CHAPTER 2

Ancestral sequence reconstruction as a tool to understand natural history and guide synthetic biology: realizing and extending the vision of Zuckerkandl and Pauling

Eric A. Gaucher

2.1 Historical context

The recent accumulation of DNA sequence data, combined with advances in evolutionary theory and computational power, have paved the way for innovative approaches to understanding the origins, evolution, and distribution of life and its constituent biomolecules (Pauling and Zuckerkandl, 1963; Benner *et al*., 2002; Gaucher *et al*., 2004). One approach to understanding ancestral states follows a present-day-backwards strategy, whereby genomic sequences from extant (modern) organisms are incorporated into evolutionary models that estimate the extinct (ancient) character states of genes no longer present on Earth (Fitch, 1971; Shih *et al*., 1993; Benner, 1995; Koshi and Goldstein, 1996; Schultz *et al*., 1996; Cunningham, 1999; Omland, 1999; Pagel, 1999; Schultz and Churchill, 1999; Chang and Donoghue, 2000; Thornton, 2004; Hall, 2006). These inferred ancestral gene sequences act as hypotheses that can be tested in the laboratory through the resurrection of the ancestral proteins themselves (paleogenetics). Results from functional assays of the protein products from these ancient genes permit us to accept or reject null hypotheses about the sequences themselves, or about their interactions, binding specificities, environments, etc. And beyond narratives relating ancient phenotypes and environments, paleogenetics provides the unique opportunity to "replay" the molecular tape of life (Gould, 1989).

It is probably appropriate that any discussion of ancestral sequence reconstruction be placed within a historical framework. By doing so in this chapter, I hope to convey how the field has made recent advances, and emphasize the need to continue this progress so that our concepts enjoy wider recognition and create greater impact than the achievements to date. The chapter will begin with a brief account of the field during the 1990s. I will continue with a discussion on one of the resurrected gene families that has been studied, and then conclude with a discussion on future directions in the field, with a particular emphasis on evolutionary synthetic biology.

In 1963, Emile Zuckerkandl and Linus Pauling published an intriguing article entitled "Chemical paleogenetics molecular restoration studies of extinct forms of life" (Pauling and Zuckerkandl, 1963). In it, they put forward the notion of reconstructing amino acid sequences of ancestral proteins by virtue of a comparison between sequences of related proteins found in contemporary organisms and subsequent synthesis (and thereby resurrection) of these sequences in the laboratory. While limits in technology prohibited the actual resurrection, Zuckerkandl and Pauling presented a sequence reconstruction of ancient mammalian

hemoglobins. The duo then suggested that a future resurrection of ancient hemoglobins assayed for dioxygen affinity and pH dependence would generate higher-order inferences of biological interpretation. Or, more broadly, the joining of chemical, biological, and structural models to natural history would provide a more accurate description of macromolecular behavior beyond that supplied by studying individual molecules disconnected from the selective forces governing their evolution (this approach is also termed planetary biology).

It is important to note the significance of Zuckerkandl and Pauling's proposal. Molecular reconstructions and experimental resurrections are fundamentally exercises in theory/computation/algorithmic development and technological advances, respectively. Proposing these approaches would have been sufficiently significant. But Zuckerkandl and Pauling realized the potential of creating a paradigm that attempted to connect chemical structure to natural selection at the molecular, cellular, organismal, population, and planetary levels.

The first examples of molecular resurrections were performed on artiodactyl pancreatic ribonucleases and lysozymes (Malcolm *et al*., 1990; Stackhouse *et al*., 1990; Jermann *et al*., 1995; see Chapters 1 and 18 in this volume for further discussion). The former study was particularly notable in that it not only put Zuckerkandl and Pauling's theory into practice, it did so in a manner that connected chemical reactivity (RNA hydrolysis), molecular biology (single- versus double-stranded RNA-binding affinities), organismal evolution (bacterial fermentation and foregut digestion in ruminants), and planetary biology (diversification of grasses during the Oligocene cooling, and the ability of artiodactyls to digest and extract nutrients from these grasses via bacterial fermentation). These connections were the first to exemplify Zuckerkandl and Pauling's proposed paradigm, and by the late 1990s we were intent on extending the paradigm as far back in history as possible—ideally to the earliest life forms on Earth.

Successful studies on ancestral reconstructions require a sufficient base of knowledge in evolutionary theory and models. Computational simulations and experimental phylogenetics were used to differentiate accuracy and consistency between distance-based (e.g. neighbor-joining) and character-based (e.g. parsimony, likelihood, and Bayesian) approaches to phylogenetics. Further, several discussions about reconstructing ancestral character states were also appearing in the literature during this time. For instance, a symposium lead by Cunningham, Oakley, Omland, and others on behalf of the Society of Systematic Biology focused on this topic (Omland, 1999). Meanwhile, Pagel, Goldstein, Yang, and others were independently developing maximum-likelihood approaches based on Bayesian statistics (called ML or empirical Bayesian reconstruction; Yang *et al*., 1995; Koshi and Goldstein, 1996; Pupko *et al*., 2000). These authors advocated the use of maximum likelihood over parsimony for inferring ancestral states (see Pagel *et al*., 2004, and Schluter, 1995, for a criticism of previous work on resurrected proteins). Similarly, Ronquist, Huelsenbeck, Schultz, and others were developing hierarchical Bayesian algorithms for character-state inferences (Schultz and Churchill, 1999; Huelsenbeck and Bollback, 2001; Ronquist, 2004). A fuller discussion of methods for ancestral sequence reconstruction is presented in Chapters 4–9 in this volume.

