, 180: 4765–4774.

Hurley, J. H., and Dean, A. M. (1994). Structure of 3-isopropylmalate dehydrogenase in complex with NAD(+): ligand induced loop closing and mechanism for cofactor specificity, *Structure*, 2: 1007–1016.

Hurley, J. H., Thorsness, P. E., Ramalingam, V., Helmers, N. H., Koshland, D. E., Jr., and Stroud, R. M. (1989). Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase, *Proc. Natl. Acad. Sci. U.S.A.*, 86: 8635–8639.

Hurley, J. H., Dean, A. M., Koshland, D. E., and Stroud, R. M. (1991). Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium isocitrate and NADP(+) complexes, *Biochemistry*, 30: 8671–8678.

Imada, K., Sato, M., Tanaka, N., Katsube, Y., Matsuura, Y., and Oshima, T. (1991). Three-dimensional structure of a highly thermostable enzyme, 3-isopropylmalate dehydrogenase of *Thermus thermophilus* at 2.2 Å resolution, *J. Mol. Biol.*, 222: 725–738.

Ipata, P. L., and Felicioli, R. A. (1968). A spectrophotometric assay for ribonuclease activity using cytidylyl-(3′,5′)-adenosine and uridylyl-(3′,5′)-adenosine as substrates, *FEBS Lett.*, 1: 29–31.

Ivics, Z., Hackett, P. B., Plasterk, R. H., and Izsvak, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells, *Cell*, 91: 501–510.

Ivics, Z., Kaufman, C. D., Zayed, H., Miskey, C., Walisko, O., and Izsvak, Z. (2004). The Sleeping Beauty transposable element: evolution, regulation and genetic applications, *Curr. Issues Mol. Biol.*, 6: 43–55.

Iwabata, H., Watanabe, K., Ohkuri, T., Yokobori, S., and Yamagishi, A. (2005). Thermostability of ancestral mutants of *Caldococcus noboribetus* isocitrate dehydrogenase, *FEMS Microbiol. Lett.*, 243: 393–398.

James, K., and Hargreave, T. B. (1984). Immunosuppression by seminal plasma and its possible clinical significance, *Immunol. Today*, 5: 357.

Janis, C. M., Effinger, J. E., Harrison, J. A., Honey, J. G., Kron, D. G., Lander, B., Manning, E., Prothero, D. R., Stevens, M. S., Stucky, R. K., Webb, S. D., and Wright, D. B. (1998). Artiodactyla, in *Evolution of Tertiary Mammals of North America*, Cambridge University Press, Cambridge, pp. 337–357.

Jekel, P. A., Sips, H. J., Lenstra, J. A., and Beintema, J. J. (1979). Amino acid sequence of hamster pancreatic ribonuclease, *Biochimie*, 61: 827–839.

Jenkins, S. R., Nutt, R. F., Dewey, R. S., Veber, D. F., Holly, F. W., Paleveda, W. J., Lanza, T., Strachan, R. G., Schoenew, E. F., Barkemey, H., Dickinson, M. J., Sondey, J., Hirschma, R., and Walton, E. (1969). Studies on total synthesis of an enzyme. 3. Synthesis of a protected hexacontapeptide corresponding to 65–124 sequence of ribonuclease A, *J. Am. Chem. Soc.*, 91: 505–506.

Jermann, T. M. (1995), *Der Ursprung und die Evolution der Ribonuklease aus dem Pankreas und aus der Samenfluessigkeit*, Dissertation 11059, ETH, Zurich.

Jermann, T. M., Opitz, J. G., Stackhouse, J., and Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily, *Nature*, 374: 57–59.

Jukes, T. H., and Kimura, M. (1984). Evolutionary constraints and the neutral theory, *J. Mol. Evol.*, 21: 90–92.

Kelemen, B. R., Klink, T. A., Behlke, M. A., Eubanks, S. R., Leland, P. A., and Raines, R. T. (1999). Hypersensitive substrate for ribonuclease, *Nucleic Acids Res.*, 27: 3696–3701.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*, *Nature*, 428: 617–624.

Kelly, R. W., and Critchley, H. O. (1997). Immunomodulation by human seminal plasma: a benefit for spermatozoon and pathogen? *Hum. Reprod.*, 12: 2200–2207.

