CHAPTER 15

Dealing with model uncertainty in reconstructing ancestral proteins in the laboratory: examples from archosaur visual pigments and coral fluorescent proteins

Belinda S.W. Chang, Mikhail V. Matz, Steven F. Field, Johannes Müller, and Ilke van Hazel

15.1 Introduction

Resurrecting ancestral proteins in the laboratory can be a powerful tool in studies of protein structure and function as they can offer a rare glimpse into the evolutionary history of molecular function (Malcolm *et al*., 1990; Adey *et al*., 1994; Chandrasekharan *et al*., 1996; Dean and Golding, 1997; Bishop *et al*., 2000; Chang and Donoghue, 2000; Sun *et al*., 2002; Zhang and Rosenberg, 2002; Thornton, 2004). Another, perhaps even more intriguing reason for reconstructing ancestral proteins lies in the hope of achieving a better understanding of the biology of ancient animals that may have possessed these proteins (Jermann *et al*., 1995; Messier and Stewart, 1997; Galtier *et al*., 1999; Benner, 2002; Gaucher *et al*., 2003). Proteins that are involved in sensory systems might be particularly revealing with respect to the physiology and behavior of ancient animals that can no longer be studied directly in the laboratory (Nei *et al*., 1997; Boissinot *et al*., 1998; Chang *et al*., 2002). Moreover, experimental tests of laboratory-recreated ancestral proteins would provide information different from that obtained through studies of fossils. Although interpretations based on recreations of single molecules are of course limited, under the best of circumstances one may hope to test some of the theories of ancient animal biology derived from other methods such as paleontological studies (Chang *et al*., 2002).

Reconstructions of the past depend entirely on the accuracies and limitations of the statistical methods and models employed. However, even in cases of deep divergences, when the accuracy of the reconstruction may be low, the experimental outcome remains valuable, as the effects of altering specific amino acids on a protein's structure and function can be interesting, independently of whether or not the sequence represents the true ancestor. The general approach of using site-directed mutagenesis methods to alter amino acids in order to assess shifts in function is one of the most widely employed in studying protein function.

Advances in phylogenetic methods of ancestral reconstruction, particularly the development of likelihood/Bayesian models that incorporate many different aspects of sequence evolution, have led to a plethora of models and methods available for use in phylogenetic reconstruction in recent years (Thorne, 2000; Huelsenbeck and Bollback, 2001; Whelan *et al*., 2001; Nielsen, 2005). This chapter briefly discusses some of the models available for use in ancestral reconstruction, then describes ways to address variation in reconstructed ancestral sequences when the intent is to recreate proteins

experimentally in the laboratory. The primary purpose of this chapter is to focus on efficient experimental strategies to explore variation in ancestral sequence reconstructions. The statistical basis of this variation is addressed in detail in Chapter 8 in this volume. The experimental strategies described here are illustrated with two examples, ancestral rhodopsins in archosaurs and green fluorescent protein (GFP)-like proteins in corals.

15.2 Likelihood/Bayesian methods of ancestral reconstruction

Ancestral reconstruction methods based on a likelihood/Bayesian framework, such as those implemented in PAML (Yang, 1997), use as an optimality criterion a likelihood score, calculated according to a specified model of evolution (Felsenstein, 2004). Optimization of the likelihood score can be used to specify topology and parameters such as branch lengths, character-state frequencies, and ancestral states. Bayesian methods can also be used to estimate posterior probabilities of ancestral states. This can be done using the maximum-likelihood topology, branch lengths, and model parameters as priors (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 (Huelsenbeck and Bollback, 2001). These likelihood/Bayesian approaches can have considerable advantages over parsimony (Koshi and Goldstein, 1996; Lewis, 1998). 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, it is important to explore different models of molecular evolution to determine how robust the reconstruction results are. Oversimplified or unrealistic models can lead to incorrect or otherwise misleading phylogenetic reconstructions (Cao *et al*., 1994; Huelsenbeck, 1997; Buckley, 2002), emphasizing the importance of model selection. Likelihood models can be generally divided into three different types: nucleotide-, amino acid-, and codon-based models. Nucleotide models range from the simplest, such as Jukes–Cantor (Jukes and 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 non-stationary 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. Models have also 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 rates of substitutions in the rate matrix for all the different classes of amino acids (Yang, 1997). Rate matrices have been calculated for a number of data-sets, including those of Dayhoff (Dayhoff *et al*., 1978; Kishino *et al*., 1990) and Jones (Jones *et al*., 1992; Cao *et al*., 1994) for globular proteins, and mitochondrial transmembrane proteins (Adachi and Hasegawa, 1996). This allows a substantial reduction in the number of parameters in the model of evolution. Models have also been developed that allow replacement rates to be proportional to the frequencies of both the replaced and resulting residues ($+gwF$ model; Goldman and Whelan, 2002).

Codon-based models have been the subject of much recent development, particularly in the context of detecting positive selection, or changes in selective constraint in a phylogenetic context (Bielawski and Yang, 2003; Nielsen, 2005). These models can be among the most complex models, and have the potential to incorporate both nucleotide and amino acid information. The original codon-based models assumed equal non-synonymous to synonymous rate ratios among sites and lineages (Goldman and Yang, 1994; Muse and Gaut, 1994). Subsequent models have allowed that ratio to vary across lineages (Yang, 1998), or sites in the protein (Nielsen and Yang, 1998; Yang

et al., 2000; Wong *et al*., 2004), or both (Yang *et al*., 2005; Zhang *et al*., 2005), as well as across amino acids with different physiochemical characters such as charge, polarity, or volume (Sainudiin *et al*., 2005; Wong *et al*., 2006).

Finally, the use of genome-based approaches has enabled more extensive investigations of sources of systematic bias, or inconsistency in currently implemented methods of phylogenetic analyses (Phillips *et al*., 2004; Philippe *et al*., 2005a, 2005b; Jeffroy *et al*., 2006) and identified new effects difficult to detect in smaller data-sets, such as site-specific changes in evolutionary rates among lineages, or heterotachy (Lopez *et al*., 2002; Misof *et al*., 2002; Baele *et al*., 2006). However, these issues are only just being investigated and addressed (Kolaczkowski and Thornton, 2004; Philippe *et al*., 2005b; Steel, 2005; Thornton and Kolaczkowski, 2005; Lockhart *et al*., 2006).

Given the diversity of models now available, the exploration of different models for use in ancestral-state inference is critical. In certain cases likelihood ratio tests can be used to statistically compare two models of evolution that are nested with respect to one another, in order to determine whether the more complex model fits the sequence data significantly better than the simpler model (Navidi *et al*., 1991; Felsenstein, 2004). However, under many circumstances the models being compared are not nested, and an important alternative is to directly compare the results of ancestral reconstruction variants synthesized in the laboratory.

