Thompson, J. D., Gibson, T. J., and Higgins, D. G. (1994). The ClustalW: 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.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). The ClustalX-Windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882.

Walker, I. D., and Bridgen, J. (1976). The keratin chains of avian scale tissue, sequence heterogeneity and the number of scale keratin genes. *Eur. J. Biochem.* **67,** 283–293.

Whitbread, L. A., Gregg, K., and Rogers, G. E. (1991). The structure and expression of a gene encoding chick claw keratin. *Gene* **101,** 223–229.

Wilton, S. D., Crocker, L. A., and Rogers, G. E. (1985). Isolation and characterization of keratin m RNA from the scale epidermis of the embryonic chick. *Biochim. Biophys. Acta* **824,** 201–201.

Zola, H. (1987). ''Monoclonal Antibodies: A Manual of Techniques.'' CRC Press, Boca Raton, FL.

[34] Applications of Ancestral Protein Reconstruction in Understanding Protein Function: GFP-Like Proteins

By BELINDA S. W. CHANG, JUAN A. UGALDE, and MIKHAIL V. MATZ

Abstract

Recreating ancestral proteins in the laboratory increasingly is being used to study the evolutionary history of protein function. More efficient gene synthesis techniques and the decreasing costs of commercial oligo-synthesis are making this approach both simpler and less expensive to perform. Developments in ancestral reconstruction methods, particularly more realistic likelihood models of molecular evolution, allow for the accurate reconstruction of more ancient proteins than previously possible. This chapter reviews phylogenetic methods of ancestral inference, strategies for investigating alternative reconstructions, gene synthesis, and design, and an application of these methods to the reconstruction of an ancestor in the green fluorescent protein family.

Introduction

Ancestral protein reconstruction allows for the recreation of protein evolution in the laboratory so that it can be studied directly. This approach is a natural extension of experimental studies that examine present-day

 0076-6879/05 $35.00

protein function, from which the evolutionary history of function may then be extrapolated. Reconstructing ancestral proteins can provide additional information not easily obtainable from studies of extant proteins, which are necessarily limited to the range of function available today. This kind of experimental approach has been used as a window into the evolutionary past of proteins in an increasing number of systems (Chandrasekharan *et al*., 1996; Sun *et al*., 2002; Thornton *et al*., 2003; Zhang and Rosenberg, 2002; for reviews, see Chang and Donoghue, 2000; Stewart, 1995; Thornton, 2004). In addition to providing valuable information about the evolution of present-day molecular structure and function, it also can lead to the discovery of new aspects of biochemical function that have subsequently been lost in extant proteins or that exist only in obscure or difficult to obtain organisms (Adey *et al*., 1994; Jermann *et al*., 1995; Malcolm *et al*., 1990). It can lead to new insights into the biology, or even the environment, of extinct organisms (Chang *et al*., 2002b; Gaucher *et al*., 2003).

Traditional methods of studying the structure and function of proteins generally have employed mutagenesis methods to identify residues or regions of the protein that are important for function. Although the targets of mutagenesis may often be directed by knowledge of the three-dimensional structure of the protein of interest, researchers are nonetheless faced with choosing from a vast number of possible mutations, either singly or in combination. Finding a number of mutations that produce properly folded and functional proteins, never mind those that show interesting and substantial effects in biochemical assays, can be difficult indeed. Multiple site mutants are, therefore, likely to be limited to combinations of single mutants that produce interesting phenotypes. Although this can be a reasonable approach given the expense and effort involved to make and express hundreds of mutants, it necessarily precludes combinations of mutations that by themselves do not produce any effect worth noting but together may create a novel or, better yet, interesting phenotype.

Reconstructing the evolutionary history of proteins in the laboratory offers several intriguing advantages compared to these more traditional approaches. Because the process of natural selection tends to eliminate the vast majority of mutations producing dysfunctional proteins, this approach effectively screens out mutations resulting in misfolded proteins and focuses on those changes that may have altered protein function during its evolutionary history. Moreover, the problem of assessing which combinations of mutations may be interesting from a functional point of view is addressed nicely using this approach; ancestral proteins are, in effect, combinations of mutations that, if properly chosen, have been selected to produce marked shifts in evolutionary function.

Here, we review some of the methods of ancestral inference, as well as the design and synthesis of ancestral genes in the laboratory. We also touch on some of the issues that can arise, particularly if different methods do not agree in their inference of ancestral states, and how these issues can be overcome. Finally, we highlight an example of how these methods can be applied to the green fluorescent protein (GFP)–like family of proteins.

Ancestral Gene Inference Methods

Generally speaking, two types of phylogenetic methods are used to infer ancestral sequences: parsimony and likelihood/bayesian methods (Table I). Parsimony methods (Swofford, 2002) minimize the amount of evolutionary change along each branch, assuming slow and consistent rates of evolutionary change. Likelihood methods, on the other hand, incorporate an explicit model of substitution, which allows for statistical comparisons among models in order to determine which is a better fit to the data at hand, at least among nested models (likelihood ratio tests; see later discussion). Parsimony methods, with reference to the reconstruction of ancestral states, have been extensively discussed and reviewed elsewhere (Cunningham *et al.*, 1998; Maddison, 1995; Omland, 1999; Swofford *et al.*, 1996) and are not discussed further here.