Kelmanson, I. V., and Matz, M. V. (2003). Molecular basis and evolutionary origins of color diversity in great star coral *Montastraea cavernosa* (Scleractinia: Faviida), *Mol. Biol. Evol.*, 20: 1125–1133.

Kim, J. S., Soucek, J., Matousek, J., and Raines, R. T. (1995). Mechanism of ribonuclease cytotoxicity, *J. Biol. Chem.*, 270: 31097–31102.

Kleineidam, R. G., Jekel, P. A., Beintema, J. J., and Situmorang, P. (1999). Seminal type ribonuclease genes in ruminants, sequence conservation without protein expression? *Gene*, 231: 147–153.

Knauth, L. P. (2005). Temperature and salinity history of the Precambrian ocean: implications for the course of microbial evolution, *Palaeogeogr. Palaeoclimatol. Palaeoecol.*, 219: 53–69.

Knight, P. A., Wright, S. H., Lawrence, C. E., Paterson, Y. Y., and Miller, H. R. (2000). Delayed expulsion of the nematode *Trichinella spiralis* in mice lacking the mucosal mast cell specific granule chymase, mouse mast cell protease 1, *J. Exp. Med.*, 192: 1849–1856.

Koshi, J. M., and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences, *J. Mol. Evol.*, 42: 313–320.

Kreitman, M., and Akashi, H. (1995). Molecular evidence for natural selection, *Ann. Rev. Ecol. Syst.*, 26: 403–422.

Krishnan, N. M., Seligmann, H., Stewart, C. B., de Koning, A. P. J., and Pollock, D. D. (2004). Ancestral sequence reconstruction in primate mitochondrial DNA: compositional bias and effect on functional inference, *Mol. Biol. Evol.*, 21: 1871–1883.

Kruiswijk, C. P., Hermsen, T. T., Westphal, A. H., Savelkoul, H. F. J., and Stet, R. J. M. (2002). A novel functional class I lineage in zebrafish (*Danio rerio*), carp (*Cyprinus carpio*), and large barbus (*Barbus intermedius*) showing an unusual conservation of the peptide binding domains, *J. Immunol.*, 169: 1936–1947.

Kumar, S., and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution, *Nature*, 392: 917–920.

Kuper, H., and Beintema, J. J. (1976). Amino acid sequence of topi pancreatic ribonuclease, *Biochim. Biophys. Acta.*, 446: 337–344.

Lang, K., and Schmidt, F. X. (1986). Use of a trypsin pulse method to study the refolding pathway of ribonuclease, *Eur. J. Biochem.*, 159: 275–281.

Lee, J. E., and Raines, R. T. (2005). Cytotoxicity of bovine seminal ribonuclease: monomer versus dimer, *Biochemistry*, 44: 15760–15767.

Lenstra, J. A., and Beintema, J. J. (1979). Amino acid sequence of mouse pancreatic ribonuclease: extremely rapid evolutionary rates of the myomorph rodent ribonucleases, *Eur. J. Biochem.*, 98: 399–408.

Li, W. H., Gojobori, T., and Nei, M. (1981). Pseudogenes as a paradigm of neutral evolution, *Nature*, 292: 237–239.

Lorenz, E. N. (1963). Deterministic nonperiodic flow, *J. Atmos. Sci.*, 20: 130–141.

Lorenz, E. N. (1969). Predictability of a flow which possesses many scales of motion, *Tellus*, 21: 289–307.

Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes, *Science*, 290: 1151–1155.

Maddison, W. P., and Maddison, D. R. (1989). Interactive analysis of phylogeny and character evolution using the computer program MacClade, *Folia Primatol.*, 53: 190–202.

Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F., and Wilson, A. C. (1990). Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing, *Nature*, 345: 86–89.

Margoliash, E. (1963). Primary structure and evolution of cytochrome C, *Proc. Natl. Acad. Sci. U.S.A.*, 50: 672–679.

Margoliash, E. (1964). Amino acid sequence of cytochrome C in relation to its function and evolution, *Can. J. Biochem. Physiol.*, 42: 745–753.

Margulis, L., and Guerrero, R. (1995). Life as a planetary phenomenon: the colonization of Mars, *Microbiologia*, 11: 173–184.

