36 De la Rúa, P. *et al.* (1998) Mitochondrial DNA variability in the Canary Islands honeybees (*Apis mellifera* L.). *Mol. Ecol.* 7, 1543–1547

37 Widmer, A. *et al.* (1998) Population genetic structure and colonization history of *Bombus terrestris* s.l. (Hymenoptera: Apidae) from the Canary Islands and Madeira. *Heredity* 81, 563–572

38 Arnedo, M.A. *et al.* (1996) Radiation of the genus *Dysdera* (Araneae, Haplogynae, Dysderidae) in the Canary Islands: the western islands. *Zool. Scripta* 25, 241–274

39 Avanzati, A.M. *et al.* (1994) Molecular and morphological differentiation between steganacarid mites (Acari: Oribatida) from the Canary Islands. *Biol. J. Linn. Soc.* 52, 325–340

40 Kim, S.C. *et al.* (1996) A common origin for woody *Sonchus* and five related genera in the Macaronesian islands: molecular evidence for extensive radiation. *Proc. Natl. Acad. Sci. U. S. A.* 93, 7743–7748

41 Böhle, U.R. *et al.* (1996) Island colonization and evolution of the insular woody habit in *Echium* L. (Boraginaceae). *Proc. Natl. Acad. Sci. U. S. A.* 93, 11740–11745

42 Francisco-Ortega, J. *et al.* (1995) Chloroplast DNA evidence for intergeneric relationships of the Macaronesian endemic genus *Argyranthemum* (Asteraceae). *Syst. Not.* 20, 413–422

43 Francisco-Ortega, J. *et al.* (1995) Genetic divergence among Mediterranean and Macaronesian genera of the subtribe Crysanthemidae (Asteraceae). *Am. J. Bot.* 82, 1321–1328

44 Francisco-Ortega, J. *et al.* (1996) Chloroplast DNA evidence of colonization, adaptive radiation, and hybridization in the evolution of the Macaronesian flora. *Proc. Natl. Acad. Sci. U. S. A.* 93, 4085–4090

45 Francisco-Ortega, J. *et al.* (1996) Isozyme differentiation in the endemic genus *Argyranthemum* (Asteraceae: Anthemideae) in the Macaronesian Islands. *Plant Syst. Evol.* 202, 137–152

46 Francisco-Ortega, J. *et al.* (1997) Molecular evidence for a Mediterranean origin of the Macaronesian endemic genus *Argyranthemum* (Asteraceae). *Am. J. Bot.* 84, 1595–1613

47 Francisco-Ortega, J. *et al.* (1997) Origin and evolution of *Argyranthemum* (Asteraceae: Anthemideae) in Macaronesia. In *Molecular Evolution and Adaptive Radiation* (Givnish, T.J. and Sytsma, K.J., eds), pp. 407–431, Cambridge University Press

48 Roderick, G.K. and Gillespie, R.G. (1998) Speciation and phylogeography of Hawaiian terrestrial arthropods. *Mol. Ecol.* 7, 519–531

49 Oromí, P. *et al.* (1991) The evolution of the hypogean fauna in the Canary Islands. In *The Unity of Evolutionary Biology* (Dudley, E.C., ed.) pp. 380–395, Dioscorides Press

50 Carracedo, J.C. *et al.* (1998) Hotspot volcanism close to a passive continental margin: the Canary Islands. *Geol. Mag.* 135, 591–604

Recreating ancestral proteins

Belinda S.W. Chang and Michael J. Donoghue

Tracing the history of molecular changes using phylogenetic methods can provide powerful insights into how and why molecules work the way they do. It is now possible to recreate inferred ancestral proteins in the laboratory and study the function of these molecules. This provides a unique opportunity to study the paths and the mechanisms of functional change during molecular evolution. What insights have already emerged from such phylogenetic studies of protein evolution and function, what are the impediments to progress and what are the prospects for the future?

Belinda S.W. Chang and Michael J. Donoghue are at the Dept of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 02138, USA (chang24@oeb.harvard.edu; mdonoghue@oeb.harvard.edu).