TABLE I
STATISTICAL METHODS OF INFERENCE OF ANCESTRAL STATES

Method	Programs available	Computer programs	References
Maximum Parsimony	http://macclade.org/macclade.html	MacClade (Maddison and Maddison, 1993)	(Maddison, 1995; Swofford *et al.*, 1996)
	http://paup.csit.fsu.edu/	PAUP* (Swofford, 2002)	
Maximum likelihood/ bayesian	http://abacus.gene.ucl.ac.uk/software/paml.html	PAML (Yang, 1997)	Empirical Bayes (Yang *et al.*, 1995)
	http://morphbank.ebc.uu.se/mrbayes/	MrBayes (Huelsenbeck and Ronquist, 2001)	Hierarchical Bayes (Huelsenbeck and Bollback, 2001a)

Phylogenetic methods based on maximum likelihood analysis (implemented in programs such as PHYLIP [Felsenstein, 1991], MOLPHY [Adachi and Hasegawa, 1994], PAML [Yang, 1997], and NHML [Galtier and Gouy, 1998]) use a likelihood score as an optimality criterion. This likelihood score is calculated according to a specified model of evolution (Felsenstein, 1981). Optimizing the likelihood score can be used to infer the most likely tree topology, as well as parameters such as branch lengths, character state frequencies, and ancestral states. Bayesian methods are then used to calculate ancestral states with the highest posterior probability. This can be done by using the maximum likelihood topology, branch lengths, and model parameters as priors (empirical Bayes method [Yang *et al*., 1995]), or alternatively the posterior probabilities can be calculated by taking into account the uncertainty in the maximum likelihood topology and parameters using a Markov chain Monte Carlo approach (hierarchical Bayes method, (Huelsenbeck and Bollback, 2001a) (see Table I). This approach has some advantages over parsimony methods (Koshi and Goldstein, 1996; Lewis, 1998; Yang *et al*., 1995). In using an explicit model of molecular evolution, stochastic methods allow for the incorporation of knowledge of the mechanisms and constraints acting on coding sequences, as well as the possibility of comparing the performance of different models, ultimately resulting in the development of more realistic models (Goldman, 1993).

With stochastic methods such as maximum likelihood and Bayesian analysis, it is important to explore different models of molecular evolution to determine how robust the ancestral reconstruction results are (Huelsenbeck and Bollback, 2001b; Huelsenbeck *et al*., 2002). Oversimplified or unrealistic models have been shown in certain cases to yield spurious phylogenetic reconstructions (Buckley, 2002; Cao *et al*., 1994; Huelsenbeck, 1997), emphasizing the importance of model selection. Models can be grouped into different classes: nucleotide, amino acid, and codon. Nucleotide models range from the simplest (Jukes-Cantor, 1969), which assumes equal base frequencies and rates of transitions and transversions, to much more complex models allowing unequal base frequencies (Felsenstein, 1981), transition/transversion bias (Kimura, 1980), among-site rate heterogeneity (Yang, 1994), and/or nonstationary base composition (Galtier and Gouy, 1998).

The simplest amino acid model is the Poisson, which assumes equal amino acid frequencies and rates of substitution among amino acids. This model is clearly unrealistic and would not be expected to perform well. More realistic models have been developed that allow unequal amino acid frequencies (Hasegawa and Fujiwara, 1993), and among-site rate heterogeneity (Yang, 1994), in addition to a general time-reversible (GTR) model

for amino acids, which allows for unequal numbers of substitutions in the rate matrix for all the different classes of amino acid substitutions (Yang, 1997). Fixed rate matrices have also been calculated for a number of datasets, including globular proteins (Cao *et al*., 1994; Dayhoff, 1978; Jones *et al*., 1992; Kishino *et al*., 1990), and mitochondrial transmembrane proteins (Adachi and Hasegawa, 1996). The use of fixed or constant parameters in the rate matrix can be advantageous because it allows for a reduction in the number of parameters in the model of evolution being used. Recent developments include amino acid models that allow replacement rates to be proportional to the frequencies of both the replaced and the resulting residues (+*gwF* model; (Goldman and Whelan, 2002).

Codon-based models of molecular evolution show the most promise, as they have the potential to incorporate information about different types of nucleotide substitutions and whether they change the amino acid or not. The original codon-based models assumed equal nonsynonymous-to-synonymous rate ratios among sites and lineages (Goldman and Yang, 1994; Muse and Gaut, 1994). Subsequently, models were developed that allowed that ratio to vary across lineages in a phylogeny (Yang, 1998), across sites in a protein (Nielsen and Yang, 1998), and across both sites and lineages (Yang and Nielsen, 2002). Several models, each employing a different statistical distribution of nonsynonymous-to-synonymous rate ratios (dN/dS) across sites, have been developed as tools to detect positively selected sites using likelihood ratio tests (''random sites models'' [Yang *et al*., 2000]; for reviews see Bielawski and Yang [2003] and Yang and Bielawski [2000]). If, for example, there is some *a priori* information available based on tertiary protein structure that can be used to partition sites in the protein into different classes, then fixed-sites models may also be used (Yang and Swanson, 2002).

Given the diversity of models now available, the choice of a particular model for use in phylogenetic analysis and ancestral inference is critical. An inappropriate model of evolution can lead to inconsistency in the likelihood analysis and convergence to an incorrect result (Huelsenbeck, 1998). Ancestral inference methods are particularly sensitive to model choice. The possibility of an incorrect result can be reduced by selecting a model of evolution that displays the best fit to the sequence data at hand. To this end, likelihood ratio tests can be used to compare two models of evolution that are nested with respect to each other, to determine whether the more complex model fits the sequence data significantly better than the simpler model (Felsenstein, 1981; Huelsenbeck and Rannala, 1997; Yang *et al*., 1994). For nested models, a more complex model (H_1) will contain all the parameters of the original model (H_0), as well as additional parameters. If the models are not nested, they cannot be directly compared using a

likelihood ratio test, and other methods, such as the generation of the distribution of the test statistic using Monte Carlo simulation, must be used (Goldman, 1993).