Margulis, L., and West, O. (1993). Gaia and the colonization of Mars, *GSA Today*, 3: 277–280, 291.

Marshall, C. R., Raff, E. C., and Raff, R. A. (1994). Dollo's law and the death and resurrection of genes, *Proc. Natl. Acad. Sci. U.S.A.*, 91: 12283–12287.

Martins, E. P. (1999). Estimation of ancestral states of continuous characters: a computer simulation study, *Syst. Biol.*, 48: 642–650.

Matousek, J. (1973). The effect of bovine seminal ribonuclease (AS RNase) on cells of Crocker tumour in mice, *Experientia*, 29: 858–859.

McGovern, P. E. (2004). Fermented beverages of pre- and proto-historic China, *Proc. Natl. Acad. Sci. U.S.A.*, 101: 17593–17598.

Menon, S. T., Han, M., and Sakmar, T. P. (2001). Rhodopsin: structural basis of molecular physiology, *Physiol. Rev.*, 81: 1659–1688.

Miller, H. R. (1996). Mucosal mast cells and the allergic response against nematode parasites, *Vet. Immunol. Immunopathol.*, 54: 331–336.

Miskey, C., Izsvak, Z., Plasterk, R. H., and Ivics, Z. (2003). The Frog Prince: a reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells, *Nucleic Acids Res.*, 31: 6873–6881.

Miyazaki, J., Nakaya, S., Suzuki, T., Tamakoshi, M., Oshima, T., and Yamagishi, A. (2001). Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis, *J. Biochem.*, 129: 777–782.

Mooers, A. O., and Schluter, D. (1999). Reconstructing ancestor states with maximum likelihood: support for one- and two-rate models, *Syst. Biol.*, 48: 623–633.

Moore, S., and Stein, W. H. (1973). Chemical structures of pancreatic ribonuclease and deoxyribonuclease, *Science*, 180: 458–464.

Muskiet, F. A. J., Welling, G. W., and Beintema, J. J. (1976). Studies on primary structure of bison pancreatic ribonuclease, *Int. J. Pept. Protein Res.*, 8: 345–348.

Nambiar, K. P., Stackhouse, J., Stauffer, D. M., Kennedy, W. P., Eldredge, J. K., and Benner, S. A. (1984). Total synthesis and cloning of a gene coding for the ribonuclease S protein, *Science*, 223: 1299–1301.

Navidi, W. C., Churchill, G. A., and von Haeseler, A. (1991). Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants, *Mol. Biol. Evol.*, 8: 128–143.

Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. (1990), *Physiology of the Bacterial Cell: A Molecular Approach*, Sinauer Associates, Sunderland, Mass.

Nielsen, R. (2002). Mapping mutations on phylogenies, *Syst. Biol.*, 51: 729–739.

Nock, S., Grillenbeck, N., Ahmadian, M. R., Ribeiro, S., Kreutzer, R., and Sprinzl, M. (1995). Properties of isolated domains of the elongation factor Tu from *Thermus thermophilus* HB8, *Eur. J. Biochem.*, 234: 132–139.

O'Harra, C. C. (1930). A fossil mammal with unborn twins, *Science*, 71: 341.

Oberdorster, E., and McClellan-Green, P. (2002). Mechanisms of imposex induction in the mud snail, *Ilyanassa obsoleta*: TBT as a neurotoxin and aromatase inhibitor, *Mar. Environ. Res.*, 54: 715–718.

Ohno, S. (1970). *Evolution by Gene Duplication*, Springer-Verlag, New York.

Ohno, S., Muramoto, J., Christia, L., and Atkin, N. B. (1967). Diploid tetraploid relationship among Old World members of fish family Cyprinidae, *Chromosoma*, 23: 1–9.

Okabe, Y., Katayama, N., Iwama, M., Watanabe, H., Ohgi, K., Irie, M., Nitta, K., Kawauchi, H., Takayanagi, Y., Oyama, F., Titani, K., Abe, Y., Okazaki, T., Inokuchi, N., and Koyama, T. (1991). Comparative base specificity, stability, and lectin activity of 2 lectins from eggs of

Rana catesbeiana and *R. japonica* and liver ribonuclease from *R. catesbeiana*, *J. Biochem.*, 109: 786–790.