Regardless of the growing literature, it remained unclear at the time which approach would best guide an experimental design of ancestral sequences. Maximum likelihood's superior performance under certain models of evolution was apparent from numerous simulation studies and laboratory evolution experiments with viruses (Bull *et al*., 1993; Zhang and Nei, 1997). Hierarchical Bayesian methods could theoretically outperform likelihood approaches, although only Yang's empirical approach was available to us at the time. Ideally, a hierarchical Bayesian approach accounts for uncertainty from phylogenetic hypotheses (these include uncertainty in the topology, branch lengths, and any other parameter estimates). A heuristic work-around, however, could take advantage of an empirical Bayesian approach in which alternative models, parameters, and topologies could be analyzed separately. The output from these separate analyses could then be

combined with the expectation of resolving uncertainty or bias arising from any of the individual analyses alone.

The heuristic hierarchical Bayesian approach was initially applied to the alcohol dehydrogenase (ADH) gene family (the same approach was later applied to elongation factors (EFs) and seminal ribonucleases). Ancestral character states were inferred from the ADH phylogeny using multiple models each for DNA-, codon-, and amino acid-based evolution, all for two competing topologies (Thomson *et al*., 2005). The results from these separate analyses required that 12 putative ancestral genes be synthesized to account for the differences (or uncertainty) at individual sites among the various analyses. Although this approach did not eliminate all uncertainty and bias associated with the ancestral reconstruction methods, it was an improvement in the ability to capture the true ancestral state.

During our gene synthesis experiments of ancestral ADHs, Chang and colleagues had formulated a similar heuristic hierarchical Bayesian approach towards the resurrection of ancestral rhodopsins (Chang *et al*., 2002), and later with fluorescent proteins (Ugalde *et al*., 2004). Although their analysis did not account for alternative topologies, they did consider various models of sequence evolution and their effects on the ancestral state inferences.

We should note that the choice of individual amino acid residues at sites in an ancestral sequence remains under debate to this day. For instance, Pollock, Goldstein, and colleagues have suggested that selecting the ancestral residue with the highest probability at each site (most-probabilistic ancestral sequence, MPAS) can lead to erroneous inferences of the ancestral states (Williams *et al*., 2006; see Chapter 8). These authors advocate the synthesis of random sequences weighted from the posterior distribution. For instance, consider a pentapeptide in which each site in the ancestral sequence has a posterior probability for alanine of 80% and serine of 20%. The MPAS would be a penta-alanine (i.e. AAAAA), whereas the average weighted sample from the distribution would generate four alanines and one serine (e.g. ASAAA).

Recent discussions over which of these two sequences introduces less bias during the reconstruction procedure have resulted in lively exchanges, and will undoubtedly continue to do so for the next few years. The center issue remains unresolved: under what phylogenetic and evolutionary conditions will the two sequences be biased and thus no longer able to capture the true ancestral behavior/function? To what extent do sequence divergence, episodes of adaptive evolution (e.g. high nonsynonymous/synonymous ratios or heterotachy), heterogeneous processes, etc., lead to bias during the reconstruction process?

For instance, consider an ancient gene-duplication event resulting in neofunctionalization whereby the novel behavior is determined by amino acid replacements at two sites only. Each of the two sites experiences an alanine-to-serine replacement along the branch leading to the new function. Reconstructing the ancestral sequence at the node representing the common ancestor of the paralogs infers alanine at 80% and serine at 20%. The MPAS would correctly have alanines at these two positions. Sequences sampled from the posterior distribution, however, would only have alanines at these two positions 64% of the time (0.8×0.8). Therefore, one-third of the sampled sequences would have a behavior not analogous to the ancestral behavior, making it nearly impossible to generate accurate interpretations from the sampled sequences.

Alternatively, lack of phylogenetic signal and parallel/convergent evolution may cause the MPAS to be biased away from the true ancestral behavior. Here, ancestral states can be driven by the evolutionary models themselves when the data fail to provide sufficient signal to extract the true ancestral states. Any bias associated with the evolutionary models (e.g. preponderance of hydrophobic residues) may be propagated throughout the inference process, ultimately influencing the inferred protein behavior away from the true ancestral state. We tend to expect this for sites that are more rapidly evolving (e.g. coils on the protein surface) although simulations are required to confirm this notion. We anticipate that many of these issues will be addressed in the near future through extensive computational

simulations that exploit biologically realistic models of protein evolution as well as experimental phylogenetics studies.

By no means do the above paragraphs serve as a definitive account of the field. Many other people were involved in the development of reconstruction theory, evolutionary models, algorithms, and software development. Rather, it serves as one historical account intended to convey the dynamic growth occurring within the field.

In fact, approximately 20 narratives have emerged to date where specific molecular systems from extinct organisms have been resurrected for study in the laboratory (Sassi *et al*., 2007). These include digestive proteins (ribonucleases, proteases, and lysozymes) in ruminants and primates to illustrate how digestive function arose from non-digestive 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 different environments, steroid hormone receptors adapting to changing function in steroid-based regulation of metazoans, fluorescent proteins from ocean-dwelling invertebrates, enzyme cofactor evolution, and proteins from very ancient bacteria helping to define environments where the earliest forms of bacterial life lived.

2.2 Temperature conditions of early life

By the end of the last decade, the most ancient paleomolecular resurrections had traveled back in time only *c*.200–300 million years. This had left untouched many of the most widely discussed questions about the nature of early life on Earth. We identified and concluded that the EF-Tu family had the greatest potential to address these questions. Further, this gene family could generate a robust statistical reconstruction in conjunction with biological interpretations (correlation between protein thermostability and optimal growth temperature of the host organism) and planetary integrations (correlating the temperature histories of early life from molecular and geologic records; Gaucher *et al*., 2003).