Choosing an Ancestral Sequence to Reconstruct

What happens when different likelihood models or even methods of inferring ancestral sequences result in different amino acid reconstructions at particular sites in the protein? This may not happen if the sequences are closely related, where we would expect inferences about ancestral states to be fairly robust to changes in the particular model of evolution used. Additionally, it may be possible to compare likelihood models using likelihood ratio tests, if the models are nested, and then only the reconstructions from the model with the better fit would be considered. However, it is not always possible to compare models this way, and there usually will be a certain proportion of sites for which feasible alternative reconstructions exist, among which it may be difficult to choose.

There are different approaches that can be taken to address this problem. The simplest is to randomly choose one reconstruction (per site) among all the alternatives. This has obvious drawbacks if the true reconstruction is not among the alternatives chosen. In some cases, the probability that the ancestral sequence chosen will match the true ancestor across all sites may be quite small. This may have important consequences for subsequent interpretation of the results of functional assays of the ancestral protein. It seems more appropriate to incorporate at least some exploration of alternative methods and/or models in reconstructing the ancestral protein in the laboratory. This can be done in a number of ways. For the purpose of synthesizing ancestral proteins in the laboratory, different types of models may be most suitable, depending on how deep the ancestral node is in the tree. More recent ancestors that are not too diverged from existing sequences may be best reconstructed using nucleotide models, where parameters like transition/transversion rate ratios predominate. In contrast, as divergence increases, the more ancient nodes may be best reconstructed using the amino acid models, where factors such as side-chain properties predominate. For this reason, a useful approach is to synthesize several variants of the gene predicted to be the best reconstruction under different models, so the results of different models can be compared in functional assays in the laboratory (Chang and Donoghue, 2005).

Another approach entirely would be to incorporate degeneracies at sites where alternative reconstructions exist *during* the gene synthesis (degenerate reconstruction). In this way, it is possible to pool together the predictions of different models, which mitigates the problem of choosing

the most appropriate model to a certain extent. As a result, a library of possible ancestral genes is obtained instead of just a single gene that is the most probable according to a particular model. Although the degenerate reconstruction approach is not really aimed at directly comparing the most likely ancestral reconstructions inferred using different likelihood models, it may be useful to compare models' performances by evaluating the probabilities of different phenotypes according to each of the models. If the different phenotypes are observed in the combinatorial library, the corresponding genes can be sequenced to determine the combinations of degenerate sites that are responsible for the phenotypes. It is then possible to go back and evaluate how probable these particular combinations are relative to each other according to the model's predictions. In this way, one can obtain the result in the form such as ''at the node N, model A predicts phenotype X with probability P1, phenotype Y with probability P2, and so on.'' The necessary information concerning probabilities of each state at each site at each node can be extracted with the PAML 3.13 package by specifying ''verbose = 2'' along with ''Rate Ancestor = 1'' in a control file for *codeml* or *baseml*. The biggest consideration in estimating phenotypic probabilities is that it requires an efficient method for screening the combinatorial library for different phenotypes, which may not be feasible for many proteins. Therefore, the degenerate and targeted reconstruction approaches are equally valid and, in fact, somewhat complementary to each other in that they explore alternative reconstructions in different ways.

Ancestral Gene Design and Synthesis

Artificial Gene Design

Once an ancestral protein sequence or an array of possible sequences has been inferred, the degeneracy of the genetic code can be used to design artificial genes with properties useful in the synthesis and expression of the ancestral gene. Unique restriction sites and potential primer sites that later will aid in the characterization and construction of the gene can be incorporated. Codon usage bias can be optimized for a particular species or cell type (Sharp *et al.*, 1988). In many expression systems, rare codons are known to cause translational problems caused by limited tRNA availability, resulting in misincorporations, truncated proteins, and overall reduced translational efficiency (Kane, 1995). Conversely, although the goal of optimizing codon usage frequencies is usually increased expression levels, the incorporation of unpreferred codons is occasionally useful in slowing translation of signal sequences so that cellular membrane translocation

systems are not saturated (Karnik *et al*., 1987). The secondary structure of mRNA also has been implicated in lowered expression levels in *Escherichia coli* (Griswold *et al*., 2003). GC content can affect levels of heterologously expressed proteins (Sinclair and Choy, 2002) and may need to be adjusted to minimize potential difficulties in later molecular biology manipulations such as cloning and sequencing. Epitopes for antibodies or other tags that would aid in protein purification also can be introduced in the design of an artificial gene.

Gene Synthesis Incorporating Degenerate Sites

The principle of the method described here is depicted in Fig. 1. It uses an array of overlapping oligonucleotides 30–35 bases long to assemble both strands of the synthesized gene by means of ligation, followed by PCR amplification of the target product using flanking oligonucleotides as primers. Note that degenerate sites, if they are to be incorporated into the gene synthesis, should be positioned as far as possible from the ligation points. Though very simple, the method has important advantages over previously reported techniques that rely on longer oligonucleotides (Chang *et al*., 2002a; Ferretti *et al*., 1986), because it minimizes the number of errors introduced during the chemical synthesis of the oligomers and allows for them to be ordered commercially instead of requiring an in-house oligonucleotide synthesizer. Long oligonucleotides tend to form secondary structures during synthesis, resulting in frequent errors that can include deletions and insertions of varying sizes. In contrast, shorter oligonucleotides are usually synthesized for a minimal cost with accuracy that is high enough to skip the expensive purification procedures, which are often necessary for long oligonucleotides. In addition, in our method, the oligonucleotides do not need to be modified (e.g., they do not require 5′ phosphates), which further decreases the cost of the project.

