CHAPTER 8

Dealing with uncertainty in ancestral sequence reconstruction: sampling from the posterior distribution

David D. Pollock and Belinda S.W. Chang

8.1 Introduction

Whereas the evolution of morphology, particularly bone morphology, can be studied by digging up fossilized remains, DNA and proteins unfortunately do not survive the ravages of time as well. Nevertheless, the evolution of ancient protein function can be studied by inferring ancestral sequences using phylogenetic techniques (Fitch, 1971; Yang *et al*., 1995; Koshi and Goldstein, 1996) and applying gene-synthesis methods to resurrect them and assay their function *in vitro* (Geman and Geman, 1984; Malcolm *et al*., 1990; Jermann *et al*., 1995; Krawczak *et al*., 1996; Ivics *et al*., 1997; Messier and Stewart, 1997; Benner *et al*., 2000; Benner, 2002; Chang *et al*., 2002a, 2002b; Zhang and Rosenberg, 2002; Gaucher *et al*., 2003a, 2003b; Thornton *et al*., 2003; Thornton, 2004). The power of this approach lies in the opportunity to directly test hypotheses concerning the evolution of ancestral protein structure and function. Molecular evolution is a field dominated by inferences about the past based primarily on examination of present-day protein function, making this hypothesis-testing ability particularly important. Since an experimentally recreated ancestral sequence is inferred rather than observed, however, the question remains as to whether the functional features of the reconstructed protein are a true approximation of the functional features of the ancestral protein. A thorough statistical assessment and justification of the method used is therefore critical to success.

Parsimony methods have been used for ancestral reconstruction due to their ease of implementation; however, certain limitations such as the lack of an explicit model of evolution, and bias towards more frequent amino acids (or nucleotides, in the case of DNA or codon reconstructions; Collins *et al*., 1994; Zhang and Nei, 1997; Eyre-Walker, 1998; Sanderson *et al*., 2000; Krishnan *et al*., 2004), have rendered the use of model-based maximum likelihood in an empirical Bayesian approach more prevalent (Chang *et al*., 2002a, 2002b; Gaucher *et al*., 2003b; Thornton *et al*., 2003; Thornton, 2004). More recently, full Bayesian methods have also been implemented (Huelsenbeck *et al*., 2003; Ronquist and Huelsenbeck, 2003; see also Chapter 16 in this volume).

Depending on the levels of sequence divergence at which ancestral nodes are being reconstructed, there can be substantial variation at certain amino acid sites in the ancestral sequences inferred under different models of evolution. This issue of model variability in reconstructed ancestral sequences has been addressed experimentally using a variety of methods. These methods have ranged from using site-directed mutagenesis techniques to generate variants at sites differing among maximum-likelihood results from different models (Chang *et al*., 2002a), to the incorporation of degenerate oligonucleotides into the synthesis of the most likely ancestral gene, allowing for random sampling of sites that vary under different models (Ugalde *et al*., 2004). At the levels of divergence investigated thus far, these

experiments have not demonstrated significant functional differences among reconstructed ancestral protein variants. However, this issue of model variability in experimental recreation of ancestral proteins is distinct from the uncertainty that can arise under a single model of evolution, which is the main concern of this chapter. Model variability, and how best to address it experimentally, is the subject of Chapter 15 in this volume, and will not be discussed at length here.

Due to the constraints imposed by the effort and resources required to reconstruct ancestral proteins in the laboratory, most studies of ancestral protein function have by necessity tended to focus on a single optimal ancestral sequence. This is either the most parsimonious ancestor for the maximum-parsimony approaches or the most probable ancestor (MPA) for the Bayesian and empirical Bayesian approaches (Krishnan *et al*., 2004). If time has permitted, variants based on that initial sequence have also been synthesized. This was assumed to make sense, given the high costs of synthesizing and expressing proteins *in vitro*, but the potential pitfalls of focusing on the MPA were not thoroughly considered. Unfortunately, just as with maximum-parsimony reconstructions, optimality-based MPAs can be biased toward more frequent amino acid states, even when the model is correct (Krishnan *et al*., 2004). This bias in amino acid frequencies may in turn lead to biases in the inferred function of the reconstructed ancestors (Krishnan *et al*., 2004; Williams *et al*., 2006).