15.3 Laboratory synthesis of ancestral proteins

How can the variability in results from ancestral reconstruction by different models be most efficiently addressed in attempting to reconstruct ancestral proteins in the laboratory? We can distinguish two types of variability in ancestral-reconstruction results. The first is due to sites for which there are alternative reconstructions with significant posterior probabilities under a single model of evolution. However, recreation of only the most probable variant for each site does not significantly improve the chance of the whole sequence being correct, and moreover, some have recently argued that such a strategy may lead to biases in estimating the protein's properties (Williams *et al*., 2006). A more desirable approach would be to allow the ambiguous positions to vary and experimentally determine whether this would affect the protein phenotype. If the phenotype is robust to such variations, it may be assumed that the errors of ancestral sequence prediction would not matter. Ideally, to address this issue, the best approach would be to experimentally recreate the predicted posterior distribution of ancestral sequences. This would mean synthesizing a library of genes in which the proportion of each variant at each site would be equal to the posterior probability of this variant according to the prediction. Unfortunately, the construction of a degenerate gene in which the variants are represented at unequal pre-defined proportions represents a significant technical challenge. In the current version of the gene-synthesis protocol (see below), site variation is introduced through the use of commercially synthesized degenerate oligonucleotides, which, in turn, are achieved by including more than one type of nucleotide precursor at a particular step of the oligonucleotide synthesis. It is usually assumed that by controlling the proportions of the precursors in the mixture one may manipulate the proportions of the corresponding incorporation products, but this represents a significant technical challenge not routinely offered for commercially synthesized oligonucleotides. In addition to these problems, there is also the question of how many sequences would be considered adequate sampling from the posterior. These issues are addressed in detail in another chapter (see Chapter 8), and will not be discussed further here.

The second type of variability in ancestral reconstructions results from sites for which the reconstructions vary when different models of evolution are employed. Investigating the effects of this type of model variation is more easily addressed experimentally, and will be the focus of the experimental strategies discussed further here. In cases where the number of variable sites is not too high, the results of different models can be synthesized in their entirety from an initial variant using site-directed mutagenesis methods, and the

various proteins assayed for any changes in function. The major advantage of this strategy is that proteins resulting from different models can be compared directly. This strategy/method was employed for reconstructing the ancestral archosaur visual pigment protein, described in detail below. The disadvantage of this approach is that it is only feasible if the number of variable sites in the protein is small. If the number of variable sites exceeds an amount that can be easily incorporated using mutagenesis techniques, as is the case with the ancestral coral GFP-like proteins, then a different strategy needs to be employed.

In situations where the number of variable sites is high, an efficient strategy is to incorporate the variable sites directly during the gene-synthesis, instead of mutating the sites after the first protein has already been synthesized. Degenerate oligonucleotides are incorporated into the gene-synthesis methods that allow for variation found in the ancestral reconstructions. This means that sites found to be variable in the reconstructions are incorporated in a random combinatorial fashion into the synthesis of the gene. This method utilizes 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. The degenerate sites should be positioned as far as possible from the ligation points (see example in Figure 15.1). Although very simple, the method has the important advantage that the relatively short oligonucleotides can be ordered commercially, in contrast to other techniques that rely on longer oligonucleotides (Ferretti *et al*., 1986). Moreover, the oligonucleotides do not need to be modified (for example, 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 of the fact that 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-quarters of the gene length is actually proofread at the ligation step since DNA ligase is sensitive to mismatches up to at least 6 bases from the ligation site (Roth *et al*., 2004). As a result, the mutated clones in our experiments comprised 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. The advantage of this approach is that a large number of variable sites can be incorporated into the synthesis fairly easily, and many ancestral variants efficiently obtained in only one synthesis step. As long as enough variants are assayed so that a sufficient number of variable sites are represented, this approach can be extremely efficient in determining the average phenotype of an ancestral

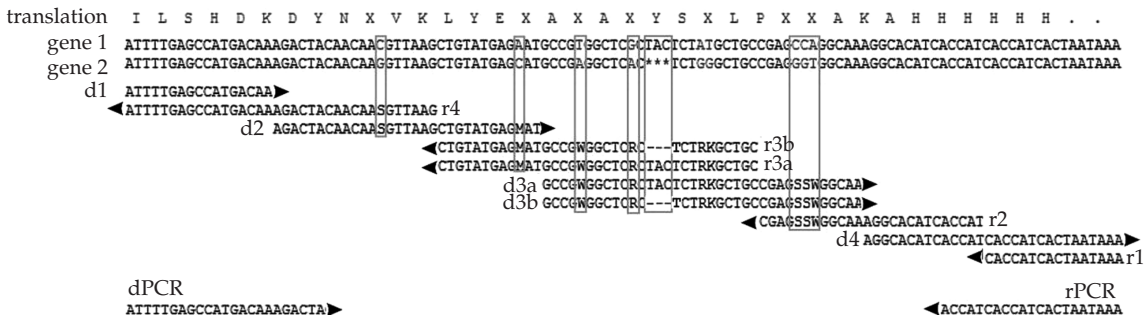

Figure 15.1 A practical example of oligonucleotide design for degenerate gene synthesis. The fragment that is being synthesized is the actual 3′-terminal portion of the highly degenerate gene incorporating transitional mutations between the ancestral identified gene 1 and the extant gene 2 (see text, Example 2). dPCR and rPCR are the PCR primers that would be used to amplify the final product. Arrowheads indicate 3′-termini of the oligonucleotides; the sequences show the corresponding portion of the sense DNA strand. Antisense oligonucleotides (on the right-hand end of sequences, starting with r) actually have the sequence complementary to the one shown. Note that to model a three-nucleotide deletion (asterisks in gene 2) two sets of sense and antisense oligonucleotides are prepared—with and without the deletion, which are then mixed in the synthesis reaction in equal proportions.

protein in the face of a large number of reconstruction variants. Unlike site-directed mutagenesis methods which can be used to synthesize the results of any one model in its entirety, the use of degenerate oligonucleotides in gene synthesis means that no one clone is likely to have the results of any one model because the variability is incorporated randomly among the sites. However, this is not necessarily a disadvantage, as the best model may vary across sites, and it is not clear that any one model should give the most accurate reconstruction for all sites.

15.4 Ancestral archosaur visual pigment

Archosaurs are a major branch of diapsid reptiles that include modern-day birds, crocodiles, and alligators (Figure 15.2). In addition to these extant taxa, the archosaur ancestors also gave rise to impressive reptiles now long vanished including several lineages of dinosaurs and pterosaurs. The closest relatives of archosaurs can be traced back into the Upper Permian, and together with the former constitute the clade Archosauriformes (Reisz and Müller, 2004; Müller and Reisz, 2005). Archosaurs in particular are thought to have originated in the Early Triassic; more specifically, the fossil record, our current understanding of archosauriform phylogeny, and the consideration of errors in stratigraphic dating techniques make it possible to constrain the date of archosaur origin to a time frame of only a few million years, i.e. between 251–243 million years ago (Müller and Reisz, 2005). Shortly after their origin, archosauriform reptiles rapidly became one of the dominant components of terrestrial ecosystems; however, what little is known of the paleobiology and paleoecology of these archosaur ancestors is inferred from the fossil record, and by analogy to their living descendants. It remains largely a mystery why they diverged so spectacularly from other diapsid reptiles such as lepidosaurs, which gave rise to modern snakes and lizards. Clearly, more knowledge about the physiology and behaviour of early archosaurs is needed, but this is difficult to gather from fossilized hard parts alone. For example, additional information about the visual adaptations of basal archosaurs would be highly desirable; in addition to its general relevance for terrestrial animals, the visual system may be of special importance for archosaurs due to their role as large predators in their respective fossil ecosystems.