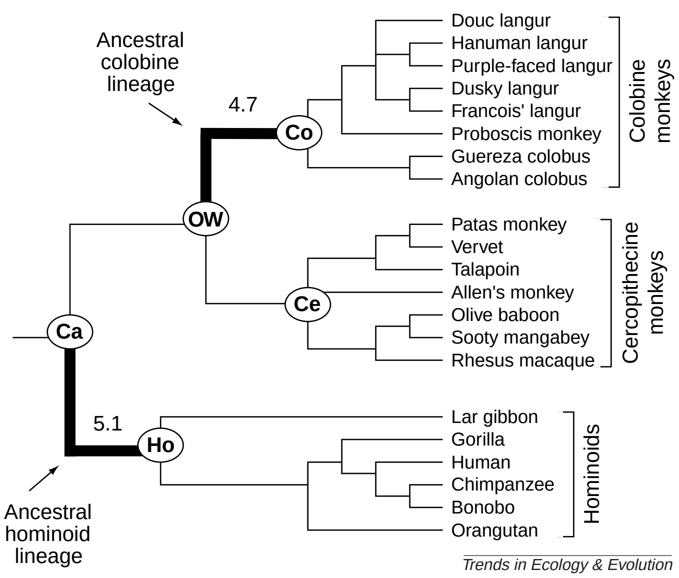
Molecular evolution leaves behind a trail of amino acid substitutions potentially rich in information about molecular function. Tracing changes in protein structure along the branches of a phylogenetic tree can provide important insights into molecular function, and the role of selection in shaping the relationship between molecular structure and function. Recreation of inferred ancestral proteins using gene synthesis and protein expression methods, whose biochemical functions can then be directly measured *in vitro*, provides a powerful approach to this problem.

Phylogenies, molecular function and natural selection

Phylogenies can be used in several ways to infer the effects of natural selection on molecular function. Because directional selection is known to elevate the ratio of nonsynonymous to synonymous nucleotide substitutions, this ratio can be used as a tool to detect lineages in a molecular phylogeny along which selection has occurred. Cows and langur monkeys convergently evolved foregut fermentation as a mechanism to digest the large amounts of plant material in their diets; a key component of this involved the recruitment of the lysozyme enzyme to digest foregut bacteria. Selection for this specialized function of lysozyme should be evident in a phylogeny of primate lysozyme sequences, on the lineage leading to colobine monkeys. By inferring and comparing ancestral lysozyme sequences, Messier and Stewart[1] were able to demonstrate adaptive change in this lineage. Specifically, they estimated the numbers of nonsynonymous (d_N) and synonymous (d_S) substitutions along each branch using Li's method[2], and found a d_N/d_S ratio much greater than one in the lineage leading to the colobine monkeys (Fig. 1). This analysis also revealed a previously unsuspected episode of positive selection in the ancestral hominoid lineage.

Subsequently, Yang[3] developed a more rigorous statistical approach to this problem, using a codon-based maximum likelihood model of evolution[4] to detect elevated d_N/d_S ratios along lineages in a phylogeny. Yang's method[3] uses likelihood ratio tests to compare the performance of different likelihood models, and to determine if the d_N/d_S ratio for the lineage of interest is elevated compared with other lineages in the phylogeny. This approach avoids using reconstructed ancestral states as if these were actual observations in the calculations of nonsynonymous and synonymous substitution rates. When applied to the primate lysozyme data, this method showed strong evidence for positive selection

0169-5347/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved. PII: S0169-5347(99)01778-4



Trends in Ecology & Evolution

Fig. 1. The phylogeny of primate lysozyme sequences, showing episodes of inferred adaptive evolution along the lineages leading to colobine monkeys and hominoids. d_N/d_S ratios of lineages significantly higher than one are shown above the thicker lines. Nodes representing the ancestors of extant primate groups of interest are labeled as follows: Co, colobines; Ce, cercopithecines; Ho, hominoids; OW, Old World monkeys; Ca, catarrhines. Ancestral nucleotide sequences were reconstructed using maximum likelihood, and pairwise d_N/d_S ratios were calculated using Li's method[2]. *Modified, with permission, from Ref. 1.*

in the hominoid lineage; however, although higher than the ratio calculated over all other lineages, the elevated d_N/d_S ratio in the colobine lineage was not significantly greater than one.