The goal of ancestral inference, of course, is to have as accurate a picture of ancestral function as possible, so it is worthwhile to try to understand the nature and cause of the sequence and functional bias, and how to overcome this bias. The principle source of bias in the MPA and maximum-parsimony reconstructions is (as their names imply) the choice of a ''best'' reconstruction for every site in the protein. Although this sounds intuitively preferable, the result of repeatedly making the optimal or best choice at every site is that you will preferentially choose amino acids that are more frequent at every site. If these more frequent amino acids are preferentially associated with some aspect of function, such as thermodynamic stability, then the cumulative effect will be an error in reconstruction of that aspect of function. Although an optimist might prefer to assume that such association is rare, in nearly-neutral population genetics models and in thermodynamic-based population genetics simulations, slightly deleterious variants tend to be incorporated into substitutions less often than the more fit alternatives. Simulations show that this can lead to reconstructed ancestors that are *more* stable than the true ancestral sequences (Williams *et al*., 2006).

It is important to note that the bias toward more frequent amino acids in ancestral reconstruction is due to the choice of the most probable amino acid residue at each site, from the posterior probability distribution of all amino acids at that site. To our knowledge, it has not yet been shown that there are important differences in ancestral reconstruction depending on the whether the posterior sampling comes from a full Bayesian analysis of topology and other parameters, or from an empirical Bayesian analysis that produces a marginal posterior distribution of amino acid frequencies at each node and site. In this paper, we will refer to the bias in ancestral reconstruction as optimization bias, regardless of the method of generating the posterior distribution.

In considering how ancestral amino acid frequency optimization biases might lead to ancestral functional biases, it might be assumed that biases are only a problem if the frequency of a particular amino acid residue is biased across the entire protein under consideration. This is not the case, however, since the bias arises according to the frequency of the particular amino acid residues at each site. Thus, if slightly deleterious variants are the less frequent variant at every site, it does not matter which residue is slightly deleterious, and it does not matter if these slightly deleterious residues are consistently one or a few particular amino acids, or not.

8.2 A case study

Consider the case of the ancestral archosaur visual pigment, rhodopsin (Chang *et al*., 2002a). This protein was chosen for ancestral reconstruction

analysis partly because there is very little divergence (no more than 16%) among all vertebrates, and indeed the posterior probability for the most likely amino acid reconstructions at almost all sites is in the range of 0.9–1.0 (see Figure 2 from Chang *et al*., 2002a), with only six sites having values of less than 0.9. However, in total there are, on average, nearly three (2.44) sites that would have been sampled differently if sampled from the posterior (or marginal posterior) rather than choosing the MPA at each site (on average, two of these differently sampled sites would have come from the six sites with the least likely maxima). In a random sample of 10 sequences from the posterior, the MPA sequence was never sampled, and the sequences sampled differed from the MPA sequence by up to five sites (Figure 8.1a).

We can also consider whether the amino acid frequencies are different in the most likely ancestor than in the extant vertebrates, but this turns out to be somewhat difficult to answer with any certainty. The average amino acid frequencies are different between the sites that are conserved among all vertebrates and those that are variable (Figure 8.2a), particularly for the amino acids alanine, cysteine, glycine, isoleucine, proline, arginine, serine, valine, and tyrosine. The rare variants (those observed only once) also have a notably different profile: for the amino acids noted above, the rare variants match the conserved frequencies for alanine, glycine, proline, and tyrosine, match the variable frequencies for isoleucine and arginine, and are uniquely different for cysteine, serine, and valine. Furthermore, they are notably different from both conserved and variable site averages for aspartic acid, histidine, methionine, and serine, for which the frequencies are greater, and glutamic acid, lysine, leucine, and valine, for which the frequencies are less.