The laboratory recreation of ancestral proteins involved in sensory pathways such as vision offers the opportunity to experimentally study a protein from an ancient animal to understand better the aspects of its physiology and behavior that are difficult to achieve by more traditional methods. Sensory proteins are particularly well suited for this type of study, as small changes in biochemical function can have profound consequences for the sensory capabilities of an animal (Wilkie *et al.*, 1998; Hunt *et al.*, 2001).

Visual pigments form the first critical step in the primary visual transduction cascade (Menon *et al.*, 2001). They are composed of an opsin protein moiety to which a retinal chromophore is covalently attached. It is this chromophore, 11-*cis*-retinal or its derivatives, that isomerizes in response to light, inducing a conformational change in its associated opsin protein, activating the second-messenger G-protein transducin, and triggering the biochemical cascade of events in retinal photoreceptors which constitute the signal that light has been perceived (Baylor, 1996).

In order to achieve the greater photosensitivity required for vision at low light levels, the visual system has many specializations, including a visual pigment expressly adapted for this purpose, rhodopsin. This visual pigment is found in rod photoreceptors, which are active only under dim light conditions. The ability of an animal to see well at night is thus determined, at least in part, by the functional properties of the rhodopsin contained within its rod photoreceptors. An ancestral archosaur rhodopsin sequence was inferred using phylogenetic methods and synthesized in the laboratory in order to investigate the nocturnal visual capabilities of these ancient animals (Chang *et al.*, 2002).

In this study, the ancestral archosaur rhodopsin amino acid sequence was inferred on a phylogeny reflecting systematic relationships among vertebrates for which rhodopsin sequences were available (Figure 15.3). This phylogeny has been

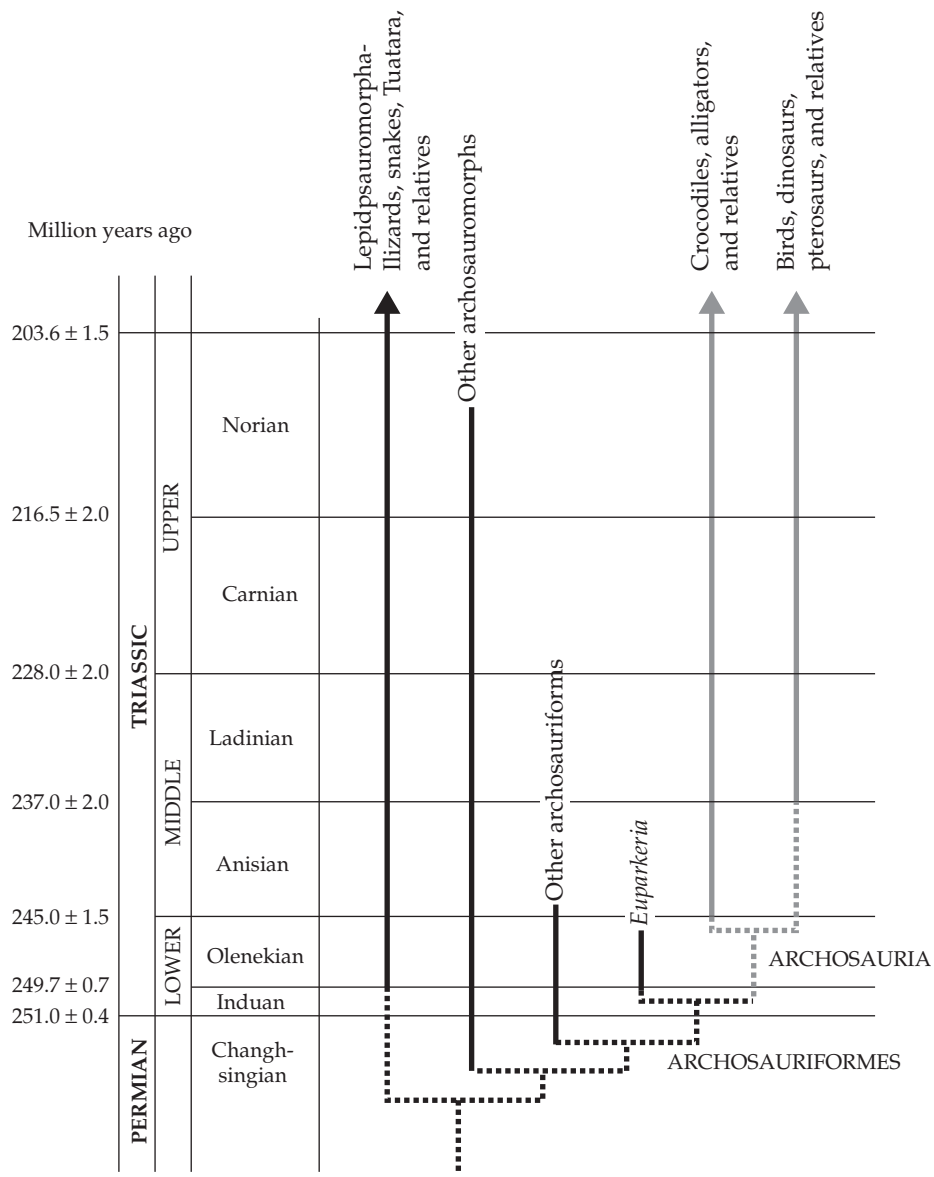

Figure 15.2 Phylogeny depicting the relationships among archosaurs (indicated in grey) and their origins from diapsid outgroups, mapped on a stratigraphy of the Late Permian and Triassic. Phylogeny, stratigraphy, and the fossil record suggest that the divergence between the two major archosaur lineages occurred in the Early Triassic, between 243 and 251 million years ago, while archosauriforms as a whole can be traced back into the Upper Permian. Current evidence also suggests that the split between archosauromorphs and lepidosauromorphs took place in the Late Permian (Reisz and Muller, 2004; Muller and Reisz, 2005).

slightly updated to reflect the current understanding of vertebrate systematics (Garcia-Moreno *et al*., 2002). Ancestral sequences were estimated for the archosaur node on this updated phylogeny using empirical Bayesian methods as implemented in PAML (Yang, 1997), with likelihood ratio tests performed where possible to determine the relative fit of the models (Navidi *et al*., 1991; Felsenstein, 2004). For the ancestral archosaur node, the amino acid reconstructions of the three best-fitting