Finally, a gene synthesis strategy that incorporates so many ligation points per gene is particularly useful because the ligation efficiency is significantly diminished by the presence of mismatches in the vicinity of the ligation site. In our protocol, the separation between the ligation sites is only 16–17 bases, which means that almost three-fourths of the gene length is actually ‘‘proofread’’ at the ligation step because DNA ligase is sensitive to mismatches at least up to 6 bases from the ligation site (Roth *et al*., 2004). For example, the mutated clones incorporating accidental errors at this step in our experiments made up less than 50% of the total number, and even in those clones, the mutations were likely to be PCR errors rather than gene assembly artifacts.

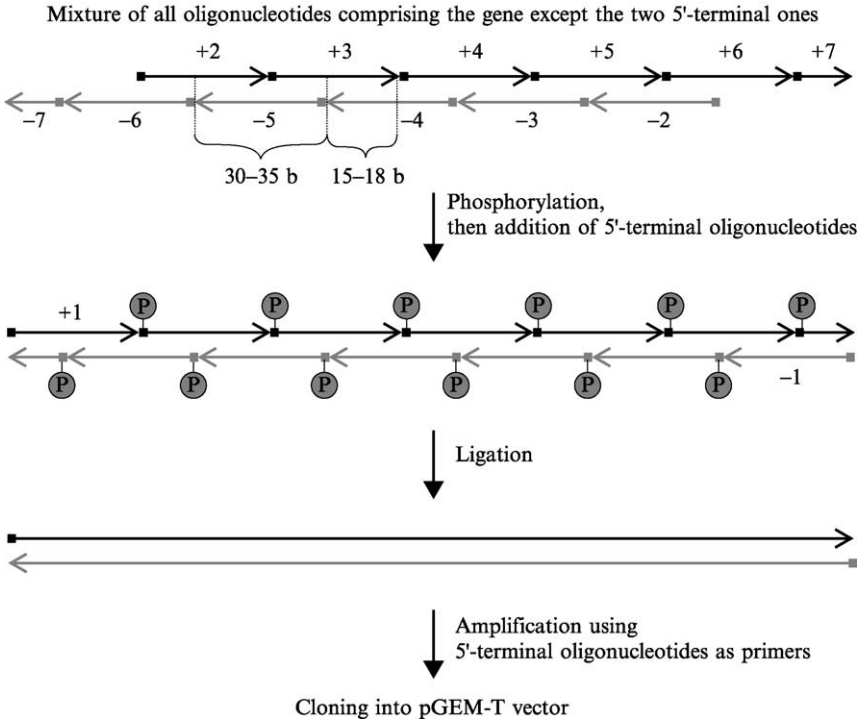

FIG. 1. Schematic outline of the described gene synthesis strategy. Oligonucleotides corresponding to plus and minus DNA strands are shown as black and gray arrows, respectively. Arrowheads correspond to free 3′ termini, squares–to free 5′ termini. For simplicity of representation, the scheme shows the synthesis of a short fragment about 210–250 bp in length; however, the strategy will work for the longer genes as well. In our experiment, as discussed in the text, the synthesized genes were 730 bp long.

Gene Synthesis Protocol

Oligonucleotides: The artificially designed ancestral gene should be divided into overlapping oligonucleotide fragments of about 30–35 bases in length. These oligonucleotides can be ordered from any reliable commercial service. No additional purification or modification is required; the smallest offered synthesis scale (usually 25 nmol) is sufficient.

Phosphorylation: In a 0.5-ml tube, combine 5 μl of 2$\times$ buffer for T4 ligase, 4 μl of the oligonucleotide mixture (all oligonucleotides that comprise the gene in a concentration of 0.1 μM each, except the two 5′-terminal ones that will not be ligated by their 5′ ends,

see Fig. 1) and 1 μl of T4 polynucleotide kinase (New England BioLabs, Beverly, MA). We used the buffer provided within the pGEM-T PCR cloning kit (Promega) because it is similar in composition to the standard T4 polynucleotide kinase buffer but already contains ATP in appropriate concentration. Incubate the reaction at 37° for 30 min, and then incubate it at 65° for 20 min to deactivate the enzyme.

Ligation: To the completed phosphorylation reaction, add 5 μl of 2× Ligation Buffer (Promega, Madison, WI), 4 μl of the terminal oligonucleotides mixture (Fig. 1, 0.1 μM each) and 1 μl of the T4 DNA ligase (New England BioLabs, Beverly, MA). Incubate the reaction for 2 h at 37°.

PCR amplification of the ligated products: It is important to use a polymerase or polymerase mixture exhibiting proofreading activity, to minimize PCR errors. In our experiments, we used Advantage 2 polymerase mixture (BD Biosciences Clontech, San Jose, CA) with the buffer provided. To perform the amplification, combine in an 0.5 thin-walled PCR tube: 2 μl of the ligation reaction, 2 μl of each of the 5′-terminal oligonucleotides diluted to 1 μM, 2 μl of the 10× reaction buffer, 2 μl of 5 mM dNTP mixture, 12 μl deionized water, and 0.5 μl of the Advantage 2 polymerase mix. Perform cycling according to the following program: 45 s at 94°, 1 min at the annealing temperature (depends on the sequence of the primers), 1 min at 72° (add 1 min per each 1000 bp of the synthesized gene over 1500 bp); 15–20 cycles. The accumulation of the PCR product should be monitored to keep the number of PCR cycles to the necessary minimum. The product should become visible on a standard agarose gel after 15–20 cycles, when 1/10 of the reaction volume is loaded into the well. The PCR product is then cloned into pGEM-T (Promega) in order to obtain bacterial expression libraries.