Using the marginal posterior probabilities calculated in PAML (from the analysis in Chang *et al*. 2002; see also Chapter 15 in this volume), it can be seen that there are also frequency differences between the sites that are certain (posterior probability, 1.0) and uncertain at the archosaur ancestral node (Figure 8.2b). These are moderate for cysteine, phenylalanine, histidine, isoleucine, lysine, and tryptophan, but quite large for glutamic acid, glycine, proline, valine, and tyrosine. In contrast, the frequencies of amino acids in

(a)

37	54	107	112	137	173	189	213	308
F	I	A	I	V	V	V	T	V
F	V	A	I	V	V	V	T	V
F	I	A	I	V	V	I	A	M
Y	I	A	I	V	V	V	T	V
Y	V	I	I	V	V	V	A	V
F	I	I	I	I	V	V	T	V
Y	I	A	V	V	V	I	T	V
Y	V	I	I	V	V	I	T	V
F	I	A	I	V	I	I	T	V
F	V	I	I	V	V	I	A	V

(b)

```
AEFLLLIIPYVATYIKAVGEIFIVVLQTEAQMSVKT
GEFLLLVIAYVATYIWVVGEIFIVVLPTEAHFSVKT
AEFLLLIIPYVATSIWAVGEICIVILQADAQFSMKT
AEYLLFIIPYVATYIWVVGEIFIVVLPTEAQFSVKT
AEYVLLVIPYMITYIWVVGEIFIVVLPAEAQFCVKT
AEFLLLIIPYVITYIWVIGEIFIVVLPTEAQFSVKM
AEYLLLILPYVATYVWVVGEIFIVIRPTERQFSVKT
ADYLMLVIPYVIIYPWVVTNIFAVILPTEAQFSVKT
AEFLLLIIPLVATYIWVVGELFIIILPTEAQFSVST
AEFLLLVIPYVIAYIWVVGEIFIVILPAEAQFSVKT
```

Figure 8.1 Amino acid variation among 10 sampled ancestors. In (a) the sequences are random samples from the posterior distribution under the general time-reversible (GTR) model. Only the sites that were variable among the sampled sequences are shown, and residues that differed from the MPA sequence are highlighted. The MPA sequence was never sampled; three sequences differed from the MPA at one site, four sequences differed at two sites, two differed at three sites, and one sequence differed from the MPA sequence at five sites. The alignment number of each variable site is shown above the site column. In (b) a sampling of rare variants (residues observed only once at a site) is added to each random sequence. The rare variant is highlighted in dark gray, and these new sequences differ from the MPA sequence by three (two sequences), four (two sequences), five (three sequences), seven (two sequences), or 10 (one sequence) sites. The number of rare variants per sequence was determined by random sampling from a Poisson distribution with a mean of 3.43 (see legend for Figure 8.3). Each rare variant was chosen randomly from among the 103 residues observed only once at a site. For aesthetic reasons, we did not display the rare variant sites selected, but in order these are 32, 33, 40, 49, 50, 63, 71, 74, 81, 108, 111, 112, 126, 130, 150, 154, 159, 162, 194, 196, 235, 237, 273, 281, 311, 232, 319, and 349. Note that, as expected, we have occasionally sampled the same rare variant (130, 196), different rare variants at the same site (108), or rare variants at sites that were already sampled differently based on the posterior (112).