Omland, K. E. (1999). The assumptions and challenges of ancestral state reconstructions, *Syst. Biol.*, 48: 604–611.

Onorato, J., Scovena, E., Airaghi, S., Morandi, B., Morelli, M., Pizzi, M., and Principi, N. (1996). Role of serum eosinophil cationic protein (s-ECP), neutrophil myeloperoxidase (s-MPO) and mast cell triptase (s-TRY) in children with allergic, infective asthma and atopic dermatitis, *Riv. Ital. Pediatr.*, 22: 900–911.

Opitz, J. G. (1995). *Maximum Parsimony: Ein neuer Ansatz zum besseren Verstaendnis von Protein/Nukleinsaeure-Wechselwirkungen*, Dissertation 10952, ETH, Zurich.

Opitz, J. G., Ciglic, M. I., Haugg, M., Trautwein-Fritz, K., Raillard, S. A., Jermann, T. M., and Benner, S. A. (1998). Origin of the catalytic activity of bovine seminal ribonuclease against double-stranded RNA, *Biochemistry*, 37: 4023–4033.

Pagel, M. (1999a). Inferring the historical patterns of biological evolution, *Nature*, 401: 877–884.

Pagel, M. (1999b). The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies, *Syst. Biol.*, 48: 612–622.

Pagel, M., Meade, A., and Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies, *Syst. Biol.*, 53: 673–684.

Pauling, L., and Zuckerkandl, E. (1963). Chemical paleogenetics molecular restoration studies of extinct forms of life, *Acta Chem. Scand.*, 17: S9–S16.

Peterson, K. J., and Eernisse, D. J. (2001). Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences, *Evol. Dev.*, 3: 170–205.

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests, *Syst. Biol.*, 53: 793–808.

Presnell, S. R., and Benner, S. A. (1988). The design of synthetic genes, *Nucleic Acids Res.*, 16: 1693–1702.

Pretorius, I. S. (2000). Tailoring wine yeasts for the new millennium: novel approaches to the ancient art of winemaking, *Yeast*, 16: 675–729.

Preuss, K. D., Wagner, S., Freudenstein, J., and Scheit, K. H. (1990). Cloning of cDNA encoding the complete precursor for bovine seminal ribonuclease, *Nucleic Acids Res.*, 18: 1057.

Pupko, T., Pe'er, I., Shamir, R., and Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences, *Mol. Biol. Evol.*, 17: 890–896.

Raillard, S. A. (1993). *Veraenderung der Struktur und der biologischen Aktivitaet in RNase A mit Hilfe von gezielter Mutagenese*, Dissertation 10022, ETH, Zurich.

Ree, R. H., and Donoghue, M. J. (1999). Inferring rates of change in flower symmetry in asterid angiosperms, *Syst. Biol.*, 48: 633–641.

Richards, F. M., and Logue, A. D. (1962). Changes in absorption spectra in the ribonuclease S system, *J. Biol. Chem.*, 237: 3693–3697.

Riggs, A. (1959). Molecular adaptation in haemoglobins: nature of the Bohr effect, *Nature*, 183: 1037–1038.

Rosenberg, H. F., and Domachowske, J. B. (2001). Eosinophils, eosinophil ribonucleases, and their role in host defense against respiratory virus pathogens, *J. Leukoc. Biol.*, 70: 691–698.

Rost, B. (2001). Review: protein secondary structure prediction continues to rise, *J. Struct. Biol.*, 134: 204–218.

Runnegar, B. (2000). Loophole for snowball earth, *Nature*, 405: 403–404.

Sanangelantoni, A. M., Cammarano, P., and Tiboni, O. (1996). Manipulation of the tuf gene provides clues to the localization of sequence element(s) involved in the thermal stability of *Thermotoga maritima* elongation factor Tu, *Microbiology U.K.*, 142: 2525–2532.

Sassi, S. O. (2005). Paleogentics and the past as a key to unlock the present: The resurrection of ancestral proteins to elucidate the function of seminal ribonuclease, University of Florida, Gainesville.

Saunders, M., Wishnia, A., and Kirkwood, J. G. (1957). The nuclear magnetic resonance spectrum of ribonuclease, *J. Am. Chem. Soc.*, 79: 3289–3290.