These studies[1–4] attempt to detect selection along particular lineages in a phylogeny. Such statistical analyses can establish *when* in the evolution of a protein selection might have acted, but not *where* in the protein it might have caused amino acid changes. Nielsen and Yang[5] developed a codon-based likelihood method to address the problem of detecting positively selected amino acid sites in a protein. Because positive selection should increase the number of nonsynonymous substitutions relative to synonymous substitutions at a site, they developed likelihood models that incorporated variable d_N/d_S ratios across sites (Box 1). They applied this method to HIV-1 envelope genes from viral variants sequenced at different years after the infection of a patient. The hypervariable region (V3) of these envelope genes is thought to be under positive diversifying selection during the course of HIV infection. Using their likelihood approach, they found strong support for variable selection intensity among amino acid sites, and identified positively selected sites not only in the V3 region but also in the flanking regions of the external glycoprotein (gp120). Their analysis indicated that selection acted on a broader region of the envelope gene than previously imagined, and that most of the variability in the hypervariable V3 region appeared to be neutral.

Box 1. Likelihood models for detecting positively selected amino acid sites in a protein

Positively selected amino acid sites are usually identified by an elevated ratio of nonsynonymous-to-synonymous substitutions (d_N/d_S or ω), with a ratio significantly greater than one implying strong evidence of selection at that site[44]. This is in contrast to sites that are highly conserved owing to functional constraints and thus have few nonsynonymous substitutions, and to sites evolving neutrally, which would tend towards more equal ratios of nonsynonymous and synonymous substitutions. In their codon-based maximum likelihood model, Nielsen and Yang[5] allow for three categories of amino acid sites: sites that are conserved ($\omega = 0$), sites evolving neutrally ($\omega = 1$) and sites that are the targets of positive selection ($\omega > 1$). The numerical value of ω for the category of positively selected sites is optimized in the likelihood analysis, which indicates the strength of selection for this category of sites. The accuracy of the inferred assignment of amino acid sites can be assessed using an empirical Bayesian approach to calculate the posterior probabilities of a site belonging to a particular category. This provides a measure of confidence that a particular amino acid site was the target of selection.

The likelihood approach developed by Nielsen and Yang[5] has recently been used in identifying specific sites in plant chitinases that have been subject to selection. In this case, diversifying selection has been implicated in connection with defense against fungal pathogens. Using this approach[5] on a phylogeny of crucifers related to *Arabidopsis*, Bishop and colleagues[6] were able to identify residues within the active site cleft with elevated d_N/d_S ratios, thus implying positive selection. They suggest that this selection at the active site results from changes in pathogen defenses against plant chitinolytic activity.

Finally, in addition to determining *when* in the course of molecular evolution selection might have acted, and *where* in the protein this might have occurred, one would like to establish with *what* functional changes specific amino acid substitutions were associated? These structure–function questions can be addressed using phylogeny-based comparative methods, but such approaches have received surprisingly little attention in this context. An analysis of the evolution of opsin pigments provides an example of the insights that can be gained in this way[7].

Visual pigments, which mediate vision in the eye, occur with a diversity of wavelength sensitivities that reflect the visual needs of an animal in its particular environment. Using a comparative phylogenetic approach, Chang and colleagues[7] were able to correlate specific amino acid substitutions with changes in visual pigment spectral sensitivities. This involved reconstructing amino acid changes at particular sites in the protein and looking for those amino acid changes that were repeatedly associated with shifts in sensitivity towards shorter wavelengths (Fig. 2). Four sites were identified as showing evidence of convergently evolved replacements. At these sites, changes from nonpolar to polar amino acid residues were repeatedly associated with blue shifts in opsin wavelength sensitivity. Moreover, the location of these amino acid residues within the chromophore-binding pocket of the opsin protein suggested a mechanism for these blue shifts, involving stabilization of the ground state chromophore through interactions of polar side chains with its protonated Schiff base moiety. This hypothesis has subsequently been confirmed in laboratory experiments using *in vitro* expression techniques and resonance Raman spectroscopy[8]. This kind of approach, combining comparative sequencing and mutagenesis techniques, is increasingly popular for identifying sites in the protein that affect opsin wavelength sensitivity[9,10].