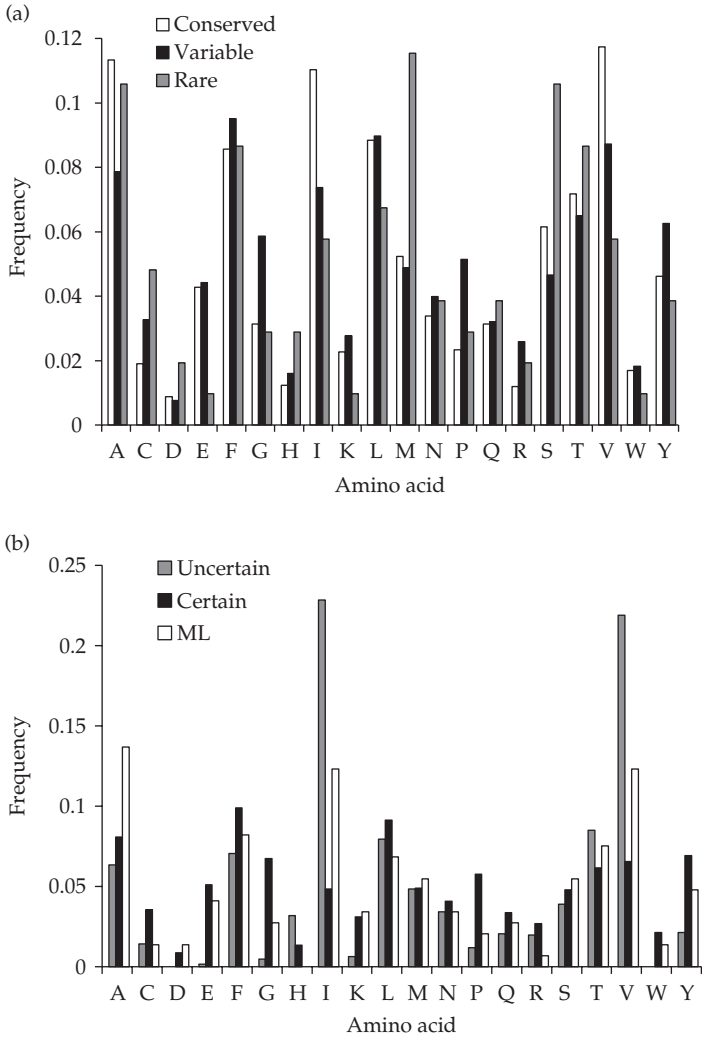

Figure 8.2 Amino acid frequencies for extant and ancestral sequences. Frequencies in extant organisms were divided into those sites that were variable or conserved among all sequences from extant organisms (a), and were also divided into those sites that were certain (posterior probability = 1.0) or uncertain (posterior probability <1.0) in the archosaur ancestor under the general time-reversible (GTR) model (b). Amino acid frequencies in the rare variants (those residues observed only once in the extant sequences at a site) and in the MPA sequence are also compared in a and b, respectively. For the uncertain sites, the frequencies are calculated based on the extant sequences, not the posterior probability at that site. ML, maximum likelihood.

the most likely archosaur reconstruction at variable sites are often midway between the certain and uncertain frequencies, but there are a number of notable exceptions, including alanine, histidine, and arginine.

The question of the effect of rare variants on ancestral reconstruction is a thorny one, mostly because it is difficult to assess their prevalence in ancestral sequences with any degree of accuracy (precisely because of their rarity). Models determined by averaging over many sites will tend to obscure variants that are rare at some sites but not at others, while the best conceivable site-specific models will not accurately assess the true frequencies of rare variants because there are not enough data at an individual site. Nevertheless, we wanted to provide an example to illustrate the possibility for rare variants to be under-sampled in this data-set, regardless of *which* amino acid is the rare variant. With over 100 variants observed only once at a site, it is reasonable to wonder whether a substantial number of low-frequency or rare amino acid variants are missing from the reconstructed ancestor. If such variants tend to affect function, then their absence could bias results.

If the number of times that a variant is observed among all vertebrates is used to estimate its expected frequency, then rare variants clearly tend to be under-represented in the reconstructed ancestor. The distribution of variant counts on a logarithmic scale (Figure 8.3a) is somewhat U-shaped, meaning that at many sites it is common to have a single dominant variant and one or more rare variants. A rough estimate of the expected number of times that a variant should be sampled in the ancestor can be calculated by weighting the variant count by the number of times each variant is observed in the vertebrates. We did this by first counting over all L sites the number of times, N_x, that any residue is observed x times at each site; that is, $N_x = \sum_i^L C_x^i$, where C_x^i is the count of variants observed x times at site i. Assuming that there is no bias in ancestral reconstruction, the expected representation of these variants in the maximum-likelihood ancestor is just the frequency of each variant count among all sequences in the alignment, $E_x = xN_x/S$, where S is the number of sequences (in this case, 30).

