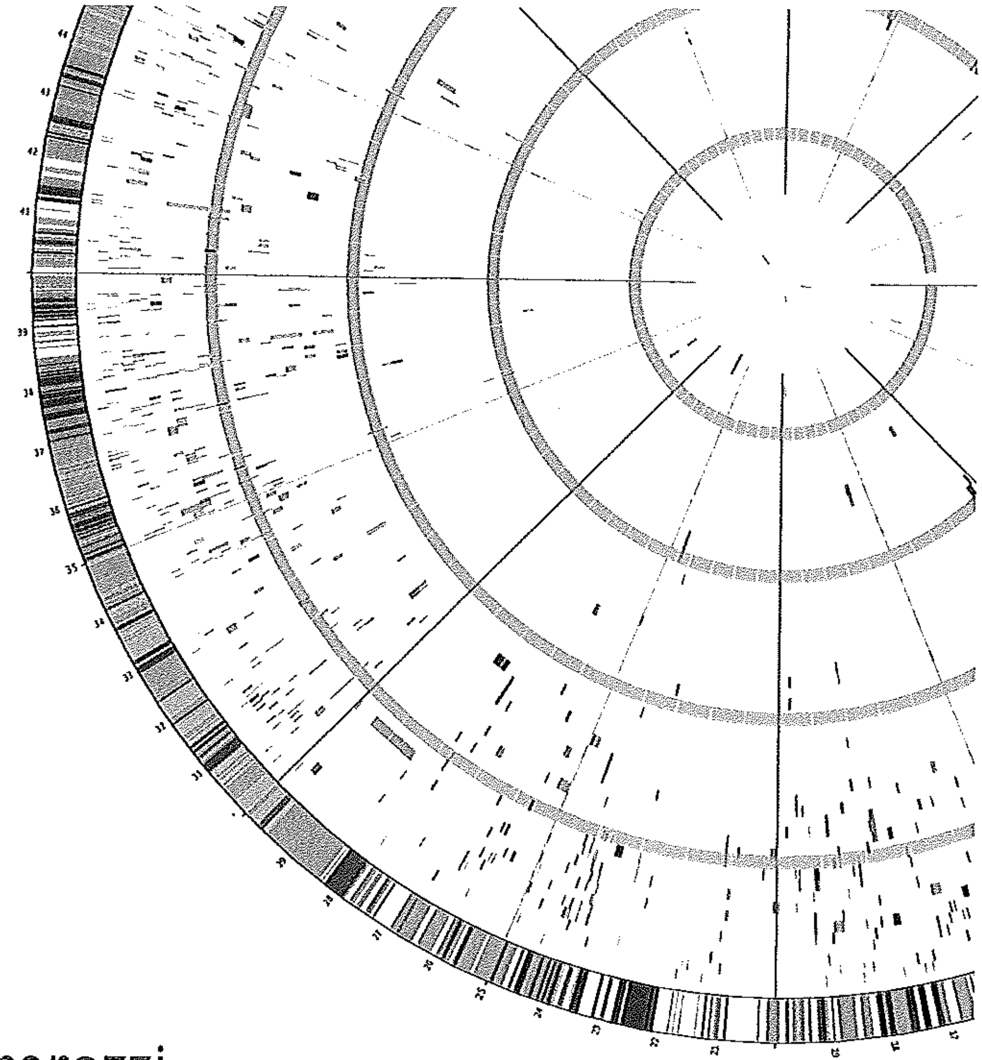
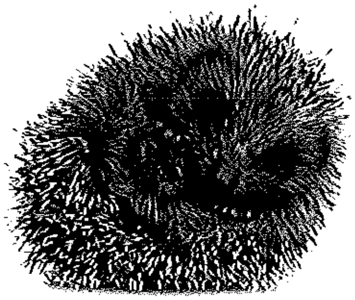
OXFORD



Edited by
Gina M. **Cannarozzi**
Adrian **Schneider**



codon evolution

MECHANISMS AND MODELS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Oxford University Press 2012

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First published 2012

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Library of Congress Control Number: 2011944051

Typeset by SPI Publisher Services, Pondicherry, India
Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