2.3 EFs

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 robustly reconstructed for proteins from deep nodes in the Bacterial phylogeny. 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 20–40, 40–80, and $>80^{\circ}$C, respectively, and represented by species of *Escherichia*, *Thermus*, and *Thermotoga*, have temperature optima in their respective ranges. 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.

Figures 2.1a and 2.1b show 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, 3–20, did not affect the amino acid reconstructions at these nodes: ML-stem (maximum-likelihood tree for the stem bacteria) and Alt-stem (alternative tree for the stem bacteria). The ancestral sequence at the node representing the most recent common ancestor of only mesophilic bacterial lineages was also reconstructed, and named ML-meso (maximum-likelihood tree for mesophiles only). This node captures one feature of models that have concluded that the last common ancestor of the Bacteria was mesophilic. In all, these reconstructed ancestral sequences did not appear to be influenced by long-branch attraction or non-homogeneous modes of molecular evolution, such as changes in the mutability of individual sites in different branches of the bacterial subtree (see Chapter 9 for further discussion of this).

A BLAST search was then performed to identify the most similar extant sequences to the inferred ancestors. ML-stem and Alt-stem are most similar to the sequences of EFs from *Thermoanaerobacter tengcongensis* (a thermophile) and *Thermotoga*

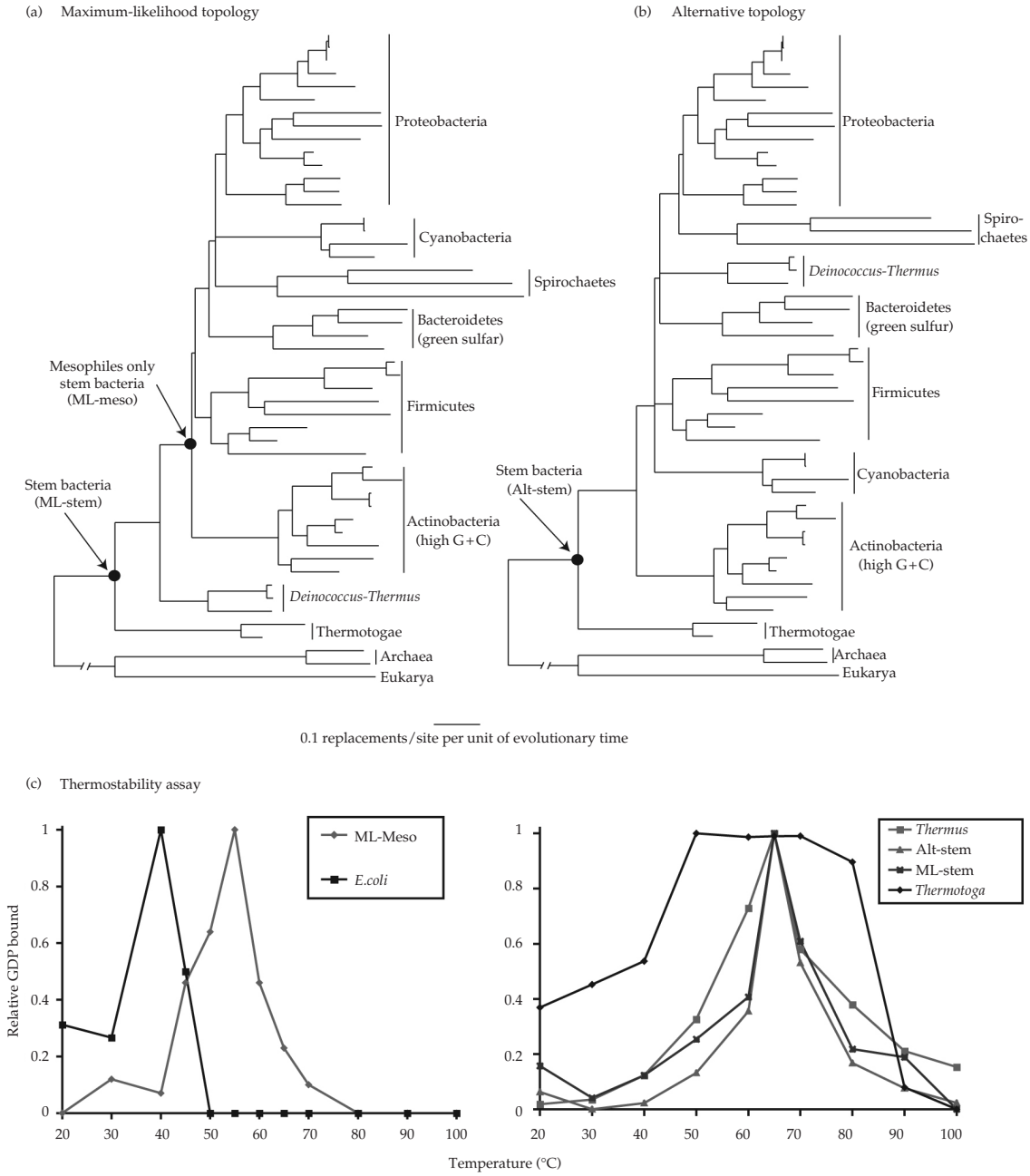

Figure 2.1 The two un-rooted universal trees used to reconstruct ancestral bacterial sequences. Archaea and Eukarya serve to provide a node within the bacterial subtree from which ancient sequences can be inferred. Lineages containing thermophiles (as known at the time of the original study) are highlighted in Italic. (a) Maximum-likelihood topology used to reconstruct the stem elongation factors (EFs) from 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 elongation factors from bacteria (Alt-stem). (c) GDP-binding assay to test thermostability of ancestral and modern EF proteins. The amount of tritium-labeled GDP bound at 0°C 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. EF thermostability profiles for *Escherichia coli*, *Thermus aquaticus*, and *Thermotoga maritima* are shown with the three ancestral EF profiles. Reprinted from Gaucher *et al.* (2003) Inferring the paleoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425**: 285–288.