For example, there were 103 rare variants that were observed only once in the 30 extant vertebrates sampled, and we expect that on average $1 \times 103/30 = 3.43$ of these should have been sampled in the maximum-likelihood archosaur ancestor (whereas in practice, no rare variants were sampled in the maximum-likelihood ancestor). When expectations are compared with the number of times that variants were sampled in the

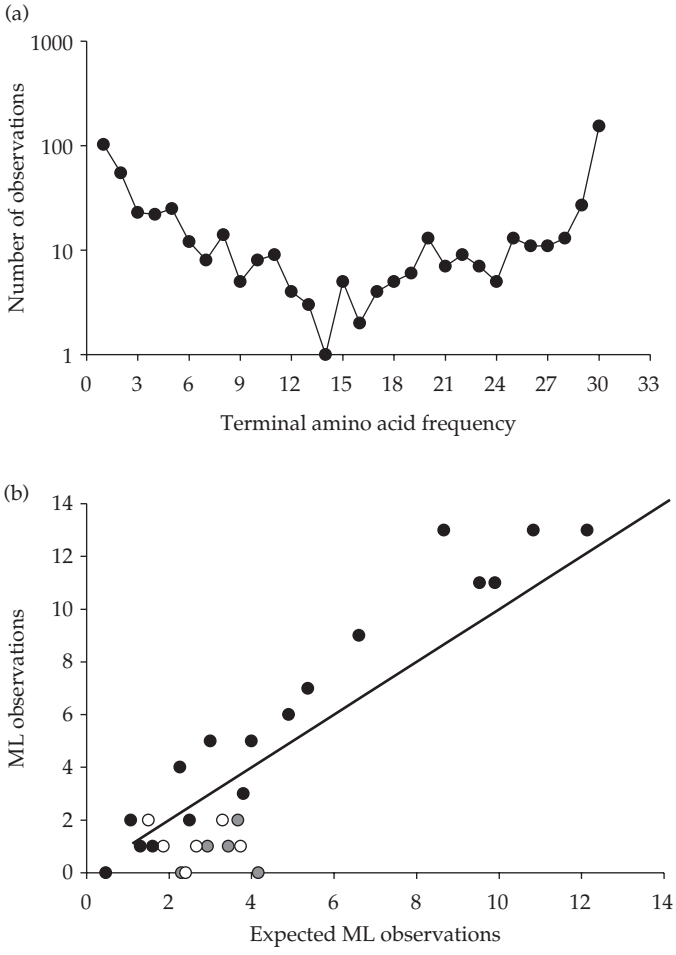

Figure 8.3 Distribution of variant counts among extant rhodopsins and predicted under-sampling of rare variants in the ancestral archosaur. The distribution of variant count observations in the extant archosaurs was assessed (a) and used to determine whether variants of a specific count were under- or over-represented in the MPA archosaur compared to expectation (b). The observed versus expected points for variants observed five or fewer times are shown as gray circles, and the points for variants observed between six and 11 times are shown as white circles. Cumulatively, variants at frequencies of 11 or fewer are observed almost 20 times less than expected. ML, maximum likelihood.

maximum-likelihood reconstruction at variable sites, it can be seen that most of the variants that were under-sampled relative to expectation were observed 11 or fewer times in the vertebrate sequences sampled (19.7 variants). An even greater proportion of under-sampling was found among those variants observed five or fewer times (13.5 variants; Figure 8.3b).

These estimates seem high, and this may be due partly to our use of a rough non-phylogenetic estimator of variant frequencies. The underestimate of rare variants at a particular node will of course depend upon the phylogenetic structure around that node, the frequencies of rare variants at each site in the alignment, and the rates of substitution to and from those rare variants at each site. Nevertheless, the estimate serves as an illustration of the potential for a substantial number of rare variants to be omitted from the MPA. Whereas it is true that the degree that rare variants are absent will depend somewhat on the phylogenetic position of a node, and the assumption that site-specific frequencies are well estimated by this (or any other method) is almost certainly wrong, it is also true that less-frequent variants will be systematically missing from MPA or maximum-parsimony reconstructions, regardless of phylogenetic position.

We took the frequency of rare variants observed only once (3.44) as a much more conservative estimate of the possible number of missing rare variants, and added a sample of those rare variants to the 10 previously sampled sequences from the posterior (Figure 8.1b). These new sequences differ from the MPA sequence at up to 10 sites. As with the bias of MPA sampling, missing rare variants may or may not contribute to functional bias in ancestral reconstruction, and it would be worthwhile to test their effect empirically. We suggest that although the number of missing rare variants is difficult to estimate, and precise knowledge of which rare variant is missing at each site may be nearly impossible to acquire, empirical testing of the cumulative effect of substituting variants that are rare in the extant species would help to gauge the potential severity of the problem.