ISBN 978–0–19–960116–5

1 3 5 7 9 10 8 6 4 2

CHAPTER 10

Use of codon models in molecular dating and functional analysis

Steven A. Benner

10.1 Introduction

Modelling the divergence of the sequences of families of homologous genes and proteins is an interesting challenge for bioinformaticians. Ultimately, however, the value of models resides in their use by biologists to solve problems that they find interesting. Evolutionary biologists, for example, use models of sequence divergence that make direct statements about the historical relationships between parts of gene and protein sequences. For example, a phylogenetic tree models the historical familial relationships between the sequences themselves and, possibly, the organisms that carry them. A multiple sequence alignment of a set of homologous sequences states that the aligned codons all descended from single codons in an ancestral gene. As another example, the sequences of ancestral genes and proteins are inferred in building a tree from an alignment. Paleogenetics find these useful targets for molecular resurrection using modern biotechnology, delivering a bit of antiquity to the laboratory where it can be studied to improve our understanding of the intimate interaction between chemistry and biology (Benner, 2007).

Other biologists have different goals. For example, structural, molecular, cellular, and organismic biologists need have no direct interest in molecular evolution. However, they are often interested in the fold, or three-dimensional structure, of members of a protein family. They might look to patterns of variation and conservation in phylogenetic models to understand the positioning of individual segments of the peptide chain in that fold (Benner *et al.*, 1997).

Functional biologists, including most molecular, cellular, and organismic biologists, are less concerned about these elements of biophysics, but ask how the protein as a whole 'functions'. According to Darwinian theory, a statement about function explains how a gene and its encoded protein confer fitness upon a host organism. This chapter is concerned with how detailed analysis of gene sequences within the context of the genetic code can help make these statements.

10.2 The level of analysis most useful for functional biology

Nearly all analysis of protein sequence divergence exploits a Markov model that describes amino acid replacements. Historically, such models had replacements occurring independently at each site, with future replacements independent of past replacements and applying the same replacement matrices to all sites. Such models are not bad approximations for the divergent evolution of most protein families. Nevertheless, when the observed divergence differs from that expected under these Markov models, these differences can be used. Most successful has been the use of 'non-Markovian' behaviour to allow structural biologists to predict the fold forms of proteins. This has been well documented in blind prediction 'contests', such as for protein kinase, protein phosphatase, interleukin, phospho-beta-galactosidase, synaptotagmin, heat shock protein 90, and ribonucleotide reductase (Benner *et al.*, 1997).

Replacing the n sites in a standard protein alignment by the $3n$ sites in the encoding gene clearly adds information. This information helps resolve the topology of trees around short recent branches, align regions around indels, and meet other goals of building phylogenetic models interesting to evolutionary biologists. Further, analysis at this level offers information about the detailed chemical processes by which mutation occur, through DNA damage, polymerase error, or failures in repair processes.

However, a full DNA analysis overlooks the very real difference between non-silent nucleotide substitution (constrained by the full power of natural selection) and silent nucleotide replacements (only weakly constrained by natural selection, if at all). The second is more likely to reflect underlying chemical processes and is more likely to behave like a molecular clock. The first is more likely to reflect functional change. These peculiarities of terran biology provide the impetus for codon models in functional analysis, where 61 characters are used as the building blocks for divergent sequence evolution. The foundational paper in this area, by Goldman and Yang (1994), is discussed in Chapter 2.

This analysis need not be sophisticated. For example, early in the genomic age, Brenner (1988) noted that a serine (Ser or S) in the active site of a family of serine proteases was encoded by AGY (Y is a pyrimidine, either C or T) in one branch of the family and by TCN (N is any nucleotide) in another. The alignment of the two codons appeared to be indisputable, as the serine itself is absolutely critical for the catalytic function of the protein, is superimposed in the crystal structure of the proteins, and is flanked in the sequence by other well-conserved amino acids. However, absent simultaneous double nucleotide substitution, either TGC (cysteine or Cys) or ACC (threonine or Thr) must have been an intermediate in the divergent evolution of this serine codon. This implies that in the past, an active protease used either cysteine or threonine instead of serine in its catalytic mechanism. The observation was so striking that it was published in *Nature* (Brenner, 1988).