Schaaff, I., Heinisch, J., and Zimmerman, F. K. (1989). Overproduction of glycolytic enzymes in yeast, *Yeast*, 5: 285–290.

Schluter, D. (1995). Uncertainty in ancient phylogenies, *Nature*, 377: 108–109.

Schroder, W., Mallmann, P., van der Ven, H., Diedrich, K., and Krebs, D. (1990). Cellular sensitization against spermatic and seminal plasma antigens in women after intrauterine insemination, *Arch. Gynecol. Obstet.*, 248: 67–74.

Schultz, T. R., and Churchill, G. A. (1999). The role of subjectivity in reconstructing ancestral character states: a Bayesian approach to unknown rates, states, and transformation asymmetries, *Syst. Biol.*, 48: 651–664.

Schultz, T. R., Cocroft, R. B., and Churchill, G. A. (1996). The reconstruction of ancestral character states. *Evolution*, 50: 504–511.

Segel, I. H. (1975). *Enzyme Kinetics*, Wiley, New York.

Shagin, D. A., Barsova, E. V., Yanushevich, Y. G., Fradkov, A. F., Lukyanov, K. A., Labas, Y. A., Semenova, T. N., Ugalde, J. A., Meyers, A., Nunez, J. M., Widder, E. A., Lukyanov, S. A., and Matz, M. V. (2004). GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity, *Mol. Biol. Evol.*, 21: 841–850.

Shi, Y., and Yokoyama, S. (2003). Molecular analysis of the evolutionary significance of ultraviolet vision in vertebrates, *Proc. Natl. Acad. Sci. U.S.A.*, 100: 8308–8313.

Singhania, N. A., Dyer, K. D., Zhang, J., Deming, M. S., Bonville, C. A., Domachowske, J. B., and Rosenberg, H. F. (1999). Rapid evolution of the ribonuclease A superfamily: adaptive expansion of independent gene clusters in rats and mice, *J. Mol. Evol.*, 49: 721–728.

Sorrentino, S., Carsana, A., Furia, A., Doskocil, J., and Libonati, M. (1980). Ionic control of enzymic degradation of double-stranded RNA, *Biochim. Biophys. Acta*, 609: 40–52.

Soucek, J., Chudomel, V., Potmesilova, I., and Novak, J. T. (1986). Effect of ribonucleases on cell-mediated lympholysis reaction and on GM-CFC colonies in bone marrow culture, *Nat. Immun. Cell Growth Regul.*, 5: 250–258.

Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P., and Benner, S. A. (1990). The ribonuclease from an extinct bovid ruminant, *FEBS Lett.*, 262: 104–106.

Strachan, R. G., Paleveda, W. J., Nutt, R. F., Vitali, R. A., Veber, D. F., Dickinson, M. J., Garsky, V., Deak, J. E., Walton, E., Jenkins, S. R., Holly, F. W., and Hirschma, R. (1969). Studies on total synthesis of an enzyme. 2. Synthesis of a protected tetratetracontapeptide corresponding to 21–64 sequence of ribonuclease A, *J. Am. Chem. Soc.*, 91: 503–504.

Stroband, H. W. J., Stevens, C., Kronnie, G. T., Samallo, J., Schipper, H., Kramer, B., and Timmermans, L. P. M. (1995). Expression of carp Cdx1, a caudal homolog, in embryos of the carp, *Cyprinus carpio*, *Rouxs Arch. Dev. Biol.*, 204: 369–377.

Strydom, D. J., Fett, J. W., Lobb, R. R., Alderman, E. M., Bethune, J. L., Riordan, J. F., and Vallee, B. L. (1985). Amino acid sequence of human tumor derived angiogenin, *Biochemistry*, 24: 5486–5494.

Stryer, L. (1995). *Biochemistry*, W.H. Freeman, New York.

Sun, G. (2002). Archaefructaceae, a new basal angiosperm family, *Science*, 296: 899–904.

Sun, H. M., Rodin, A., Zhou, Y. H., Dickinson, D. P., Harper, D. E., HewettEmmett, D., and Li, W. H. (1997). Evolution of paired domains: isolation and sequencing of jellyfish and hydra *Pax* genes related to *Pax-5* and *Pax-6*, *Proc. Natl. Acad. Sci. U.S.A.*, 94: 5156–5161.