Example of Ancestral Protein Expression: GFP-Like Proteins

The primary function of the family of GFPs (and related colored proteins), first isolated from the jellyfish *Aequorea victoria,* is coloration and/or fluorescence. This is acquired by these proteins via autocatalytic synthesis of the chromophore moiety within its own globule, using its own side chains as substrates (Heim *et al*., 1994; Matz *et al*., 2002). GFP-like proteins are the only natural pigments in which both chromophore and protein are contained within a single gene, which has earned them great popularity as biotechnology tools (see Lippincott-Schwartz and Patterson [2003] for a recent review). They come in four basic colors, roughly

corresponding to distinct types of the chromophore chemical structure: fluorescent colors including green, yellow, and red, and nonfluorescent purple-blue (Labas *et al*., 2002). In many ways GFP-like proteins represent a convenient model for basic studies in the evolution of gene families (Matz *et al*., 2002). The family contains many cases of gene duplication followed by diversification of function and may even present several cases of convergent evolution of complex features at the molecular level (Labas *et al*., 2002; Shagin *et al*., 2004). At the same time, the proteins are small (~230 amino acid residues long) and can be expressed easily in a functional form in a variety of heterologous systems including bacteria. The phenotype, which is simply the color of fluorescence, can already be precisely quantified in the bacterial colonies growing on a solid media. No further purification of the expressed proteins is necessary. This provides an excellent opportunity for high-throughput screening of expression libraries of mutants or, as we show in this chapter, reconstructed ancestral genes.

The experiment presented here is part of an ongoing study of the color evolution among paralogous lineages of GFP-like proteins found in the great star coral *Montastrea cavernosa* (Ugalde *et al*., 2004). This species possesses several (at least four and maybe up to seven) genetic loci coding for GFP-like proteins comprising four paralogous groups corresponding to cyan (emission max at 480–495 nm), shortwave green (emission max at 500–510 nm), long-wave green (emission max at 515–525 nm), and red (emission max at 575–585 nm) colors (Kelmanson and Matz, 2003). These colors share a common ancestor sometime after the first diversification of corals in early Triassic (240 million years ago) and before mid-Jurassic (180 million years ago), according to the fossil record and the phylogenetic tree topology (Kelmanson and Matz, 2003). The most ancient ancestral phenotype of GFP-like proteins is likely to be shortwave green (Shagin *et al*., 2004), while the phenotype of the common ancestor of *M. cavernosa* paralogs could be anything because more basal branches in the phylogeny lead to proteins of all colors (Kelmanson and Matz, 2003; Shagin *et al*., 2004). We set out to reconstruct proteins in the nodes representing the common ancestor of all phenotypes (ALL ancestor), the common ancestor of all the red proteins (Red ancestor, or R), and two intermediate nodes, corresponding to the two possible common ancestors of reds and longwave greens (Red/Green ancestor, or RG; and pre-Red ancestor, or pre-R; see Fig. 4B).

For the prediction of the ancestral sequences, the dataset described in Shagin *et al*. (2004), comprising most of the cnidarian GFP-like proteins known, was used. Three alternative maximum likelihood models were applied: amino acid–based JTT (Jones *et al*., 1992), codon-based M5 (Yang *et al*., 2000), and nucleotide-based GTR+G3 (Tavare, 1986). The latter model was different from the more common GTR+G in that it assumed, not one, but

three independent gamma-distributed rates of evolution at individual nucleotide sites, corresponding to their positions within codons. GTR+G3 performed best among nucleotide-based models compared in likelihood ratio tests using the Modeltest program (Posada and Crandall, 1998).

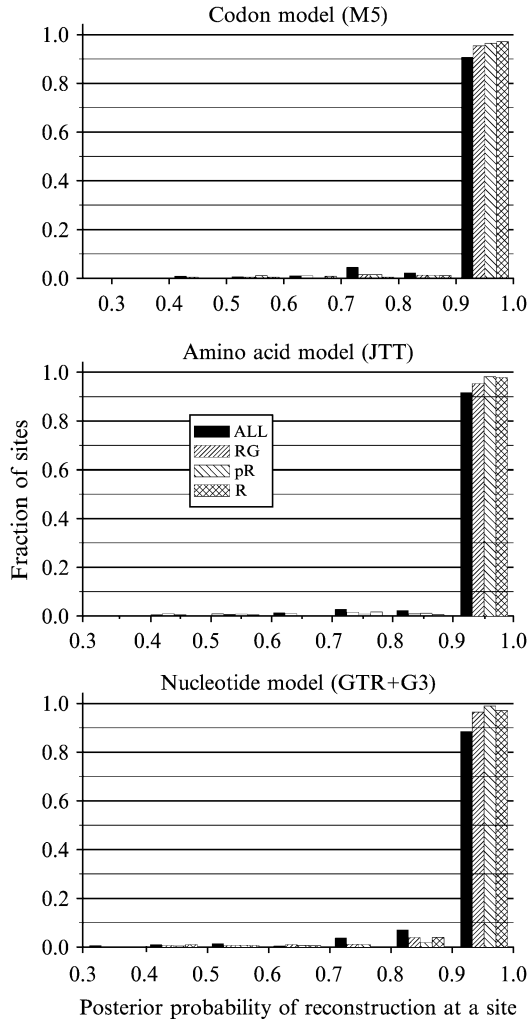

FIG. 2. Performance of different models in reconstruction of four ancestral sequences. Horizontal axis shows bin limits, so that, for example, the bars appearing between marks 0.8 and 0.9 show the fractions of sites that are predicted with posterior probability >0.8 and ≤ 0.9. In the legend, ALL, RG, pR, and R correspond to the ALL, red/green, pre-red, and red ancestors, respectively.