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). We expect that the identities of these best-hits will change as the databases themselves expand. If we had assumed, however, that similarity in sequence implies similarity in thermostability, it might have been predicted that the stem bacterium was thermophilic or hyperthermophilic, whereas the ancestral node constructed without considering thermophiles was a mesophile. To test these predictions based on this under-substantiated assumption, genes encoding the ancestral sequences were synthesized, expressed in an *Escherichia 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 (Figure 2.1c).

Each resurrected protein behaved similarly. Both ML-stem and Alt-stem EFs bound GDP with a temperature profile similar to that of the thermophilic EF from modern *Thermus aquaticus*, with optimal binding at approximately 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 *T. maritima*, which is maximally active up to approximately 85°C and has a broad optimal temperature range typical of hyperthermophiles. The observation that the amino acid sequences of ML-stem and Alt-stem shared only 93% identity but displayed the same thermostability profiles suggests that inferences of this ancestral property are robust with respect to both varying topologies and ancestral character-state predictions. Based on these given evolutionary models, this suggests that the paleoenvironment of the ancient bacterium was approximately 65°C.

Inferences were then drawn from a resurrected EF whose sequence was reconstructed from the last common ancestral sequences of contemporary organisms which, 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 2.1c). This result shows that the behavior of an ancestor need not be an average of the behaviors of its descendants.

The observation that tree-based ancestral sequence reconstructions can give results different from consensus-sequence reconstructions may be general. It underscores a fact, well known in Structure Theory, 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.

A short side note on the names of the ancestral EF proteins used in this study: an examination of the GenBank entries for ML-stem, Alt-stem and ML-meso will reveal their alternative names: BLANKET1–3. This acronym for bacterial-lineage ancestral reconstructions served as a small tribute to Pauling. It remains to be determined, however, if this was indeed a service to his name.

2.4 Conclusions from EF studies

This study pushed the experimental paleogenetics research strategy back to at least 3 billion years, to the most primitive ancestors from which descendants can be traced. Accordingly, the ambiguity encountered is substantial, and available sequence data may not be sufficient to manage it convincingly. Here, the ambiguities do not depend primarily on the details of the model used to infer ancestral states. They seem to arise rather from the uncertainty of the phylogenetic tree joining the protein family members.

It is noteworthy, therefore, that reconstructions can be made at all. Further, if the large-scale sequencing of random bacterial genomes, as undertaken by Venter and his group (Venter *et al.*, 2004), continues, there is good reason to hope that the reconstructions will improve. Indeed, the temperature history of Bacteria is already beginning to be defined by follow-up studies.

Adding interpretations from the geologic and molecular records to the results of ancestral EF proteins provides the necessary planetary

integrations to fulfill the goals of Zuckerkandl and Pauling's molecular-restoration paradigm. Here, geologic evidence based on low $\delta^{18}O$ isotopic ratios in 3.5–3.2 billion-year-old cherts from the Barberton greenstone belt in South Africa suggests that the ocean temperature of early Earth was 55–85°C (Knauth, 2005). This result is remarkably consistent with the inferred optimal growth temperature of the microorganisms hosting the ancestral EF proteins. Inferences from the evolution of amino acid frequencies, as well as other studies, suggest that the last universal ancestor lived in a hot environment, further supporting the thermophily notion of early life (see Chapter 17).

It is important to note that these studies *do not* address issues related to the origins of life. The origins of bacteria and the last universal ancestor were undoubtedly complex microorganisms far removed from life's origins. These studies do, however, provide clues to the environment that hosted life's most recent common ancestors.

2.5 Ancestral resurrections and evolutionary synthetic biology

Zuckerkandl and Pauling's proposal to connect chemical structure and reactivity to cellular, organismal, and planetary interpretations using resurrected ancestral sequences was remarkably prophetic. Nearly all implementations of this proposal have followed the natural history paradigm laid out in Zuckerkandl and Pauling's seminal discussion (1963), and will undoubtedly soon be extended to developmental biology (Colosimo *et al*., 2005). But exceptions to this paradigm exist. Most notable is the application of resurrected proteins to connect chemical and molecular reactivity to human health by Kodra, Liberles, and coworkers (*truly* applied molecular evolution; Skovgaard *et al*., 2006; see Chapter 3). The driving goal in their study is not to elucidate the evolutionary history of the insulin-response pathway per se, but rather to search ancestral functional sequence-space in hopes of developing innovative therapeutics. Regardless of the outcome from this study, the prospects of similar research aimed at integrating molecular evolution and biomedicine using ancestral resurrections are highly energizing. An additional application of ancestral reconstruction tools for biomedicine consists of generated predictions regarding the tolerability of amino acid replacements from human single-nucleotide-polymorphism data for disease-causing genes (Gaucher *et al*., 2006).

The next logical extension of molecular resurrections beyond natural history and biomedicine is to biotechnology and synthetic biology. Not surprisingly, synthetic biology means different things to different scientific disciplines (Benner and Sismour, 2005; Endy, 2005). Surprisingly, however, biologists seem to have taken a back seat to chemists and engineers in the development of this field. It seems apparent that synthetic biology would stand to benefit if so-called molecular reconstructionists contributed to its progress. In this way, an evolutionary synthetic biology is formed. A few examples come to mind: synthetic DNA/protein libraries, tools for gene integration, cellular machines, and recombinant genomes.