Sun, H., Merugu, S., Gu, X., Kang, Y. Y., Dickinson, D. P., Callaerts, P., and Li, W. H. (2002). Identification of essential amino acid changes in paired domain evolution using a novel combination of evolutionary analysis and *in vitro* and *in vivo* studies, *Mol. Biol. Evol.*, 19: 1490–1500.

Suzuki, Y., Glazko, G. V., and Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics, *Proc. Natl. Acad. Sci. U.S.A.*, 99: 16138–16143.

Swofford, D. L. (1998). *PAUP*: Phylogenetic Analysis Using Parsimony*, Version 4.

Swofford, D. L. (2001). *PAUP*: Phylogenetic Analysis Using Parsimony* (**and Other Methods*), Version 4, Sinauer Associates, Sunderland, Mass.

Swofford, D. L., Olsen, G. J. Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic inference, in *Molecular Systematics*, Sinauer Associates, Sunderland, Mass. pp 407–514.

Tauer, A., and Benner, S. A. (1997). The B-12 dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*: an evolutionary solution to the ribonucleotide reductase conundrum, *Proc. Natl. Acad. Sci. U.S.A.*, 94: 53–58.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22: 4673–4680.

Thomson, J. M. (2002). *Interpretive proteomics: experimental paleogenetics as a tool to analyze function and discover pathways in yeast*, dissertation, University of Florida.

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., and Benner, S. A. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast, *Nat. Genet.*, 37: 630–635.

Thornton, J. W. (2001). Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions, *Proc. Natl. Acad. Sci. U.S.A.*, 98: 5671–5676.

Thornton, J. W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules, *Nat. Rev. Genet.*, 5: 366–375.

Thornton, J. W., and DeSalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics, *Annu. Rev. Genomics Hum. Genet.*, 1: 41–73.

Thornton, J. W., Need, E., and Crews, D. (2003). Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling, *Science*, 301: 1714–1717.

Trabesinger-Ruef, N., Jermann, T., Zankel, T., Durrant, B., Frank, G., and Benner, S. A. (1996). Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? *FEBS Lett.*, 382: 319–322.

Trautwein, K. (1991). *Construction of an Improved Expression System for Bovine Pancreatic Ribonuclease A and Construction and Characterization of RNase A Mutants*, Dissertation 9613, ETH, Zurich.

Ugalde, J. A., Chang, B. S., and Matz, M. V. (2004). Evolution of coral pigments recreated, *Science*, 305: 1433.

Underhill, D. A. (2000). Genetic and biochemical diversity in the *Pax* gene family, *Biochem. Cell Biol.*, 78: 629–38.

Underhill, D. A., Vogan, K. J., and Gros, P. (1995). Analysis of the mouse Splotch delayed mutation indicates that the *Pax-3* paired domain can influence homeodomain DNA binding activity, *Proc. Natl. Acad. Sci. U.S.A.*, 92: 3692–3696.

Vandenberg, A., Vandenhendetimmer, L., and Beintema, J. J. (1976). Isolation, properties and primary structure of coypu and chinchilla pancreatic ribonuclease, *Biochim. Biophys. Acta.*, 453: 400–409.

Vandijk, H., Sloots, B., Vandenberg, A., Gaastra, W., and Beintema, J. J. (1976). Primary structure of muskrat pancreatic ribonuclease, *Int. J. Pept. Protein Res.*, 8: 305–316.

Veber, D. F., Varga, S. L., Milkowski, J. D., Joshua, H., Conn, J. B., Hirschma, R., and Denkewal, R. G. (1969). Studies on total synthesis of an enzyme. 4. Some factors affecting conversion of protected S protein to ribonuclease S, *J. Am. Chem. Soc.*, 91: 506–507.

Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001). The sequence of the human genome, *Science*, 291: 1304–1351.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea, *Science*, 304: 66–74.

Vos, J. P., Lopes-Cardozo, M., and Gadella, B. M. (1994). Metabolic and functional aspects of sulfogalactolipids, *Biochim. Biophys. Acta.*, 1211: 125–149.

Wade, N. (1995). Method and madness: dead sure, *New York Times*, New York, pp. 35–36.