The reconstructions of all four ancestral sequences were quite robust with any model, although the GTR+G3 model was slightly less robust in its predictions compared to the other two (Fig. 2 and Table II). The least posterior probability at a reconstructed site was 0.328, observed in the reconstruction of the ALL-ancestor under the GTR+G3 model (the other two models predicted the same site with posterior probability 0.78). Comparison of the most probable reconstructions (Fig. 3) revealed that a small number of sites were predicted differently under different models. Apparently, these sites were poorly predictable by some or all of the three models, because no disagreement was observed between models when all three of them generated the site prediction with posterior probability exceeding 0.80. When planning ancestral gene synthesis, the codons corresponding to these ambiguous sites were designed to be degenerate to incorporate the alternative predictions. As a result, the designed genes for ALL, RG, pre-R, and R ancestors contained eight, six, four, and six degenerate codons, respectively.

A total of 500–1000 fluorescent clones from each of the four combinatorial libraries were visually surveyed using a fluorescent stereomicroscope (Leica MZ FLIII) with the optical filters providing excitation in the 400–450 nm range and emission from 475 nm and up (long-pass filter). Such a filter combination allows for easy discrimination of different fluorescent phenotypes by human eye, even such similar ones as long-wave and short-wave green. In two of the four cases (ALL ancestor and Red ancestor), no phenotypic diversity was observed, while the Red/Green and Pre-Red ancestors were represented by clones appearing in slightly different shades of yellow. Twenty-four clones from each library were sequenced and plated onto new plates for spectroscopy. The fraction of clones containing no additional mutations was 0.54–0.75. Among these ''clean'' clones, there were variations at all the degenerate sites. As predicted, the common ancestor of all colors (ALL ancestor) turned out to be shortwave green. Most interestingly, all clones corresponding to the two possible common ancestors of red and green proteins (Red/Green and pre-Red) showed an

TABLE II
AVERAGE POSTERIOR PROBABILITIES OF ANCESTRAL RECONSTRUCTION AT A SITE

Ancestor	JTT	Models GTRG3	M5
ALL	0.972	0.958	0.971
Red/Green	0.979	0.981	0.983
Pre-Red	0.987	0.987	0.988
Red	0.990	0.981	0.989

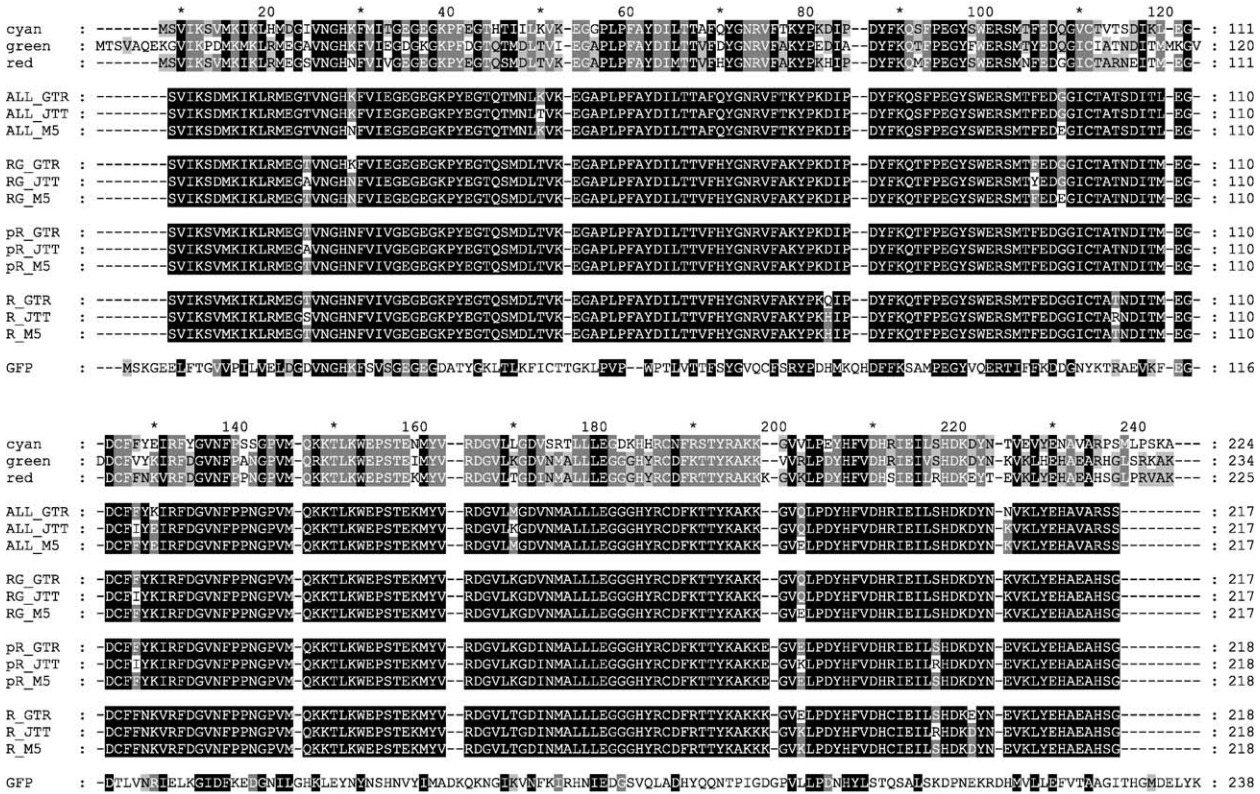

FIG. 3. Alignment of extant cyan, green, and red fluorescent proteins from *Montastrea cavernosa* (GenBank accession numbers AY181556, AY181554, and AY181552, respectively) (Kelmanson and Matz, 2003); and their ancestors predicted using three models (GTR+G3, JTT, and M5, see text for details). The ancestral sequences correspond to the nodes denoted on Fig. 2: ALL-ancestor (ALL), red/green ancestor (RG), pre-red ancestor (pR), and red ancestor (R). The green fluorescent protein (GFP) from *Aequorea victoria* (accession number M626539) is aligned below for reference.

intermediate long-wave green/red phenotype. Although the majority of the expressed protein bulk remained long-wave green, a small fraction was able to complete the third chromophore maturation stage resulting in a minor peak of red emission. Clones of the Red ancestor showed an ''imperfect red'' phenotype; although in them the red emission peak always dominated, the rate of green-to-red conversion during the last chromophore maturation stage was apparently still slower than in extant reds, resulting in a prominent minor peak of green fluorescence (Fig. 4A and C).