In addition to MPA reconstruction bias, the inability of a global model to accurately account for rare variants at each site could contribute heavily to reconstruction errors. Mixture models that incorporate sharply different models tend to reconstruct function fairly accurately in thermodynamics-based population simulations (Williams *et al.*, 2006), and these may in general be preferable so long as the taxonomic sampling density is high enough that the posterior probability of the particular model at each site is relatively high (Pollock and Bruno, 2000). Whereas Bayesian methods are preferable, their accuracy may still depend on the accuracy of the substitution model used in the reconstruction, and a detailed understanding of the conditions under which model inaccuracy can lead to biased reconstruction is largely unknown. In a situation such as the example given here with so many rare variants out of 30 sequences, it is expected that due to lack of data even the best phylogeny-based site-specific models would not be able to estimate the frequency of specific rare variants accurately.

8.3 Discussion and practical recommendations

It is apparent that even for archosaur rhodopsin, our relatively ideal case study, the bias inherent in choosing to reconstruct the ancestral sequence with the highest posterior probability, along with the optimization bias due to site-specific model inaccuracy, may have biased the frequencies with which certain amino acids are inferred. Amino acids that tend to have consistently low posterior probabilities are most probably undersampled. The lack of evidence linking aspects of rhodopsin function such as absorption spectrum to rare substitutions, along with the paucity of uncertain sites and the small difference between the predicted ancestral function and the range of extant functions, lead us to expect that this amino acid sampling bias did not strongly affect the functional inference in this case. Nevertheless, whether or not this bias may in fact have affected the functional assessment of the ancestor remains to be determined. Here we discuss some of the theoretical and practical considerations with regards to this problem, and present a simple strategy for

addressing it when the goal is to reconstruct ancestral proteins in the laboratory.

8.3.1 Theoretical considerations

Determination of functional bias

This can be determined by sampling at least one ancestral sequence from the posterior distribution, in addition to the most likely (MPA) sequence. If there is no important difference in the function assayed, then the bias inherent in reconstructing the most likely ancestral sequence may not be a problem for that particular aspect of protein function. If functional differences are found between the most likely sequence and ancestral sequences sampled from the posterior, then further sampling from the posterior may be required. The number of sequences needing to be sampled will depend on the nature of the variability contained in the data-set, but in this case at least a few sequences from the posterior should be sampled; and more if resources allow. This will provide an estimate of the variability of the functional inference, rather than accepting a point estimate with unknown error. It can then be determined whether the differences between the most likely sequence and the posterior samples are significant. Relevant questions can then be asked, including whether the uncertainty in ancestral function is less or greater than any notable differences between the ancestor and some of its descendants, and whether the magnitude of uncertainty is greater or less than the difference between the most likely and average posterior functions.

Model variability

It is important to explore different models of evolution, and use as detailed a substitution probability mixture model as is justified based on the data sampled. Although this is an important issue, it is addressed in detail in another Chapter 15, and will not be discussed extensively here. It is also important to try to detect any unusual situations, such as coevolution (Pollock *et al*., 1999; Wang and Pollock, 2005), or adaptive bursts (Stewart *et al*., 1987; Messier and Stewart, 1997; Bishop *et al*., 2000; Liberles *et al*., 2001), that might add uncertainty to the posterior inference at particular sites. If the three-dimensional structure and the structural basis of the function of the protein are known, it may be useful to sample as many plausible combinations of amino acid residues as possible at sites proximal to ligand binding or enzymatic function (Chang *et al*., 2002a, 2005).

Rare variants

As rare variants may be likely to affect function, it is important to experimentally investigate their effects if possible, and to include sampling of rare variants according to their occurrence among all extant sequences. These sequences will not represent a reconstruction of the exact ancestor, but they will represent a sampling of rare variants and their effect on function. The particular rare variants that were missed in the ancestor of interest may be unknowable.

8.3.2 Experimental considerations

Methods of site-directed mutagenesis

If the number of variable sites is not large, then this may be a feasible way of generating ancestral protein variants based on the posterior distribution. The idea would be to first synthesize the MPA, and then use site-directed mutagenic methods to synthesize variants. This approach is a good one when there are only a few variable sites, but it is difficult to scale it up when there are large numbers of variable sites involved.