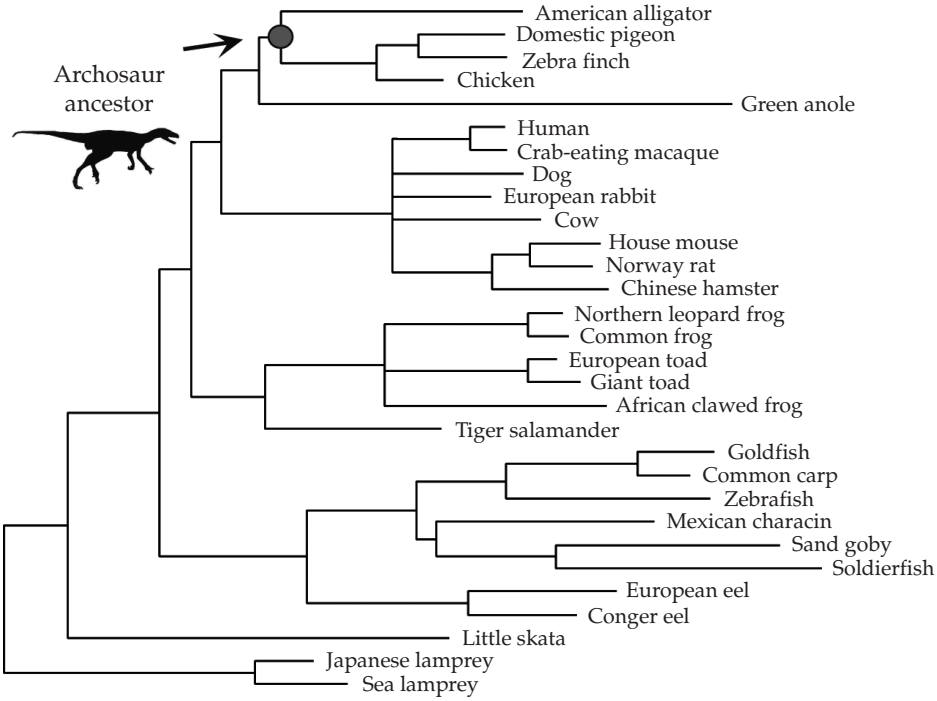

Figure 15.3 Vertebrate rhodopsin phylogeny used in inferring the sequence of the indicated ancestral archosaur node. Evolutionary relationships among vertebrates reflect current understanding of divergences among major lineages (Chang *et al*., 2002 Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**: 1483–1489, by permission of Molecular Biology Evolution; Garcia-Moreno *et al*., 2002).

models agreed at all but two sites, where one reconstruction differed from the other two. A third site for which reconstructions were found to differ on the original phylogeny now all agree for the three best-fitting models, although the alternative reconstruction remains among the possibilities with a much lower posterior probability (Figure 15.4).

The artificial gene corresponding to the inferred ancestral archosaur rhodopsin protein sequence was chemically synthesized in long fragments of up to 230 bases, and placed into a mammalian expression vector (pMT). Alternative reconstructions (T213I, T217A, and V218I; Figure 15.4) were introduced into the synthetic ancestral archosaur gene using site-directed mutagenesis methods, as implemented in the QuikChange® protocol (Stratagene). These ancestral genes were transiently transfected into monkey kidney (COS-1) cells, harvested, regenerated with 11-*cis*-retinal in the COS-cell membranes, solubilized, and immunoaffinity-purified using the 1D4 monoclonal antibody (Ferretti *et al*., 1986; Chang *et al*., 2002). The purified ancestral proteins were tested for their ability to activate transducin in a fluorescence assay, and their absorption spectra measured. Among the variants of the ancestral archosaur rhodopsin, there were no significant differences in either of these assays. All variants were able to activate transducin (original archosaur rhodopsin, 86% activation rate with respect to a bovine rhodopsin control; variants, 79–83%), and all showed a red-shifted absorption maxima relative to bovine rhodopsin of approximately 508 nm (Figure 15.5).

All variants of the ancestral archosaur rhodopsin expressed in the laboratory not only activated the second messenger in the visual transduction cascade, transducin, at least as well as a mammalian rhodopsin, but they also showed a slightly red-shifted absorption maxima characteristic of a few modern birds (Chang *et al*., 2002). These functional characteristics of the recreated ancestral archosaur protein imply that, at least for these aspects of rhodopsin function, the ancestral archosaur would have been able to see as well at night as a modern-day mammal,

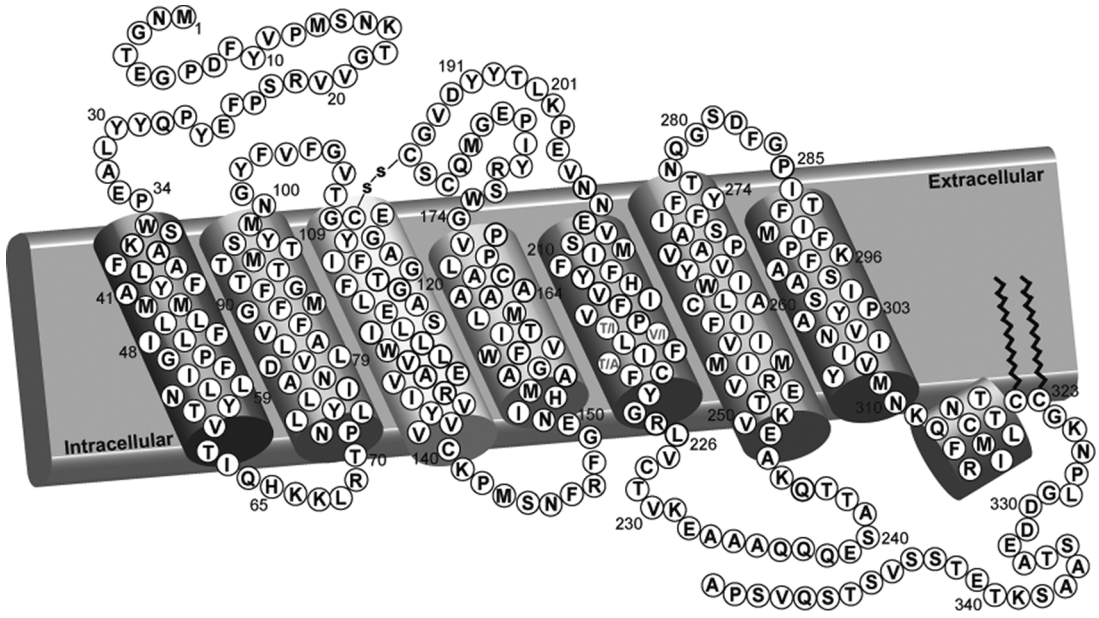

Figure 15.4 Inferred ancestral archosaur rhodopsin protein sequence, drawn in schematic form indicating the putative transmembrane domains, disulfide linkage, and palmitoylation sites. The three sites for which alternative reconstructions were experimentally introduced are indicated with slashes (in the fifth transmembrance domain); models tested agree at all other sites. For these sites at which alternative reconstructions were investigated, posterior probabilities of models employed are as follows: site 213 [F61 + G] T (0.83), I (0.078), M (0.035), V (0.028), A (0.03); [HKY + G] I (0.584), M (0.016), T (0.340), V (0.036), A (0.021); [Jones + F + G] V (0.951), I (0.048). Site 217 [F61 + G] A (0.833), T (0.102), I (0.012), M (0.004), S (0.034), V (0.015); [HKY + G]; [Jones + F + G] T (0.834), I (0.059), M (0.008), A (0.091), V (0.007). Site 218 [F61 + G] V218 (0.981), I (0.019); [HKY + G] V (0.916), I (0.083); [Jones + F + G] V (0.951), I (0.048).