Weiser, K. C., and Justice, M. J. (2005). Cancer biology: Sleeping Beauty awakens, *Nature*, 436: 184–186.

Welling, G. W., Groen, G., and Beintema, J. J. (1975). Amino acid sequence of dromedary pancreatic ribonuclease, *Biochem. J.*, 147: 505–511.

Welling, G. W., Mulder, H., and Beintema, J. J. (1976). Allelic polymorphism in arabian camel ribonuclease and amino acid sequence of bactrian camel ribonuclease, *Biochem. Genet.*, 14: 309–317.

Wilkie, S. E., Robinson, P. R., Cronin, T. W., Poopalasundaram, S., Bowmaker, J. K., and Hunt, D. M. (2000). Spectral tuning of avian violet- and ultraviolet-sensitive visual pigments, *Biochemistry*, 39: 7895–7901.

Wills, C. (1976). Production of yeast alcohol dehydrogenase isoenzymes by selection, *Nature*, 261: 26.

Woese, C. R. (1987). Bacterial evolution, *Microbiol. Rev.*, 51: 221–271.

Wolfe, K. H., and Shields, D. C. (2001). Molecular evidence for an ancient duplication of the entire yeast genome, *Nature*, 387: 708–713.

Xu, W., Rould, M. A., Jun, S., Desplan, C., and Pabo, C. O. (1995). Crystal structure of a paired domain–DNA complex at 2.5 Å resolution reveals structural basis for *Pax6* developmental mutations *Cell*, 80: 639–650.

Xu, H. E., Rould, M. A., Xu, W., Epstein, J. A., Maas, R. L., and Pabo, C. O. (1999). Crystal structure of the human *Pax6* paired domain–DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding, *Genes Dev.*, 13: 1263–1275.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput. Appl. Biosci.*, 13: 555–556.

Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences, *Genetics*, 141: 1641–1650.

Yant, S. R., Park, J., Huang, Y., Mikkelsen, J. G., and Kay, M. A. (2004). Mutational analysis of the N-terminal DNA-binding domain of Sleeping Beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells, *Mol. Cell. Biol.*, 24: 9239–9247.

Yasutake, Y., Watanabe, S., Yao, M., Takada, Y., Fukunaga, N., and Tanaka, I. (2003). Crystal structure of the monomeric isocitrate dehydrogenase in the presence of $NADP^+$: insight into the cofactor recognition, catalysis, and evolution, *J. Biol. Chem.*, 278: 36897–36904.

Yokoyama, S., Radlwimmer, F. B., and Blow, N. S. (2000). Ultraviolet pigments in birds evolved from violet pigments by a single amino acid change, *Proc. Natl. Acad. Sci. U.S.A.*, 97: 7366–7371.

Zhang, J. Z., and Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods, *J. Mol. Evol.*, 44: S139–S146.

Zhang, J. Z., and Rosenberg, H. F. (2002). Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates, *Proc. Natl. Acad. Sci. U.S.A.*, 99: 5486–5491.

Zhang, J., Rosenberg, H. F., and Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes, *Proc. Natl. Acad. Sci. U.S.A.*, 95: 3708–3713.

Zhang, J. Z., Dyer, K. D., and Rosenberg, H. F. (2002). RNase 8, a novel RNase A superfamily ribonuclease expressed uniquely in placenta, *Nucleic Acids Res.*, 30: 1169–1175.

Zhang, J. Z., Dyer, K. D., and Rosenberg, H. F. (2003). Human RNase 7: a new cationic ribonuclease of the RNase A superfamily, *Nucleic Acids Res.*, 31: 602–607.

Zhao, W., Kote-Jarai, Z., van Santen, Y., Hofsteenge, J., and Beintema, J. J. (1998). Ribonucleases from rat and bovine liver: purification, specificity and structural characterization, *Biochim. Biophys. Acta.*, 1384: 55–65.

Zhu, G., Golding, G. B., and Dean, A. M. (2005). The selective cause of an ancient adaptation, *Science*, 307: 1279–1282.

Zilliacus, J., Carlstedtduke, J., Gustafsson, J. A., and Wright, A. P. H. (1994). Evolution of distant DNA binding specificities within the nuclear receptor family of transcription factors, *Proc. Natl. Acad. Sci. U.S.A.*, 91: 4175–4179.