Degenerate oligonucleotides

Rather than introducing variability after synthesis of an ancestor, variability can instead be incorporated directly in the initial gene-synthesis steps. This is particularly easy if the gene synthesis uses relatively short oligonucleotide fragments of no longer than 50 bases in length. The use of shorter oligonucleotide fragments has been found to improve the speed and efficiency of gene synthesis over longer fragments, while retaining a relatively low error rate (Chang *et al*., 2005). Additionally, the use of shorter oligonucleotides makes it easy to introduce variability at sites in the gene-synthesis step by including oligonucleotides that are degenerate at those sites. The advantage of this approach is that even a large amount of variability

can be incorporated in a single synthesis step. It is limited, however, in that it is costly and not particularly accurate to incorporate nucleotides at unequal frequencies, such as would be necessary to sample variants proportional to their frequencies in the posterior distribution.

Variable oligonucleotide frequencies

Another means of incorporating variability during gene synthesis would be to synthesize different reasonably likely ancestral oligonucleotides, and then to mix them at the appropriate frequencies to sample from the posterior distribution. Although this method offers a significant advantage by directly incorporating ancestral variants at the frequencies in which they occur in the posterior distribution, it suffers from a major drawback as well. If more than a few variants with substantial posterior probabilities occur on the same oligonucleotide fragment, then many different oligonucleotides will need to be synthesized to encode all possible variant combinations. For a long and variable gene, this could be quite expensive. Note that this is not as much of a problem for rare variants since we can simply sample the rare variants only rarely (for example, a 5% variant in the posterior might be included at 50% frequency in an oligonucleotide mix only 10% of the time).

Sampling in silico

In contrast to the above methods, in which sampling from the posterior distribution is accomplished by experimental incorporation of proportionally sampled variants, sampling can also be separated from the gene-synthesis procedures and done instead beforehand, on a computer. In this approach, a small number of ancestral variants are sampled *in silico*. These variants can then be synthesized in the laboratory by site-directed mutagenic methods from an initial (or MPA) variant.

8.3.3 A proposed strategy

In consideration of the factors discussed above, we propose that an optimal strategy would be a hybrid approach incorporating both sampling *in silico* and a single-step gene-synthesis strategy using degenerate oligonucleotides. This approach would use sampling *in silico* to produce a reduced set of ancestral sequence variants drawn from the posterior distribution. Degenerate oligonucletides containing these restricted variants would then be mixed randomly in the synthesis step, in the proportions in which they occur in the *in silico* sampling. Such an approach would have notable advantages over creating a full library of sequences drawn from the posterior, as it would be composed of a greatly reduced number of variable amino acid sites relative to the full posterior distribution. Moreover, the number of different oligonucleotides required in the synthesis step would be greatly reduced, and degeneracy at any site could be limited to 50:50 mixes, thus greatly reducing the expense of the procedure. Note that this approach also takes advantage of the fact that sites in the protein are assumed to be independent, and can therefore be sampled independently when creating these libraries, as long as the frequencies of the states at variable sites are maintained.

It is worth noting that with a single synthesis step, the pre-sampled oligonucleotides are mixed randomly across oligonucleotide positions. In other words, although the recoverable nucleotide variants are pre-specified by the sampling *in silico*, and thus limited in number, the oligonucleotide combinations are not, and a very large number of gene sequences may be recovered. Increasing the sample size of the computer-generated sequences and thus increasing the number of variable oligonucleotides synthesized might be statistically preferable (to reduce sampling variance), but will only be worthwhile if a very large number of overall sequences are recovered and tested, which is unlikely to be feasible in many experimental systems. Using this synthesis strategy, the final result would be a set of ancestral variants that contained variable amino acid sites in the proportion that occur in the computer-sampled sequences from the posterior. Note that this synthesis strategy does not necessarily result in synthesis of the overall MPA sequence, which would need to be synthesized separately.