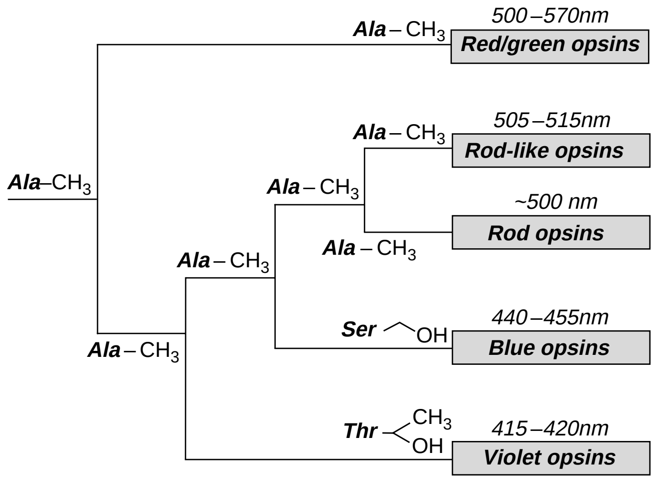
Recreating ancestral proteins

Another approach that has become feasible in recent years provides an opportunity to examine the function of inferred ancestral proteins directly in the laboratory. In the studies reviewed here, ancestral gene sequences have first been inferred, then synthesized, incorporated into an expression vector and expressed in cell culture. The resulting protein has been purified, and functional assays have been conducted to compare the activity of recreated proteins with those proteins present in extant organisms. Although moving from the statistical inference of ancestral proteins to their actual synthesis in the laboratory is a major step, this can provide unique information about the context in which adaptive replacements might have occurred, and can provide other structure–function information difficult to obtain using more traditional molecular methods. The potential payoffs have inspired a small, but rapidly growing, number of laboratory studies in this area.

An early study incorporating elements of this approach was that of Adey and colleagues[11] on L1 retroposon [long interspersed (repetitive) element 1] promoters. The F-type subfamily of L1 retroposons, found scattered throughout the mouse genome, is known to be inactive. Adey and colleagues[11] hypothesized that present-day retroposons in the mouse genome are inactive because of the accumulation of mutations in their promoter sequences. To approximate an ancestral F-type promoter they constructed a consensus sequence based on an alignment of existing promoters thought to have been active most recently. By contrast to all existing promoters in this subfamily, they found [in assays using the chloramphenicol acetyltransferase (CAT) reporter gene] that the inferred ancestral F-type promoter was indeed functional.

The use of a consensus sequence as an estimate of the ancestral promoter was possible in this study because they focused on a region with low sequence divergence. At higher levels of divergence, consensus sequences will generally contain too much uncertainty (i.e. ambiguity at particular sites), and explicitly phylogenetic methods are crucial.

One such study, by Jermann and colleagues[12], aimed to identify changes in the function of ribonucleases (RNases) related to the evolution of ruminant digestion in artiodactyls (Fig. 3). Ruminant digestion is known to generate increased amounts of RNA, requiring higher levels of specialized RNases to aid in digestion. To investigate the relationship between these specialized digestive RNases in artiodactyls, and other nondigestive RNase enzymes in animals lacking ruminant digestion, Jermann and colleagues[12] used maximum parsimony to infer the sequences of ten ancestral RNases (a–j, Fig. 3), which they then synthesized in the laboratory. Given the ancient divergences involved (which can be a problem for reconstruction methods[13]), they first conducted several experimental assays to show that these were indeed functional ancestral proteins. To do this, they demonstrated that in addition to displaying normal levels of catalytic activity against single-stranded RNA, the synthesized ancestral proteins also possessed thermal stabilities comparable with present-day digestive RNases. However, in one functional assay the ancestral proteins did show a significant difference from present-day digestive RNases in ruminants. This difference was especially marked in the early ancestral proteins (ancestors h, i and j in Fig. 3). Catalytic activity against double-stranded RNA was found to be greatly increased in these ancestors, a feature that appears to have been subsequently lost in the lineage leading to ruminants. Although the adaptive significance of this catalytic activity against duplex RNA remains unclear, it is characteristic of other non-digestive RNases, such as bovine seminal RNase (Ref. 12). This implies that contemporary digestive RNases in ruminants might have arisen from nondigestive ancestors, which were co-opted for a new digestive role concurrent with the evolution of ruminant digestion.