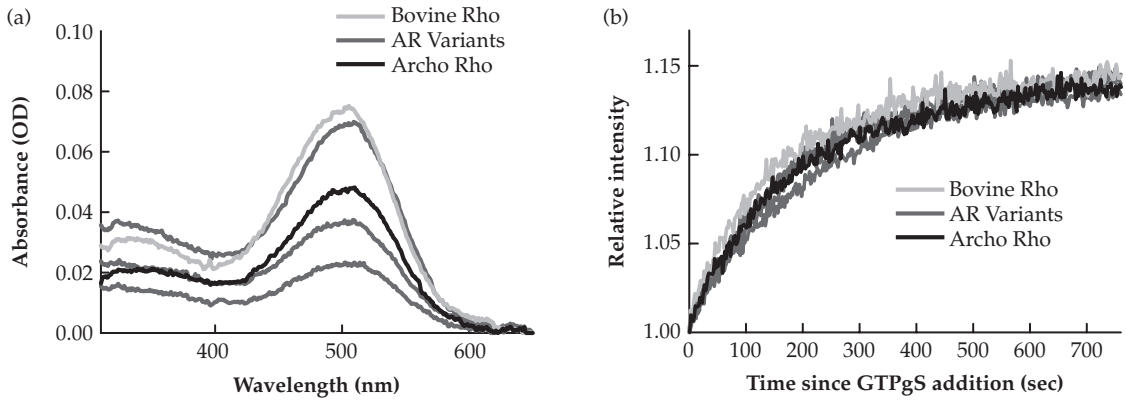

Figure 15.5 Functional assays of all variants of ancestral archosaur rhodopsin (Rho). (a) Dark absorption spectra, recorded at 25 C in a UV-visible spectrophotometer. (b) Rate of transducin activation as measured by increases in fluorescence intensity, recorded at 25 C in a spectrofluorimeter. For all assays, similarly expressed and purified bovine rhodopsin was used as a control (Chang *et al.*, 2002). Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**: 1483–1489, by permission of Molecular Biology Evolution. Original ancestral archosaur rhodopsin is indicated by black lines, variants in dark gray, and bovine rhodopsin in light gray.

which is surprising given theories of an extended nocturnal phase thought to have occurred in early mammals (Crompton *et al.*, 1978).

15.5 Ancestral GFP-like coral proteins

Coral fluorescent proteins are homologous to the GFP (Matz *et al.*, 1999; Field *et al.*, 2006). These proteins share a remarkable ability to produce a chromophore moiety autocatalytically within their own globule, using their own side chains as substrates (Heim *et al.*, 1994; Matz *et al.*, 2002). GFP-like proteins are very convenient for experimental evolutionary studies. They are small (about 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 be precisely quantified in the bacterial colonies growing on a solid medium. This provides an excellent opportunity for high-throughput screening of expression libraries, or phenotypic characterization of mutants or products of degenerate gene synthesis. Here we discuss the details of the degenerate gene-synthesis method, as well as two illustrative examples from our evolutionary studies of GFP-like protein function in corals.

15.6 Degenerate gene synthesis

15.6.1 Oligonucleotides

The artificially designed ancestral gene should be divided into overlapping oligonucleotide fragments of about 30–35 bases in length (Figures 15.1 and 15.6). 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.

15.6.2 Phosphorylation

In a 0.5-ml tube, combine 5 µl of 2 × 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 Figure 15.6), and 1 µl of T4 polynucleotide kinase (New England Biolabs). We used the buffer provided within the pGEM-T PCR

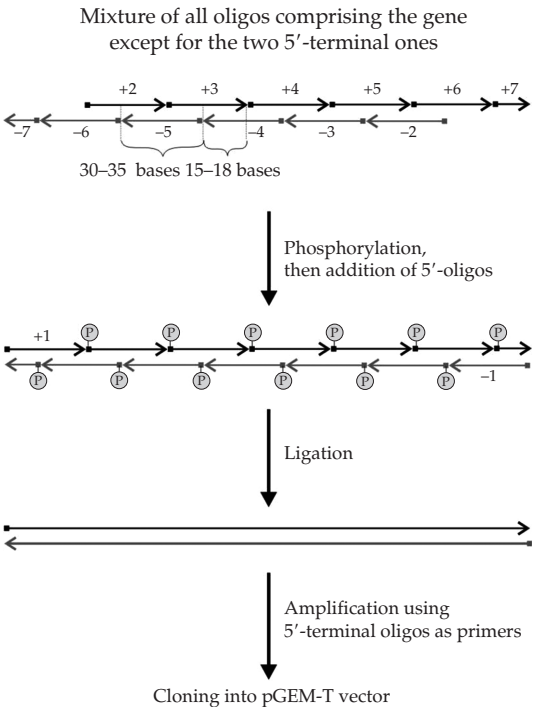

Figure 15.6 Schematic outline of the described gene-synthesis strategy. Oligonucleotides corresponding to plus and minus DNA strands are shown as black and gray arrows, respectively. Arrowheads correspond to free 3′ termini, squares to free 5′ termini. For simplicity of representation, the scheme shows the synthesis of a short fragment about 210–250 bp in length; however, the strategy will work for the longer genes as well.

cloning kit (Promega): it is similar in composition to the standard T4 polynucleotide kinase buffer, but already contains ATP in appropriate concentration. Incubate the reaction at 37°C for 30 min, then at 65°C for 20 min to deactivate the enzyme.

15.6.3 Ligation

To the completed phosphorylation reaction, add 5 μl of 2 × ligation buffer (Promega), 4 μl of the terminal oligonucleotide mixture (Figure 15.6; 0.1 μM each) and 1 μl of the T4 DNA ligase (New England Biolabs). Incubate the reaction for 2 h at 37°C.

15.6.4 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) with provided buffer. To perform the amplification, combine the following in an 0.5-ml thin-walled PCR tube: 2 μl of the ligation reaction, 2 μl of each of the 5′-terminal oligonucleotides (or specifically designed 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°C, 1 min at the annealing temperature (depends on the sequence of the primers), 1 min at 72°C (add 1 min per each 1000 bp of the synthesized gene over 1500 bp); run for 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 of amplification should become visible on a standard agarose gel after 15–20 cycles, when one-tenth of the reaction volume is loaded into the well. The PCR product is then cloned into pGEM-T (Promega) to obtain bacterial expression libraries.

15.7 Example 1: resurrecting ancestral proteins using alternative evolutionary models

The corals of the Faviina suborder exhibit four basic colors of fluorescence, each determined by a specific type of GFP-like protein: cyan (blue-shifted version of green), two slightly different shades of green, and red. Cyan proteins possess the same chromophore as greens, the blue shift being due to the modified molecular environment of the chromophore (Gurskaya *et al.*, 2001; Henderson and Remington, 2005). Red proteins, however, possess the chromophore which is the 'extended version' of the green structure (Mizuno *et al.*, 2003), requiring one additional autocatalytic reaction for its synthesis (Figure 15.7). To see how such a complex feature might have evolved, we first applied degenerate gene-synthesis methods to recreate the ancestral proteins at the nodes descending from the common ancestor of all colors (ALL ancestor) to the common ancestor of all the red proteins (Red ancestor; Figure 15.8).