Although choosing to synthesize the most likely sequence (MPA), rather than sampling from the posterior distribution, can lead to biased results if

the functional consequences are also biased, these are issues that are easily addressed, and should in no way hinder the use of ancestral reconstruction as a fundamentally useful and powerful technique in comparative functional genomics. The strategy recommended above does not require any more work than is now standard in the field (for example, recent studies have tested single mutation variants from the MPA or maximum-parsimony construct, or have tested a variety of models). This is particularly true if the variants are incorporated in the initial steps of gene synthesis (instead of using mutagenesis methods afterwards). Our suggested strategy is technically feasible and will immediately provide more reliable results than synthesizing the MPA alone. While we do not yet know, in practice, how often the MPA is functionally biased compared to a sample from the posterior, it appears preferable to us to sample from the posterior in the first place.

In addition to providing a potentially unbiased reconstruction of ancestral function, incorporation of the above recommendations would allow ancestral reconstruction to provide additional fundamental information about protein biochemistry, about the sequence/structure/function relationship, and about the context-dependent effect of variants that have been accepted at some time in the evolutionary process. The distribution of effects among evolutionarily accepted variants are different from the distribution of effects from random mutations, and a better understanding of such distributions is essential to a realistic theoretical model of protein evolution.

8.4 Acknowledgments

We thank Zhengyuan Wang, Matthew Reynolds, and Judith Beekman for technical assistance in creating practical web-based perl scripts for implementing our proposed strategy (www.EvolutionaryGenomics.com). This work was supported by grants to D.D.P. from the National Institutes of Health (GM065612–01 and GM065580–01) and the National Science Foundation through Louisiana EPSCOR and the Center for Biomodular Multi-scale Systems, and from startup funds from the University of Colorardo Health Sciences Center. This work was also supported by a grant to B.S.W.C. from the Natural Sciences and Engineering Research Council.

References

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.

Benner, S.A., Chamberlin, S.G., Liberles, D.A., Govindarajan, S., and Knecht, L. (2000) Functional inferences from reconstructed evolutionary biology involving rectified databases–an evolutionarily grounded approach to functional genomics. *Res. Microbiol.* **151**: 97–106.

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.

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

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

Chang, B.S., Ugalde, J.A., and Matz, M.V. (2005) Applications of ancestral protein reconstruction in understanding protein function: GFP-like proteins. *Methods Enzymol.* **395**: 652–670.

Collins, T.M., Wimberger, P.H., and Naylor, G.J.P. (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* **43**: 482–496.

Eyre-Walker, A. (1998) Problems with parsimony in sequences of biased base composition. *J. Mol. Evol.* **47**: 686–690.

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

Gaucher, E.A., Miyamoto, M.M., and Benner, S.A. (2003a) Evolutionary, structural and biochemical evidence for a new interaction site of the leptin obesity protein. *Genetics* **163**: 1549–1553.

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

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Machine Intelligence* **6**: 721–741.

Huelsenbeck, J.P., Nielsen, R., and Bollback, J.P. (2003) Stochastic mapping of morphological characters. *Syst. Biol.* **52**: 131–158.

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

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

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

Krawczak, M., Wacey, A., and Cooper, D.N. (1996) Molecular reconstruction and homology modelling of the catalytic domain of the common ancestor of the haemostatic vitamin-K-dependent serine proteinases. *Hum. Genet.* **98**: 351–370.

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

Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., and Benner, S.A. (2001) The adaptive evolution database (TAED). *Genome Biol.* **2**: RESEARCH0028.

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.

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

Pollock, D.D. and Bruno, W.J. (2000) Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. *Mol. Biol. Evol.* **17**: 1854–1858.

Pollock, D.D., Taylor, W.R., and Goldman, N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287**: 187–198.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Sanderson, M.J., Wojciechowski, M.F., Hu, J.M., Khan, T. S., and Brady, S.G. (2000) Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol. Biol. Evol.* **17**: 782–797.

Stewart, C.B., Schilling, J.W., and Wilson, A.C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**: 401–404.

Thornton, J.W. (2004) Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* **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**: 1433.

Wang, Z.O. and Pollock, D.D. (2005) Context dependence and coevolution among amino acid residues in proteins. *Methods Enzymol.* **395**: 779–790.

Williams, P.D., Pollock, D.D., Blackburne, B.P., and Goldstein, R.A. (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computat. Biol.* **2**: 598–605.

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

Zhang, J. and Nei, M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *J. Mol. Evol.* **44**: (suppl. 1): S139–S146.

Zhang, J. 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.