Trends in Ecology & Evolution

Fig. 2. A simplified phylogeny of vertebrate opsin sequences, showing convergent patterns of evolution at site 124 associated with blue shifts in opsin wavelength sensitivity. Amino acid reconstructions were obtained using parsimony, and are shown above each ancestral node with a representation of the amino acid side-chain conformation. Each box represents a clade of opsin sequences with similar spectral properties. Convergent replacements occurred in two lineages leading to short-wavelength opsins – blue and violet. These replacements were from nonpolar amino acids, such as alanine, to highly polar amino acids containing a hydroxyl group (-OH), such as serine and threonine. These convergent substitutions were inferred to be adaptive because they are both associated with shifts in opsin spectral sensitivity towards shorter wavelengths. This theory, along with the model explaining the molecular mechanisms underlying these blue shifts, has subsequently been confirmed by laboratory studies[8]. *Modified, with permission, from Ref. 7.*

Synthesizing an entire family of ancestral sequences is particularly useful for examining patterns of molecular function across the phylogeny. It can also be useful (and practical) to focus on one or more nodes of special interest, where a particular change in function is predicted to have occurred. Chandrasekharan and colleagues[14] focused on synthesizing the protein inferred for one ancestral node of particular interest in a phylogeny of leukocyte serine proteases. These enzymes are known to cleave peptide bonds at the C-termini of aromatic residues, such as Phe (phenylalanine), Tyr (tyrosine) and Trp (tryptophan). Chymases, which form a distinct group within the larger family of mast cell serine proteases, include some members, such as human chymase, that display highly selective hydrolysis of specific peptide bonds such as the Phe8-His9 bond in Angiotensin I to form Angiotensin II. By contrast, other members, such as rat chymase-1, do not show such highly specific activity and instead readily degrade both Angiotensin I and II.

Chandrasekharan *et al.*[14] tested the hypothesis that early serine proteases were simple degradative enzymes, and that more specialized functions, such as restricted substrate specificities, were only acquired later in evolution. They used parsimony to infer the sequence of the ancestor to the chymase family, and showed in the laboratory that it could convert a specific substrate (i.e. Angiotensin I to Angiotensin II) with high efficiency. This indicates that chymases evolved from an ancestor with highly specific activity, with some members of the family losing this specificity later in evolution. This argues against the hypothesis that serine proteases always evolved from general to more specific functions.

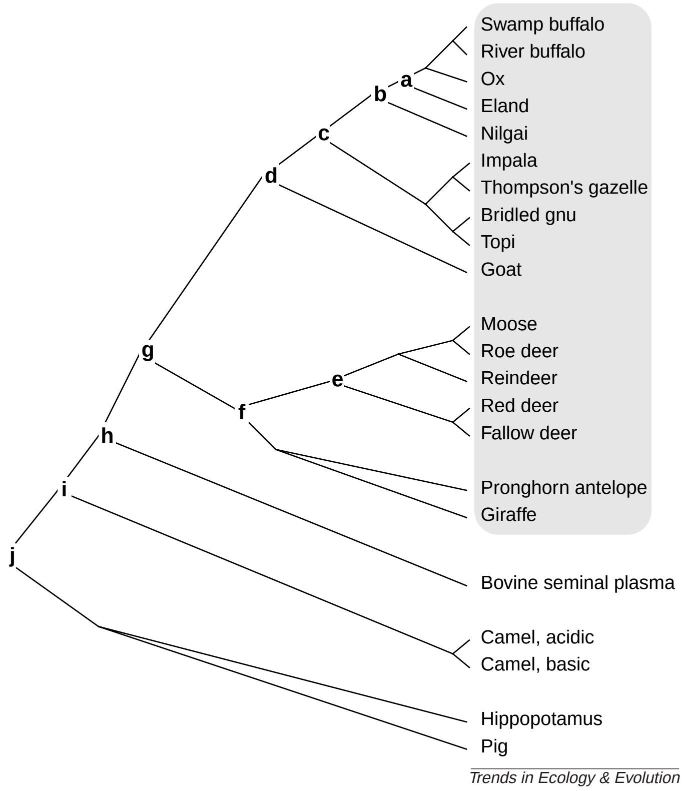


Fig. 3. Phylogeny of RNase sequences in artiodactyls. Ten ancestral proteins (a–j) were inferred using parsimony at the amino acid level. True ruminant artiodactyls are shaded in grey. The ancestors 'h', 'i' and 'j' were found to possess a fivefold increase in catalytic activity against duplex RNA, relative to more recent ruminant ancestors such as 'g'. A reduction in activity against duplex RNA is thought to have been concurrent with the evolution of ruminant digestion in artiodactyls. *Modified, with permission, from Ref. 12.*