The prediction of the ancestral sequences was done using three alternative maximum-likelihood models: amino acid-based JTT (Jones *et al.*, 1992), codon-based M5 (Yang *et al.*, 2000), and nucleotide-based GTR + G3 (Tavare, 1986). A small number of sites were predicted differently under different models. These differences were not due to model biases but rather to the fact that these sites were poorly predictable in general: 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, Red/green, Pre-red and Red ancestors contained nine, six, four, and six degenerate codons, respectively.

Some 500–1000 fluorescent clones from each of the four combinatorial libraries were visually surveyed using a Leica MZ FLIII fluorescence stereomicroscope with the optical filters providing excitation in the 400–450-nm range and emission from 475 nm and up (long-pass filter). Twenty-four clones from each library were sequenced and plated for spectroscopy. The fraction of clones containing no additional mutations was 0.54–0.75 (for different ancestral genes). Among these clones there were variations at all the degenerate sites. The common ancestor of all colors (ALL ancestor) turned out to be short-wave green. Most interestingly, all clones

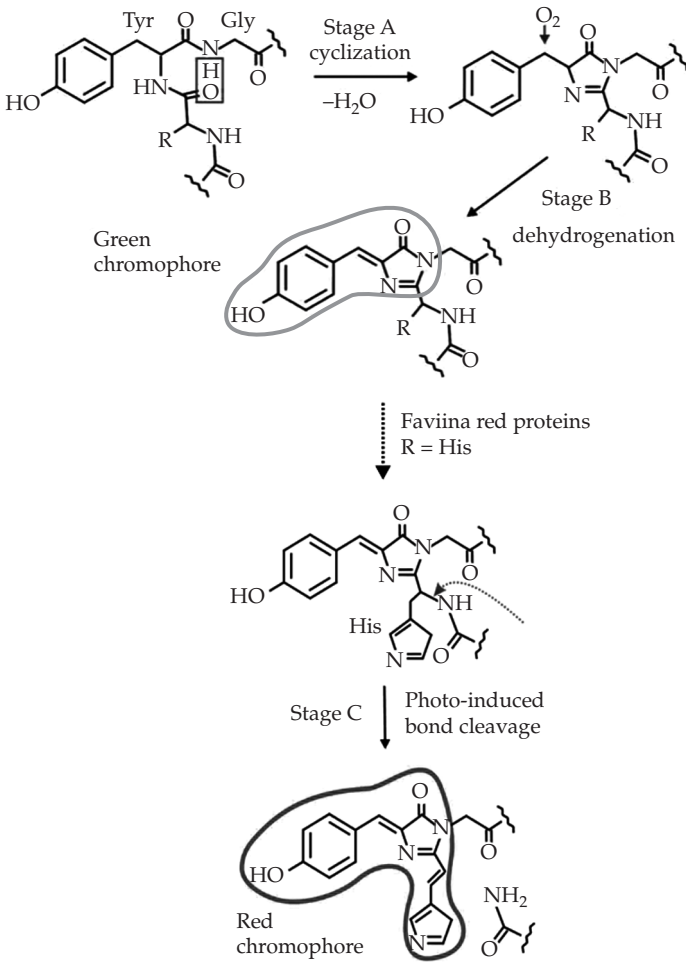

Figure 15.7 Chromophore formation pathways in green (stages A and B) and red (stages A–C) GFP-like fluorescent proteins from the corals of Faviina suborder (Heim *et al.*, 1994; Mizuno *et al.*, 2003). To synthesize the chromophore, a GFP-like protein uses its own backbone and side chains as substrates and also catalyzes all the biosynthesis reactions.

corresponding to the two possible common ancestors of red and green proteins (Red/green and Pre-red) showed an intermediate green/red phenotype: although the majority of the expressed protein remained green, a small fraction was able to complete the third chromophore maturation stage resulting in a minor peak of red emission.

These results (Ugalde *et al.*, 2004) indicate that the evolution of red emission color, which corresponds to an increase in functional and structural complexity (Shagin *et al.*, 2004), progressed through a series of intermediate stages. The follow-up analysis of color evolution in fluorescent proteins, which includes studies of selection pressure across individual sites and mutagenesis experiments, has recently been published (Field *et al.*, 2006).

15.8 Example 2: evolutionary structure–function study

Our second example of the use of degenerate gene synthesis is slightly different. To acquire further insight into the evolutionary pathway of red fluorescence from green, the minimal set of mutations, both necessary and sufficient for the green-to-red phenotype transformation among the 37 mutations separating the green common ancestor of all Faviina colors (ALL ancestor; see Figure 15.8) and the least divergent of the extant red proteins was determined. Since the evolution of red from green included intermediate stages, more than one mutation must be responsible. However, the identification of the correct combination of

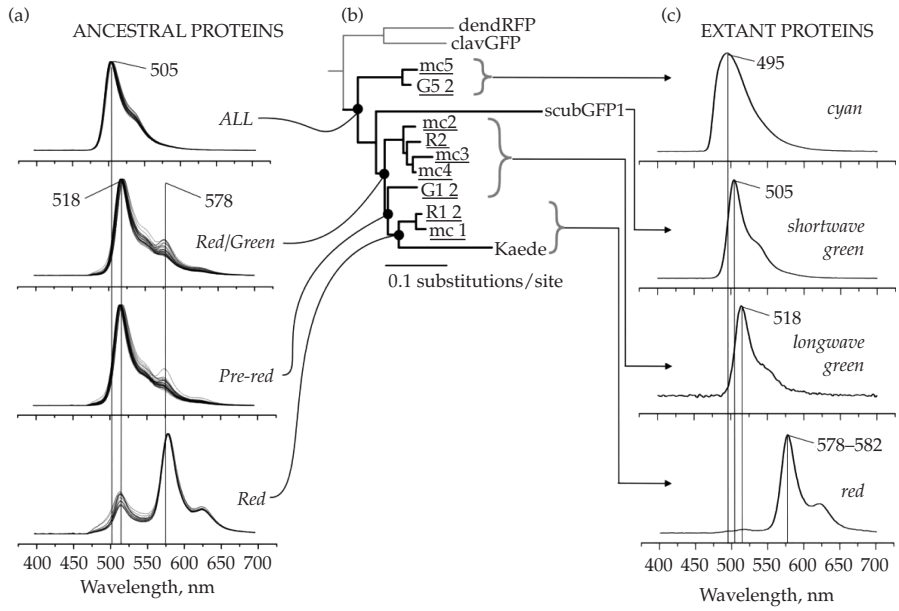

Figure 15.8 (a) Fluorescence spectra of the reconstructed ancestral proteins. Multiple curves correspond to clones bearing variations at degenerate sites. (b) Phylogeny of GFP-like proteins from the great star coral *Montastraea cavernosa* (underlined sequence names) and closely related coral species. The red and green proteins from soft corals (dendRFP and clavGFP) represent an outgroup. (c) Fluorescence spectra of extant proteins. (d) Phylogenetic tree of GFP-like proteins from the great star coral drawn on a Petri dish using bacteria expressing extant and ancestral genes, under UV-A light.