2.5.1 Synthetic DNA/protein libraries

As the cost of DNA-synthesis technology diminishes, the constraint of resurrecting a few ancestral nodes on a phylogeny or a few variants at any node is released. Suppose, for instance, that the novel protein identified by Skovgaard *et al.* (2006) only existed in sequence space that was slightly removed from any of the given nodes connecting humans and gila monsters. Ancestral resurrections of these particular nodes could have failed to reveal a protein with interesting properties. An exercise in directed evolution intended to capture phylogenetic information, however, may be more successful in revealing the desired properties.

Directed evolution is a powerful technique for improving the activity, specificity and/or stability of proteins. It has been applied to a wide range of proteins with uses in therapeutics, agriculture, and chemistry.

My research group have developed a new method for designing libraries to include the functional diversity of large, highly divergent protein families while still maintaining a high proportion of active members. The strategy is to include only those changes that are associated with

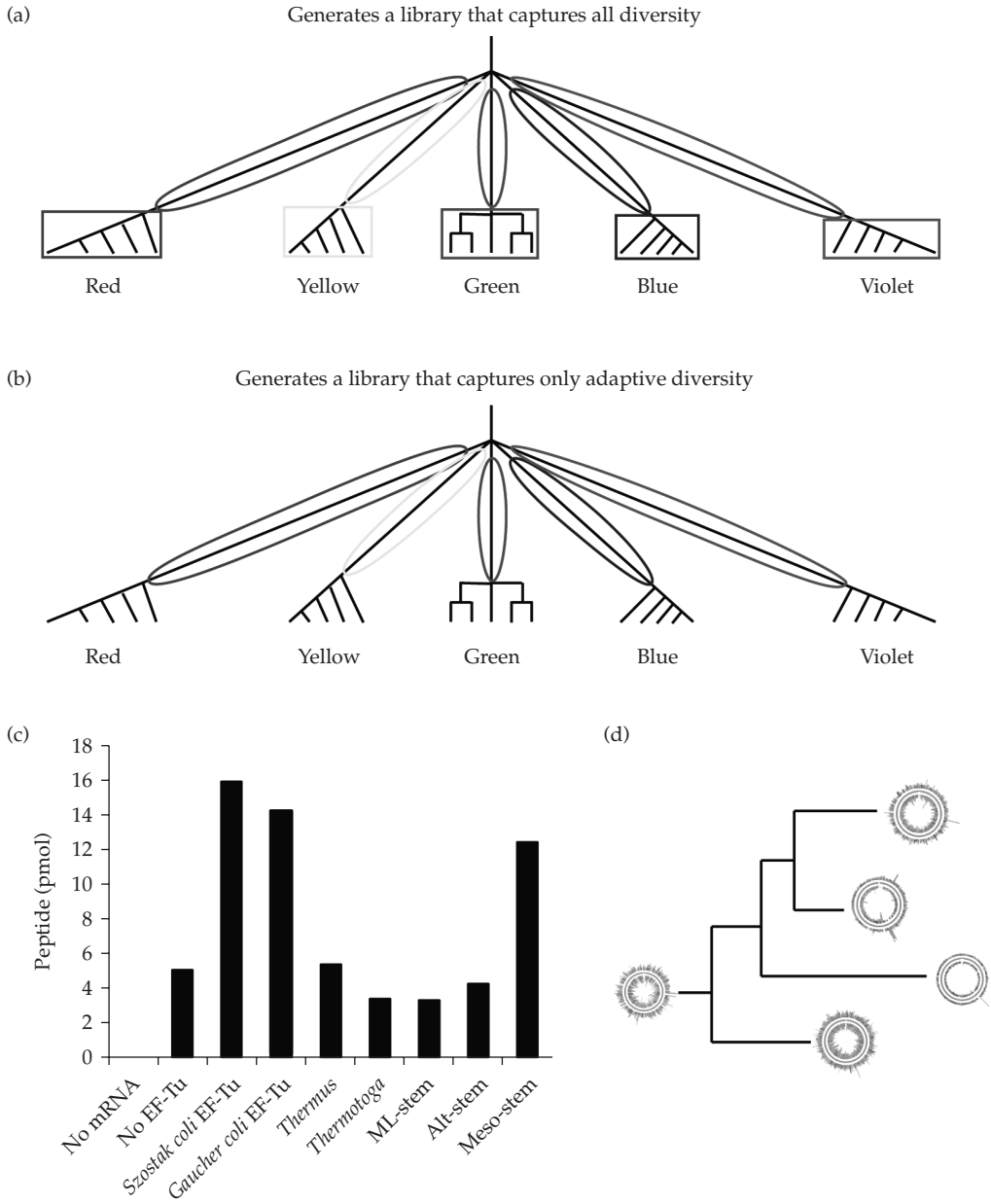

Figure 2.2 Ancestral sequence reconstruction as a tool for synthetic biology. (a) Accepted-diversity approach for the synthesis of DNA libraries intended to expand protein functionality. A library built using the approach integrates all the diversity found within extant sequences. (b) REAP approach for library synthesis. Note that a library built using this approach attempts to capture only sequence diversity responsible for differences between the five fluorescent subfamilies. (c) The amount of peptide synthesized from a reconstituted *in vitro* translation system that makes use of modern and ancestral EF-Tu proteins. *E. coli* EF-Tu was purified from both the Jack Szostak (Harvard Medical School) and Gaucher laboratories to compare purification protocols. (d) Resurrection of the ancestral genome.

new activities during the evolution of a protein family and to exclude the much larger set of changes that do not lead to new functions (phenotypically neutral). In other words, the approach identifies the amino acid changes that were associated with functional divergence during the evolutionary history of a family and builds libraries containing only those specific changes. The new method is referred to as reconstructing evolutionary adaptive paths (REAP).