Where possible, it is preferable to reconstruct entire sequences, because this allows changes to be viewed in the context of the entire ancestral protein. However, it is not always possible to reconstruct an entire ancestral sequence reliably at all sites in the protein. In such cases, a complementary approach that is becoming popular is to use site-directed mutagenesis to reconstitute a few key amino acid substitutions without reconstructing and synthesizing entire ancestral molecules. For example, Dean and Golding[15] investigated the evolution of eubacterial NADP-dependent isocitrate dehydrogenases using this approach. Comparing the X-ray crystallographic structure of *Escherichia coli* isocitrate dehydrogenase (which uses NADP as a substrate) with that of a distantly related protein, isopropylmalate dehydrogenase (which uses NAD as a substrate), they identified key residues in the coenzyme binding pockets that appeared to interact with the nucleotide substrate. The same residues that appeared important for coenzyme specificity based on such structural predictions, were also found to be amino acid replacements in ancestral sequences inferred using maximum likelihood. Using site-directed mutagenesis, they showed that specifically changing these residues, in either coenzyme, inverted its binding specificity.

Inferring ancestral sequences

A crucial step in all such studies is the inference of ancestral protein sequences. If great care is not taken at this stage in the analysis, the rest of the analysis will be compromised and the results could even be misleading. The inference step is perhaps especially crucial when the plan is to proceed to recreate ancestral proteins in the lab. Such studies require a greater commitment of time and of money, and their feasibility and value will depend heavily on obtaining as accurate and unambiguous an inference as possible.

In an ideal situation, one would start with a well supported phylogeny and the relevant ancestral sequences would be unambiguously determined. However, this is rarely the case. Given that a typical protein is composed of hundreds of amino acid sites, all of whose ancestral states must be inferred, it is extremely rare that all sites can be reconstructed unambiguously. Moreover, often the most interesting evolutionary changes in biochemical function do not occur at extremely low levels of sequence divergence, where ancestral states are more easily inferred. This can be a problem, particularly when not only are the sequences highly diverged but also the rates of evolution vary[13].

Our intention is not to provide a comprehensive review of phylogenetic methods or methods for inferring ancestral states, because these can be found elsewhere[16–19]. Instead, we highlight methodological considerations we believe to be most directly relevant when the goal is to proceed to reconstruct sequences in the lab. One obvious source of error in inferring ancestral states is the lack of resolution of the tree itself or the existence of a variety of plausible trees. If this problem cannot be overcome, it seems highly advisable to thoroughly explore the sensitivity of the inferred ancestral sequences to alternative resolutions of poorly supported nodes in the tree[20,21]. Even when one has great confidence in a single tree, there might still be ambiguity in inferences of ancestral states at particular internal nodes. Here, we focus attention on ways to proceed in the face of such ambiguity.

Parsimony methods evaluate phylogenetic relationships and ancestral state assignments based on the amount of evolutionary change along the branches of the tree – specifically, trees or ancestral states that require the fewest changes are preferred[16]. Although weighted parsimony methods and the use of step-matrices can accommodate rather complex models of character change, it is difficult to correct for multiple substitutions at a site in an explicit model of evolution or to take branch lengths into consideration[17,22–26]. Furthermore, in practice it is clear that equivocal assessments of ancestral states are common using parsimony. Even a small percentage of such ambiguities scattered along an entire sequence could yield a large number of combinations, which might then require the examination of a large number of proteins in the lab. For these reasons, it is important to consider alternative strategies to narrow down the possibilities[20,23,27].