It is clear from these experiments that the evolution of red emission color, which corresponds to an increase of functional and structural complexity (Shagin *et al*., 2004), progressed through a series of intermediate stages. Also, it was possible to establish the color of the common ancestor of all the *M. cavernosa* paralogs as shortwave green. The complete molecular analysis of the color evolution in fluorescent proteins that would include studies of selection pressure across individual sites and mutagenesis experiments will be published elsewhere.

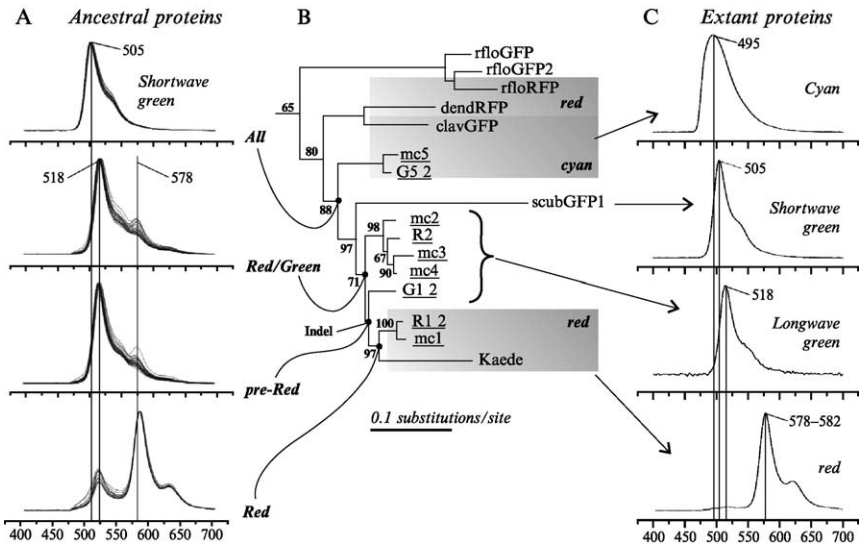

FIG. 4. Evolution of colors in a subset of cnidarian green fluorescent protein (GFP)–like proteins. (A) Fluorescence spectra of the reconstructed ancestral proteins. Multiple curves correspond to clones bearing variations at degenerate sites. (B) The portion of the phylogenetic tree of GFP-like proteins discussed here. The names of sequences originating from the great star coral *Montastrea cavernosa* are underlined. The values at the branches are nonparametric bootstrap support under the maximum likelihood criterion with GTR+G+I model. The sequence G1.2, which in an unconstrained bootstrap analysis has an uncertain affinity either to red or long-wave green cluster, has been forced to group with the red proteins on the basis of shared three-nucleotide (one codon) indel. (C) Fluorescence spectra of extant proteins.

Concluding Remarks

Laboratory synthesis of ancestral proteins is becoming a fast and fairly inexpensive method for studying the evolution of molecular structure and function. We no longer have to rely on inferences about the evolution of protein function based on amino acid substitutions thought to be historically important for functional shifts, which are then incorporated into site-directed mutagenesis studies of present day proteins. Ancestral protein synthesis offers a much more direct view of the evolutionary history of proteins, where the entire ancestral protein can be recreated and functionally assayed in the laboratory. This can shed light not only on the structure and function of present-day proteins, but also potentially on how evolution gave rise to the diversity of function seen today.

Acknowledgments

This work was supported by grants from the NSF and the National Sciences and Engineering Research Concil of Canada (B.S.W.C.), Grass Foundation (J.A.U.), and the U.S. Department of Defense and NIH (M.V.M.)

References

Adachi, J., and Hasegawa, M. (1994). ''MOLPHY.'' Institute of Statistical Mechanics, Tokyo, Japan.

Adachi, J., and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42,** 459–468.

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. USA* **91,** 1569–1573.

Bielawski, J. P., and Yang, Z. (2003). Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J. Struct. Funct. Genomics* **3,** 201–212.

Buckley, T. R. (2002). Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. *Syst. Biol.* **51,** 509–523.

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.

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., and Donoghue, M. J. (2005). Phylogenetic reconstruction of the origin of rod opsins from cone opsin ancestors. *submitted*.

Chang, B. S., Kazmi, M. A., and Sakmar, T. P. (2002a). Synthetic gene technology: Applications to ancestral gene reconstruction and structure-function studies of receptors. *Methods Enzymol.* **343,** 274–294.

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., Donoghue, M. J., and Sakmar, T. P. (2002b). Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19,** 1483–1489.

Cunningham, C. W., Omland, K. E., and Oakley, T. H. (1998). Reconstructing ancestral character states: A critical reappraisal. *Trends Ecol. Evol.* **13,** 361–366.

Dayhoff, M. O. (1978). A model of evolutionary change in proteins. Matrices for detecting distant relationships. *In* ''Atlas of Protein Sequence and Structure,'' (M. O. Dayhoff, ed.), Vol. 5(Suppl. 3), pp. 345–358. National Biomedical Research Foundation, Washington, DC.