mutations among the huge number of all possible combinations presents a problem (for 37 mutations it exceeds 10^{10}). Degenerate gene synthesis was used to generate a library of all possible combinations, from which red clones were to be identified by fluorescence screening of expressed proteins. With 37 mutations, about 16% of all codons would have to contain variations, representing a dramatic jump in degeneracy in comparison to the previous example, where the most degenerate gene contained only nine (3.8%) variable codons (see Example 1, above). Screening a sufficient number of clones by fluorescence to cover all of the possible 10^{10} combinations was simply not possible. This would have been ideal, as it would have allowed for the identification of perfectly red clones, which could then be sequenced to determine the minimal number of mutations that produces the red phenotype. However, since our screening capabilities were much more modest (10^5–10^6 clones at most), a number of clones that exhibit incomplete reddening were screened (similar to Red/green and Pre-red ancestors on Figure 15.8). Their mutational composition was compared to the green clones from the same library to identify sites that tend to be found preferentially in one of the states in the redder clones, but not in green clones. The comparison to green clones in this case provides a necessary control for the mutation bias.

The degenerate gene was successfully synthesized, and sequencing confirmed that the resulting library was composed of genes with the transitional mutations in all possible combinations (Figure 15.9a). As expected, redder phenotypes were much less frequent than green ones (note that the frequencies of phenotypes in the sample presented on Figure 15.9a do not reflect the true situation due to the clone selection bias). A total of 28 reddish and 67 green clones were selected for sequencing (note that only part of this data-set is represented on Figure 15.9). It immediately appears that some variable sites indeed are much more conserved in the redder clones and thus are likely to be responsible for converting the phenotype. One prominent example is a histidine residue at GFP position 65 (position 11 in the alignment on Figure 15.9a), which is strictly conserved in all proteins demonstrating even a slight red fluorescence. This result is

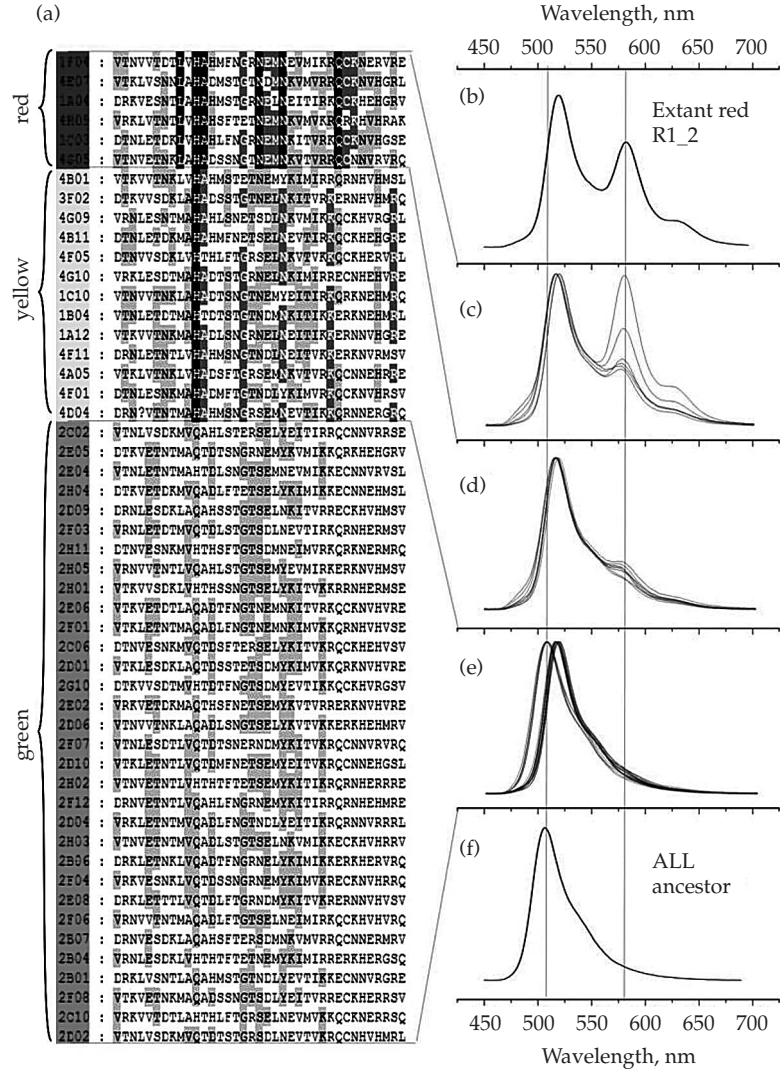

Figure 15.9 Degenerate gene synthesis to recreate a library of possible transitional variants between the green fluorescent ALL ancestor (see Figure 15.4) and the extant red fluorescent protein R1_2. (a) Alignment of only the 37 variable positions in 51 sequences from the degenerate gene library grouped according to the fluorescence color of the resulting proteins, classified as red, yellow, and green. Shading reflects the degree of within-color-group conservation. (b) Fluorescence of the extant red R1_2, measured at the intermediate stage of maturation to evaluate the efficiency of the chromophore synthesis. (c–e) Fluorescence spectra of red, yellow, and green synthetic proteins at the equivalent maturation stage. (f) Fluorescence of the ancestral green protein. On panels b–f, two vertical lines mark positions of the ancestral green and the final red fluorescence peaks.

not suprising since the side chain of His-65 forms an integral part of the red chromophore (see Figure 15.7). The Fisher exact test was used for a more rigorous comparison of site state distribution in green and red data-sets, which made it possible to rank all the variable sites in order of increasing *P* value that presumably corresponds to their importance for the red phenotype. The mutations were then introduced one by one into the green ALL ancestor in the ranking order until the protein became fully red, which required 15 mutations (interestingly, the last one had the *P* value of 0.06). Then each of the mutations was individually reversed back to the ancestral state to confirm that it was indeed

essential for the red color. This procedure eliminated three mutations, which may have appeared as high ranking in our list due to artifacts of limited sampling and multiple comparisons. As a result, we arrived at the subset of as many as 12 mutations that are needed to evolve red fluorescence from the ancestral green. Such a substantial number of mutations required for the evolution of a novel feature is remarkable: in previously investigated cases of evolution of novelties in proteins (Bridgham *et al*., 2006; Weinreich *et al*., 2006) there were no more than five mutations responsible. This result confirmed our belief that the red GFP-like proteins represent a truly unique model for studying evolution of complex protein phenotypes.

Acknowledgements

This work was funded by the National Science and Engineering Council of Canada (B.S.W.C.), the National Institutes of Health (M.V.M.), the Deutsche Forschungsgemeinschaft (J.M.), and the Canadian Institutes of Health Research (I.v.H.).