Likelihood methods use a likelihood score as an optimality criterion, calculated according to a specified model of evolution[28]. This likelihood score represents the probability of observing the sequence data, given a particular tree topology and model of evolution, and is maximized in reconstructing phylogenetic relationships and ancestral sequences. Likelihood methods rely on an explicit model of evolution, and also make use of branch length information. An explicit model allows 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 (Box 2)[21,25,27]. This means that sites that have ambiguous ancestral state assignments under parsimony can be explored under different models in likelihood[20,27], thus making use of specific probabilities associated with particular ancestral reconstructions. This information can be extremely useful in narrowing down to one or a few reconstructions for the purpose of designing ancestral proteins for synthesis in the lab.

If differences in ancestral reconstructions depend on the choice of evolutionary model, then choosing a realistic model is crucial. Not surprisingly, this has recently been a focus of attention. Likelihood models describe molecular evolution at three different levels: nucleotide, amino acid and codon. Nucleotide models[28–33] range in complexity from those assuming equal base frequencies and equal rates of transition and transversion[29], to those incorporating unequal numbers of substitutions among all the different classes of nucleotides in the rate matrix[32] and those allowing for nonstationary base composition[33]. Amino acid models tend to be even more parameter-rich, because they involve 20 states instead of only four[31,34–41]. A significant advantage of amino acid models is that they avoid many of the problems associated with models at the nucleotide level, such as nonstationary base compositional biases. However, in these models, all nucleotide-encoded information is lost, including that potentially relevant to the task of ancestral reconstruction. Codon-based models of molecular evolution are among the most recent developments, and have the advantage of incorporating information on both nucleotide and amino acid levels. The original codon-based models assumed equal nonsynonymous to synonymous rate ratios among sites and lineages[4,42]. Subsequent models have allowed that ratio to vary across lineages and among sites in the protein[3,5], and have even incorporated unequal frequencies of different types of nonsynonymous substitutions based on the nature of the amino acids involved[43].

Prospects

As the studies reviewed here have demonstrated, tracing the history of proteins through phylogeny can help determine when and how selection might have acted, and might even identify specific amino acid residues that were the targets of selection and important in molecular function. Recreating inferred ancestral proteins in the laboratory is a promising approach, allowing a direct assessment of the function of ancestral molecules. The generation of specific structure–function hypotheses using these approaches can help orient laboratory studies on extant proteins using site-directed mutagenesis and *in vitro* protein expression.

However, as we have highlighted, close attention to uncertainties associated with phylogenetic hypotheses and to the methodological issues associated with the inference of ancestral protein sequences is absolutely crucial[13]. Where it is possible to infer ancestral sequences with confidence, laboratory studies of the molecular function of ancestral proteins can provide a variety of insights that would be difficult to obtain in any other way. Inaccurate assessments severely compromise the entire approach. In view of the time and energy invested in such analyses, especially those in which proteins are synthesized in the lab, it seems wise to make every effort to thoroughly investigate alternative approaches to the inference problem, and to take advantage of newly developed maximum likelihood approaches. In this connection, attention to the development of reasonable models of evolution is of the utmost importance. This is crucial because it will often be highly desirable to focus on only a few sequences, which are examined closely, as opposed to having to examine every possible resolution of uncertainties at a variety of sites. It is also important to note that the computational intensity of likelihood methods is primarily an impediment in inferring phylogenetic relationships, and is not a major problem for the inference of ancestral states using a given tree, which can be calculated in a reasonable amount of time even for fairly large data sets[23].

Box 2. Choosing a model of evolution using likelihood ratio tests

An inappropriate model of evolution can lead to inconsistency of the likelihood analysis, and convergence to an incorrect result[16,24,45]. This possibility can be reduced by selecting a model of evolution that displays a good fit to the sequence data at hand.

Likelihood ratio tests allow the comparison of 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[28,46,47]. For nested models, a more complex model (H_1) will contain all the parameters of the original model (H_0), as well as additional parameters. If the models are not nested, they cannot be directly compared using a likelihood ratio test; other methods, such as the generation of the distribution of the test statistic using Monte Carlo simulation, must be used[47]. For nested models, a more complex model (H_1) should fit the data better than a simpler model (H_0), as judged by the likelihood score or the natural logarithm of the likelihood of each model (L_0, L_1). If H_0 is correct, this difference in fit to the data can be approximated by a χ^2 distribution, with degrees of freedom (df) equal to the difference in the number of parameters between the two models[48]:

$$2(L_1 - L_0) = \chi^2_{[df]}$$

However, if the observed difference is greater than the χ^2 critical value, then the simpler model (H_0) will be rejected and the more complex model (H_1) will be preferred. In other words, in this case, the more complex model fits the data even better than would be expected because of its additional parameters relative to the simpler model.