To illustrate this strategy, consider a hypothetical case of fluorescent proteins. There are five families with individual fluorescent spectra. Each family contains five sequences and all five families share a common ancestor (polytomy). We would like to generate a fluorescent protein with novel properties by incorporating information contained within these 25 sequences.

The standard approach (accepted diversity) takes advantage of processes governed by natural selection. Here, amino acids present in modern (extant) sequences are combinatorially used to generate a library. Highlighted in Figure 2.2a, patterns of amino acid residues that evolved either within a family (boxed regions) or along the branches that gave rise to the individual families (circled regions) are recorded. Most of the amino acid patterns observed in modern proteins presumably arose within the families (boxed regions) and thus have little to offer in terms of generating novel properties. These amino acid patterns arose mostly from neutral evolution, assuming a lack of selective pressure to diversify within a given family.

The REAP approach is based on a model of molecular evolution that attempts to eliminate amino acid patterns predicted to have minimal contributions to novel protein behaviors. This is achieved by reconstructing the ancestral patterns that arose during the evolution of 'unique' properties compared to the last common ancestor of the fluorescent proteins (circled branches only), and ignoring the amino acid patterns that arose within a family (see Figure 2.2b). In doing so, we increase the unique behaviors captured using the accepted-diversity approach, while decreasing the noise associated with its approach.

The REAP methodology is not predicted to be ideal for all library designs. The approach requires numerous homologous sequences to generate an articulated phylogeny. Further, the phylogeny needs to represent a family of sequences with diverse behaviors, otherwise the extracted amino acid patterns may not generate novel behaviors. When the appropriate information is available, however, REAP is predicted to have a number of significant advantages including the incorporation of information from highly diverse family members into a highly *functional* library and no requirements for information regarding protein structure or the mutability of sites (mutagenesis experiments) to guide the library design.

We anticipate that the REAP approach will make substantial contributions to synthetic biology. For instance, DNA libraries based on the REAP method may generate polymerases with high fidelity towards non-standard nucleotides and/or protein variants capable of supporting unnatural amino acid incorporation during protein synthesis. The resulting biopolymers will then serve as the information (novel coding systems) and catalysis (novel side-chain chemistry) components of an expanded biology.

2.5.2 Ancient transposons as tools for gene integration

Transposable elements are pieces of DNA that have the ability to jump around an organism's genome. In their simplest form, these mobile elements code for a protein that excises their encoded DNA from one chromosomal position and reintegrates it at different location in the genome. Terminal repeats flank the coding sequence and provide the necessary regulatory information for transposition. The mobile properties of these elements have considerable potential as tools for molecular biology and therapeutic gene delivery.

For instance, a member of the Tc1/mariner family has been resurrected from salmonid fish pseudogenes using a consensus approach, as opposed to a true phylogenetic reconstruction of ancestral states (Figure 2.3a; Ivics *et al*., 1997). The aptly named Sleeping Beauty (SB) element was shown to be active in transposition, and, equally important, it was active in non-fish species. Follow-up mutagenesis studies have produced a