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

Felsenstein, J. (1991). ''PHYLIP: Phylogeny Inference Package.'' University of Washington, Seattle, WA.

Ferretti, L., Karnik, S. S., Khorana, H. G., Nassal, M., and Oprian, D. D. (1986). Total synthesis of a gene for bovine rhodopsin. *Proc. Natl. Acad. Sci. USA* **83,** 599–603.

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.

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.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36,** 345–361.

Goldman, N., and Whelan, S. (2002). A novel use of equilibrium frequencies in models of sequence evolution. *Mol. Biol. Evol.* **19,** 1821–1831.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736.

Griswold, K. E., Mahmood, N. A., Iverson, B. L., and Georgiou, G. (2003). Effects of codon usage versus putative 5′-mRNA structure on the expression of Fusarium solani cutinase in the *Escherichia coli* cytoplasm. *Protein Exp. Purif.* **27,** 134–142.

Hasegawa, M., and Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* **2,** 1–5.

Heim, R., Prasher, D. C., and Tsien, R. Y. (1994). Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc. Natl. Acad. Sci.* **91,** 12501–12504.

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

Huelsenbeck, J. P. (1998). Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? *Syst. Biol.* **47,** 519–537.

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

Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* **51,** 673–688.

Huelsenbeck, J. P., and Rannala, B. (1997). Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276,** 227–232.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755.

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.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* **8,** 275–282.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. *In* ''Mammalian Protein Metabolism'' (H. N. Munro, ed.), pp. 21–132. Academic Press, New York.

Kane, J. F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotech.* **6,** 494–500.

Karnik, S. S., Nassal, M., Doi, T., Jay, E., Sgaramella, V., and Khorana, H. G. (1987). Structure-function studies on bacteriorhodopsin. II. Improved expression of the bacterio-opsin gene in *Escherichia coli*. *J. Biol. Chem.* **262,** 9255–9263.

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

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31,** 151–160.

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

Labas, Y. A., Gurskaya, N. G., Yanushevich, Y. G., Fradkov, A. F., Lukyanov, K. A., Lukyanov, S. A., and Matz, M. V. (2002). Diversity and evolution of the green fluorescent protein family. *Proc. Natl. Acad. Sci.* **99,** 4256–4261.

Lewis, P. O. (1998). Maximum likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. *In* ''Molecular Systematics of Plants II: DNA Sequencing'' (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.), pp. 132–163. Kluwer, Boston.

Lippincott-Schwartz, J., and Patterson, G. H. (2003). Development and use of fluorescent protein markers in living cells. *Science* **300,** 87–91.

Maddison, W. P. (1995). Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Syst. Biol.* **44,** 474–481.

Maddison, W. P., and Maddison, D. R. (1993). ''MacClade.'' Sinauer Associates, Sunderland, MA.

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.

Matz, M. V., Lukyanov, K. A., and Lukyanov, S. A. (2002). Family of the green fluorescent protein: Journey to the end of the rainbow. *Bioessays* **24,** 953–959.

Muse, S. V., and Gaut, B. S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11,** 715–724.

Nielsen, R., and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148,** 929–936.

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

Posada, D., and Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14,** 817–818.

Roth, M. E., Feng, L., McConnell, K. J., Schaffer, P. J., Guerra, C. E., Affourtit, J. P., Piper, K. R., Guccione, L., Hariharan, J., Ford, M. J., Powell, S. W., Krishnaswamy, H., Lane, J., Intrieri, G., Merkel, J. S., Perbost, C., Valerio, A., Zolla, B., Graham, C. D., Hnath, J., Michaelson, C., Wang, R., Ying, B., Halling, C., Parman, C. E., Raha, D., Orr, B., Jedrzkiewicz, B., Liao, J., Tevelev, A., Mattessich, M. J., Kranz, D. M., Lacey, M., Kaufman, J. C., Kim, J., Latimer, D. R., and Lizardi, P. M. (2004). Expression profiling using a hexamer-based universal microarray. *Nat. Biotechnol.* **22,** 418–426.

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.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H., and Wright, F. (1988). Codon usage patterns in *Escherichia-coli, Bacillus-subtilis, Saccharomyces-cerevisiae, Schizosaccharomyces-pombe, Drosophila-melanogaster* and Homo-Sapiens—a review of the considerable within-species diversity. *Nucleic Acids Res.* **16,** 8207–8211.

Sinclair, G., and Choy, F. Y. M. (2002). Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, *Pichia pastoris*. *Prot. Exp. Purif.* **26,** 96–105.

Stewart, C.-B. (1995). Active ancestral molecules. *Nature* **374,** 12–13.

Sun, H. M., 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.

Swofford, D. L. (2002). ''PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods).'' Sinauer, Sunderland, MA.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1996). Phylogenetic Inference. *In* ''Molecular Systematics'' (D. M. Hillis, C. Moritz, and B. K. Mable, eds.), pp. 407–514. Sinauer, Sunderland, MA.

Tavare, L. (1986). Some probabilistic and statistical problems of the analysis of DNA sequences. *Lect. Math. Life Sci.* **17,** 57–86.

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

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

Ugalde, J. A., Chang, B. S., and Matz, M. V. (2004). Evolution of coral pigments recreated. *Science* **305**(5689), 1433.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39,** 306–314.

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

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–573.

Yang, Z., and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503.

Yang, Z., Goldman, N., and Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11,** 316–324.

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

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19,** 908–917.

Yang, Z., and Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19,** 49–57.

Yang, Z. H., Nielsen, R., Goldman, N., and Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449.

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.* **99,** 5486–5491.