The studies reviewed here represent only a fraction of the possibilities. To date, the focus has largely been on identifying changes in a few key sites that have dramatically altered protein function during the course of evolution. Many questions remain. For example, are there also examples of the gradual accumulation of amino acid substitutions, eventually resulting in large functional differences? Not only can these and other such questions of biochemical function be formulated in evolutionary terms, perhaps now they can be answered in that context as well.

Acknowledgements

B.S.W.C. is supported by an NSF/Alfred P. Sloan postdoctoral fellowship in Molecular Evolution. The authors thank Z. Yang, C-B. Stewart, D. Campbell and two reviewers for their comments on this article; and S.O. Smith and T.P. Sakmar for the bovine rhodopsin coordinates pictured on the cover. B.S.W.C. is currently at the Rockefeller University, 1230 York Avenue, Box 284, New York, NY 10021, USA (changb@rockvax.rockefeller.edu).

References

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

2 Li, W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36, 96–99

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

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

5 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

6 Bishop, J. *et al.* Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution. *Proc. Natl. Acad. Sci. U. S. A.* (in press)

7 Chang, B.S.W. *et al.* (1995) Opsin phylogeny and evolution: a model for blue shifts in wavelength regulation. *Mol. Phylog. Evol.* 4, 31–43

8 Lin, S.W. *et al.* (1998) Mechanisms of spectral tuning in blue cone visual pigments. Visible and Raman spectroscopy of blue-shifted rhodopsin mutants. *J. Biol. Chem.* 273, 24583–24591

9 Fasick, J.I. *et al.* (1998) The visual pigments of the bottlenose dolphin (*Tursiops truncatus*). *Vis. Neurosci.* 15, 643–651

10 Yokoyama, S. *et al.* (1999) Adaptive evolution of color vision of the Comoran coelacanth (*Latimera chalumnae*). *Proc. Natl. Acad. Sci. U. S. A.* 96, 6279–6284

11 Adey, N.B. *et al.* (1994) Molecular resurrection of an extinct ancestral promoter for mouse L1. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1569–1573

12 Jermann, T.M. *et al.* (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374, 57–59

13 Schluter, D. (1995) Uncertainty in ancient phylogenies. *Nature* 377, 108–110

14 Chandrasekharan, U.M. *et al.* (1996) Angiotensin II – forming activity in a reconstructed ancestral chymase. *Science* 271, 502–505

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

16 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (2nd edn) (Hillis, D.M. *et al.*, eds), pp. 407–514, Sinauer

17 Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521–565

18 Schluter, D. *et al.* (1997) Likelihood of ancestor states in adaptive radiation. *Evolution* 51, 1699–1712

19 Cunningham, C.W. *et al.* (1998) Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol. Evol.* 13, 361–366

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

21 Donoghue, M.J. and Ackerly, D.D. (1996) Phylogenetic uncertainties and sensitivity analyses in comparative biology. *Philos. Trans. R. Soc. London Ser. B* 351, 1241–1249

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

23 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. *et al.*, eds), pp. 132–163, Kluwer

24 Yang, Z. (1996) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42, 294–307

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

26 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410

27 Yang, Z. *et al.* (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650

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

29 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

30 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

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

32 Yang, Z. (1994) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39, 105–111

33 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

34 Bishop, M.J. and Friday, A.E. (1985) Evolutionary trees from nucleic acid and protein sequences. *Proc. R. Soc. London B Biol. Sci.* 226, 271–302

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

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

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

38 Kishino, H. *et al.* (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 31, 151–160

39 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282

40 Cao, Y. *et al.* (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

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

42 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

43 Yang, Z. *et al.* (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* 15, 1600–1611

44 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press

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

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

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

48 Navidi, W.C. *et al.* (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.* 8, 128–143