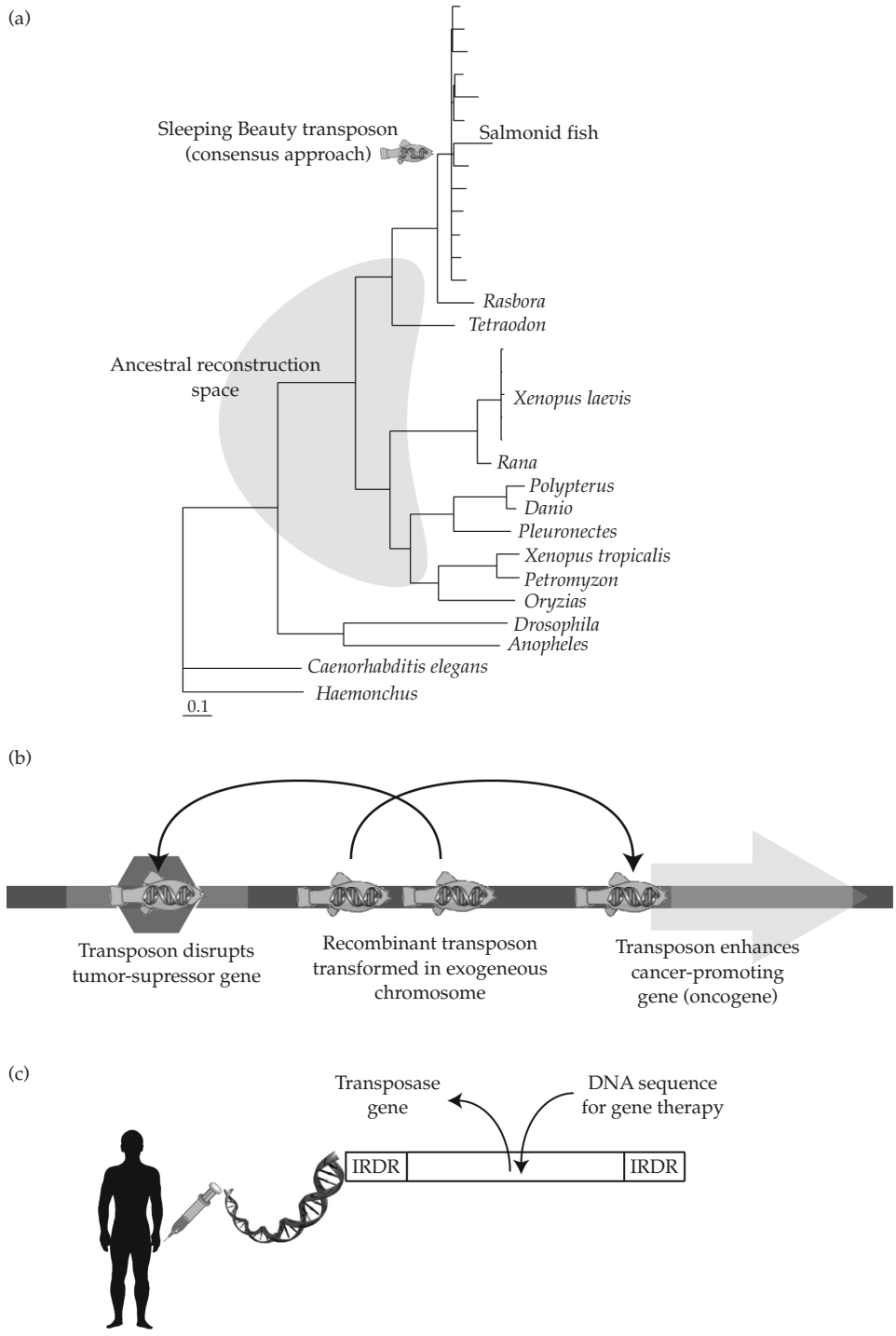

Figure 2.3 Ancestral transposons as tools. (a) Phylogeny of Tc1/mariner family of transposable elements. Sleeping Beauty (Ivics *et al.*, 1997) transposon generated by the consensus approach is shown along with ancestral sequence space of interest to our laboratory. (b) Schematic of an engineered Sleeping Beauty transposon that serves as a tool to identify genes involved in tumor suppression/growth. Adapted from Weiser and Justice (2005). (c) Transposons can be engineered as a vector for gene therapy. Inverted-repeat/direct-repeat elements (IRDRs) are regulatory/binding regions for transposition.

variant of the SB element that displays enhanced transposition activity.

One application of the SB variant has identified genes involved in mouse tumor suppression/promotion (Collier *et al*., 2005). Here, the transposase gene was replaced with a murine stem-cell-virus (MSCV) promoter, a string of poly(A)s, and slice-acceptor/-donor sites flanking the promoter and terminator signals. This transposon construct was then integrated in the mouse genome. Transposition was carried out by a transposase gene supplied *in trans*. The arrangement of splice sites in the transposon, and the location of the transposon either upstream or within an endogenous mouse gene, dictated whether the mouse gene was truncated or displayed increased expression. In this way, the transposon was used to disrupt tumor suppressors or enhance expression of cancer-promoting genes (Figure 2.3b).

The utility of the resurrected SB transposon is apparent. We are currently probing whether other ancestral sequences of the Tc1/mariner family have expanded utility (Figure 2.3a). Our motivation stems from the notion that a consensus approach may only capture evolutionary information in a heuristic manner. Reconstructions that utilize phylogenetic methods with explicit models of sequence evolution are more likely to recapitulate the ancestral form.

Overall, resurrected transposable elements have demonstrated their value by identifying genes involved in tumor growth. The next step is to develop a transposon capable of serving as a vector for gene therapy (Figure 2.3c). This will require that researchers overcome hurdles associated with the current limitations of these transposons (size of gene insert, site-integration specificity, etc.).

2.5.3 Cellular machines and recombinant (ancestral) genomes

In the interest of space limitations, let us briefly introduce the two remaining topics.

Cellular machines have a broad range of potentials; from simple expression of heterologous genes for laboratory analysis or environmental detection of explosive compounds such as TNT, to the synthesis of minimal artificial cells for small-compound drug or protein synthesis (Martin *et al*., 2003; Noireaux and Libchaber, 2004). We anticipate that ancestral reconstructed sequences will provide much of the foundation of genetic information for these machines in the future. As a first step, we have demonstrated that ancestral EF proteins can participate in a reconstituted translation system *in vitro* designed to incorporate unnatural amino acids (Figure 2.2c; Josephson *et al*., 2005). Further, experimental evolution studies of these ancestral genes introduced into laboratory organisms will enhance our biological understanding of adaptive and sequence landscapes and allow us to ask whether replaying the molecular tape of life is repetitive (Gould, 1989; Elena and Lenski, 2003; Lunzer *et al*., 2005; Weinreich *et al*., 2006). This work will have obvious extensions to natural history and the origins of (early) life.

Ancestral reconstructions will undoubtedly include larger and larger segments of a genome in the near future. As a first step in developing the appropriate technology to support this vision, the Venter Institute is in the process of constructing a minimal synthetic *Mycoplasma* genome (H. Smith, personal communication). Once the technology is available, why not construct a complete ancestral biochemical pathway (e.g. operon) or an ancestral genome (Figure 2.2d)? The ancestral reconstruction field would no longer be confined to single-gene reconstructions (Blanchette *et al*., 2004). Developing narratives based on resurrected pathways and genomes will have profound effects on our understanding of evolution (natural selection/biodiversity) and biomedicine. It is also quite possible that our understanding of what constitutes a sustaining minimal genome required to support life will be altered through ancestral reconstructions. In this way, homologous genes performing two different, but related, functions may share a single common ancestor that performed both of these functions, albeit with less efficiency or specificity.

We anticipate that the ancestral reconstruction field will drive synthetic biology once the technology permits us. We will then find ourselves in the center of the debate on artificial life, which will raise even more debate both within and outside of our field. Until then . . .