References

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

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

Baele, G., Raes, J., Van de Peer, Y., and Vansteelandt, S. (2006) An improved statistical method for detecting heterotachy in nucleotide sequences. *Mol. Biol. Evol.* **23**: 1397–1405.

Baylor, D. (1996) How photons start vision. *Proc. Natl. Acad. Sci. USA* **93**: 560–565.

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

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

Bishop, J.G., Dean, A.M., and Mitchell-Olds, T. (2000) Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. USA* **97**: 5322–5327.

Boissinot, S., Tan, Y., Shyue, S.K., Schneider, H., Sampaio, I., Neiswanger, K. *et al*. (1998) Origins and antiquity of X-linked triallelic color vision systems in New World monkeys. *Proc. Natl. Acad. Sci. USA* **95**: 13749–13754.

Bridgham, J.T., Carroll, S.M., and Thornton, J.W. (2006) Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**: 97–101.

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.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. (2002) Recreating a functional ancestral archosaur visual pigment. *Mol. Biol. Evol.* **19**: 1483–1489.

Crompton, A.W., Taylor, C.R., and Jagger, J.A. (1978) Evolution of homeothermy in mammals. *Nature* **272**: 333–336.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), pp. 345–352. National Biomedical Research Foundation, Washington DC.

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

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

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

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.

Field, S.F., Bulina, M.Y., Kelmanson, I.V., Bielawski, J.P., and Matz, M.V. (2006) Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J. Mol. Evol.* **62**: 332–339.

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

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

Garcia-Moreno, J., Sorenson, M.D., and Mindell, D.P. (2002) Congruent avian phylogenies inferred from mitochondrial and nuclear DNA Sequences. *J. Mol. Evol.* **57**: 27–37.

Gaucher, E.A., Thomson, J.M., Burgan, M.F., and Benner, S.A. (2003) Inferring the paleoenvironment 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 Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725–736.

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

Gurskaya, N.G., Savitsky, A.P., Yanushevich, Y.G., Lukyanov, S.A., and Lukyanov, K.A. (2001) Color transitions in coral's fluorescent proteins by site-directed mutagenesis. *BMC Biochemistry* **2**: 6.

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. USA* **91**: 12501–12504.

Henderson, J.N. and Remington, S.J. (2005) Crystal structures and mutational analysis of amFP486, a cyan fluorescent protein from *Anemonia majano*. *Proc. Natl. Acad. Sci. USA* **102**: 12712–12717.

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

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

Hunt, D.M., Dulai, K.S., Partridge, J.C., Cottrill, P., and Bowmaker, J.K. (2001) The molecular basis for spectral tuning of rod visual pigments in deep- sea fish. *J. Exp. Biol.* **204**: 3333–3344.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006) Phylogenomics: the beginning of incongruence? *Trends Genet.* **22**: 225–231.

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. *Comput. Appl. Biosci.* **8**: 275–282.

Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, H. N., ed.), pp. 21–132, Academic Press, New York.

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.

Kolaczkowski, B. and Thornton, J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**: 980–984.

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

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* (Soltis, P.S., Soltis, D.E., and Doyle, J.J., eds), pp. 132–163. Kluwer, Boston.

Lockhart, P., Novis, P., Milligan, B.G., Riden, J., Rambaut, A., and Larkum, T. (2006) Heterotachy and tree building: a case study with plastids and eubacteria. *Mol. Biol. Evol.* **23**: 40–45.

Lopez, P., Casane, D., and Philippe, H. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* **19**: 1–7.

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., Fradkov, A.F., Labas, Y.A., Savitsky, A.P., Zaraisky, A., Markelov, M., Lukyanov, S.A. (1999) Fluorescent proteins from non-bioluminescent Anthozoa species. *Nature Biotechnology* **17**: 969–973.

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.

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

Messier, W. and Stewart, C.B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* **385**: 151–154.

Misof, B., Anderson, C.L., Buckley, T.R., Erpenbeck, D., Rickert, A., and Misof, K. (2002) An empirical analysis of mt 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected. *J. Mol. Evol.* **55**: 460–469.

Mizuno, H., Mal, T.K., Tong, K.I., Ando, R., Furuta, T., Ikura, M., and Miyawaki, A. (2003) Photo-induced peptide cleavage in the green-to-red conversion of a fluorescent protein. *Mol. Cell* **12**: 1051–1058.

Müller, J. and Reisz, R.R. (2005) Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *Bioessays* **27**: 1069–1075.

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.

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

Nei, M., Zhang, J., and Yokoyama, S. (1997) Color vision of ancestral organisms of higher primates. *Mol. Biol. Evol.* **14**: 611–618.

Nielsen, R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**: 197–218.

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.

Philippe, H., Delsuc, F., Brinkmann, H., and Lartillot, N. (2005a) Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* **36**: 541–562.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005b) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* **5**: 50.

Phillips, M.J., Delsuc, F., and Penny, D. (2004) Genome-scale phylogeny and the detection of systematic biases. *Mol. Biol. Evol.* **21**: 1455–1458.

Reisz, R.R. and Müller, J. (2004) Molecular timescales and the fossil record: a paleontological perspective. *Trends Genet.* **20**: 237–241.

Roth, M.E., Feng, L., McConnell, K.J., Schaffer, P.J., Guerra, C.E., Affourtit, J.P. *et al.* (2004) Expression profiling using a hexamer-based universal microarray. *Nat. Biotechnol.* **22**: 418–426.

Sainudiin, R., Wong, W.S., Yogeeswaran, K., Nasrallah, J.B., Yang, Z., and Nielsen, R. (2005) Detecting site-specific physicochemical selective pressures: applications to the Class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J. Mol. Evol.* **60**: 315–326.

Shagin, D.A., Barsova, E.V., Yanushevich, Y.G., Fradkov, A.F., Lukyanov, K.A., Labas, Y.A. *et al.* (2004) GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Mol. Biol. Evol.* **21**: 841–850.

Steel, M. (2005) Should phylogenetic models be trying to ''fit an elephant''? *Trends Genet.* **21**: 307–309.

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.

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

Thorne, J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* **10**: 602–605.

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

Thornton, J.W. and Kolaczkowski, B. (2005) No magic pill for phylogenetic error. *Trends Genet.* **21**: 310–311.

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

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.

Whelan, S., Lio, P., and Goldman, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.* **17**: 262–272.

Wilkie, S.E., Vissers, P., Das, D., Degrip, W.J., Bowmaker, J.K., and Hunt, D.M. (1998) The molecular basis for uv vision in birds—spectral characteristics, cdna sequence and retinal localization of the uv-sensitive visual pigment of the budgerigar (*Melopsittacus Undulatus*). *Biochem. J.* **330**: 541–547.

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.

Wong, W.S., Yang, Z., Goldman, N., and Nielsen, R. (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041–1051.

Wong, W.S., Sainudiin, R., and Nielsen, R. (2006) Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* **7**: 148.

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

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

Yang, Z., Wong, W.S., and Nielsen, R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**: 1107–1118.

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. USA* **99**: 5486–5491.

Zhang, J., Nielsen, R., and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**: 2472–2479.