2.6 Conclusions

The reconstruction field has made tremendous strides since 1963 (Thornton, 2004). We anticipate that the past 40 years will go down in history as our lag-phase period. It is necessary then, as we enter the exponential phase and as technology harnesses our power, that we do not forget the importance of biological integration and natural history that Zuckerkandl and Pauling promoted. Some scientists have noted that the field of molecular biology had this connection but lost it in recent decades. For example, Woese has noted that *empirical* reductionism is suitable in our quest to understand the broader context of biological systems (Woese, 2004). *Fundamentalist* reductionist approaches toward molecular biology, however, trivialize the natural world and are therefore paralyzing. It is important that we do not reach such a stationary phase in our quest to understand biological systems and diversity through ancestral reconstructions. In a field full of narratives, consider this as merely one more.

2.7 Acknowledgments

This work was funded in part by a National Research Council/NASA Astrobiology fellowship and grants from the NASA Exobiology program and NIH.

I would like to thank David Liberles, David Ardell, and Giorgio Matassi for organizing the inaugural ancestral reconstruction conference held in Sweden in 2005, and for overseeing the publication of this book. I would further like to thank all of the contributors to this book for their enthusiasm to make our field as rigorous and fascinating as possible. Thanks to Mike Thomson, Slim Sassi, Michelle Burgan, Jack Szostak, and Kris Josephson for their assistance with our research, and to my undergraduate advisor (George P. Smith) for introducing me to Bayesian statistics in the early 1990s.

A special thanks to Steven Benner for adopting me into his laboratory as a graduate student. The parallels between Pauling and Benner in their adroit abilities to integrate physical sciences and natural history should not go unnoticed. Any discussion with Steve about his work on organic reactivity, biochemical pathways, stereoselectivity, non-standard nucleotides, or protein-structure prediction always lead to biological interpretations within a Darwinian framework.

References

Benner, S.A. (1995) Reconstructing ancient forms of life. *J. Cell. Biochem.* 200–200, suppl. 19A.

Benner, S.A. and Sismour, A.M. (2005) Synthetic biology. *Nat. Rev. Genet.* **6**: 533–543.

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.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* **14**: 2412–2423.

Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R., and Hillis, D.M. (1993) Experimental molecular evolution of bacteriophage-T7. *Evolution* **47**: 993–1007.

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.

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.

Colosimo, P.F., Hosemann, K.E., Balabhadra, S., Villarreal, G., Dickson, M., Grimwood, J. *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**: 1928–1933.

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

Elena, S.F. and Lenski, R.E. (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4**: 457–469.

Endy, D. (2005) Foundations for engineering biology. *Nature* **438**: 449–453.

Fitch, W. (1971) Towards defining the course of evolution. Minimum change for a specific tree topology. *Syst. Zool.* **20**: 406–416.

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.C., Simmen, F.A., Schreiber, D.R. *et al.* (2004) The planetary biology of cytochrome P450 aromatases. *BMC Biol.* **2**: 19.

Gaucher, E.A., De Kee, D.W., and Benner, S.A. (2006) Application of DETECTER, an evolutionary genomic tool to analyze genetic variation, to the cystic fibrosis gene family. *BMC Genomics* **7**: 44.

Gould, S.J. (1989) *Wonderful Life: The Burgess Shale and the Nature of History*. W.W. Norton & Company: New York.

Hall, B.G. (2006) Simple and accurate estimation of ancestral protein sequences. *Proc. Natl. Acad. Sci. USA* **103**: 5431–5436.

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

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.

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.

Josephson, K., Hartman, M.C.T., and Szostak, J.W. (2005) Ribosomal synthesis of unnatural peptides. *J. Am. Chem. Soc.* **127**: 11727–11735.

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

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

Lunzer, M., Milter, S.P., Felsheim, R., and Dean, A.M. (2005) The biochemical architecture of an ancient adaptive landscape. *Science* **310**: 499–501.

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.

Martin, V.J.J., Pitera, D.J., Withers, S.T., Newman, J.D., and Keasling, J.D. (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.* **21**: 796–802.

Noireaux, V. and Libchaber, A. (2004) A vesicle bioreactor as a step toward an artificial cell assembly. *Proc. Natl. Acad. Sci. USA* **101**: 17669–17674.

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

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

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.

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.

Ronquist, F. (2004) Bayesian inference of character evolution. *Trends Ecol. Evol.* **19**: 475–481.

Sassi, S.O., Benner, S.A., and Gaucher, E.A. (2007) Molecular paleosciences. Systems biology from the past. In *Advances in Enzymology and Related Areas of Molecular Biology: Protein Evolution*, vol. 75, pp. 1–132 (Toone, E., ed.). Wiley, Chichester.

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

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.

Shih, P., Malcolm, B.A., Rosenberg, S., Kirch, J.F., and Wilson, A.C. (1993) Reconstruction and testing of ancestral proteins. molecular evolution: producing the biochemical data. *Molecular Evolution: Producing the Biochemical Data. Methods in Enzymology*, vol. 224, pp. 3–725 (Zimmer, E.A., White, T.J., Cann, R.L., and Wilson, A.C., eds). Academic Press, San Diego.

Skovgaard, M., Kodra, J.T., Gram, D.X., Knudsen, S.M., Madsen, D., and Liberles, D.A. (2006) Using evolutionary information and ancestral sequences to understand the sequence-function relationship in GLP-1 agonists. *J. Mol. Biol.* **363**: 977–988.

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.

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

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

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006) Darwinian evolution can follow only

very few mutational paths to fitter proteins. *Science* **312**: 111–114.

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

Williams, P.D., Pollock, D.D., Blackburne, B.P., and Goldstein, R.A. (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol.* **2**, e69.

Woese, C.R. (2004) A new biology for a new century. *Microbiol. Mol. Biol. Rev.* **68**: 173–186.

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

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.