A more comprehensive tool for functional analysis exploits codon models that simply distinguish between 'silent' synonymous substitutions and 'non-silent' nonsynonymous substitutions. If one assumes that natural selection operates only at the level of the protein, synonymous and nonsynonymous substitutions are expected to accumulate at different rates within a gene family over time.

Under this assumption, synonymous nucleotide substitutions are 'neutral' with respect to the fitness of their host organism. Under the neutral theory of evolution, they therefore are fixed in a population at a rate inversely proportional to the size of the population. As the number of neutral mutations occurring in a population scales linearly with the size of a population, the extent of divergence of the two sequences at silent sites is expected to scale directly with the number of generations separating the two genes. If constant generation time is assumed, the extent of divergence of the two sequences at silent sites is expected to scale with the time separating the two proteins. Thus, the neutral divergence of sequence involving synonymous sites should provide a 'molecular clock'.

In principle, the clock-like behaviour of synonymous substitutions can be used to normalize the number of nonsynonymous substitutions to assess the rate (changes per unit time) at which a protein sequence has evolved. This, in turn, has direct value to functional biologists. If the function of a protein is changing during an episode of natural history, one might expect its amino acid sequence to change rapidly to change the properties of the protein, as new properties are needed for fitness. Conversely, if the function of a protein is not changing, then the biophysical properties needed for fitness are the same at the end of the episode as well as the beginning, meaning that no amino acid replacements are needed during the episode. Thus, a historical episode where function in a protein has changed should be reflected by a high ratio of nonsynonymous to synonymous substitutions in the encoded gene.

Conversely, amino acid replacements should often alter the biophysical behaviour of the protein in a way that destroys fitness. To the extent that this is the case, descendants having mutated proteins would be less likely to survive, with the corresponding genetic substitution being removed from the population by purifying selection. Accordingly, a historical episode where function in a protein has *not* changed, should be reflected by a low ratio of nonsynonymous to synonymous changes in the encoded gene.

Historically, this analysis was introduced into the literature during a dispute between 'neutralists' (biologists who felt that the variation seen in natural protein families generally reflected drift, and could not be explained by functional variation) and 'selectionists' (biologists who felt that variation seen in natural protein families generally reflected adaptation of the various family members to different functional constraints within different hosts). The dispute was inappropriately formulated for an age where each individual gene could be viewed experimentally, making it neither necessary nor interesting to ask questions about the structural changes of proteins 'in general'. Accordingly, by 1999, the neutralist–selectionist dispute vanished (Hey, 1999).

Set within this dispute, however, it was useful to normalize the ratio of nonsynonymous to synonymous substitutions for the ratio of nonsynonymous substitutions to synonymous sites in a gene. This normalized ratio would be equal to unity if the gene sequence were suffering substitutions without any downstream selection. It would be less than unity if function were conserved, causing purifying selection to remove nonsynonymous replacements. It would be greater than unity for proteins undergoing adaptive changes in function, where 'positive' selection fixed nonsynonymous replacements that delivered an amino acid that changed the protein's behaviour.

Even a cursory examination of the standard genetic code showed that this normalized ratio (K_a/K_s or d_N/d_S) could not be obtained trivially. Yang and Nielsen (2000) reviewed 'approximate methods' for determining these ratios. For example, the ratio requires a count of the number of synonymous (S) and nonsynonymous (N) sites in the sequences. This count must aggregate silent sites of different degeneracies, including four-fold, three-fold, and two-fold degeneracies. In some cases, whether or not a site is synonymous depends on nucleotides at other sites. Given divergence, a site might be synonymous in one homologue and nonsynonymous in another.

To complete the calculation of the ratio, approximate methods then assessed the numbers of synonymous and nonsynonymous differences separating the two sequences. They then applied a 'correction' to account for the fact that more than one substitution might have occurred at the sites being counted.

The ratios were originally applied to compare extant sequences. However, Trabesinger-Ruef *et al.* (1996) and Messier and Stewart (1997) pointed out that these ratios could also be assigned to specific episodes in evolutionary history represented by specific branches on an evolutionary tree. Soon thereafter, the first database-wide compilations were made of proteins whose function was changing as inferred from high ratios (Liberles *et al.*, 2001).

10.3 Improving codon analysis beyond the K_a/K_s and d_N/d_S ratios

Even with their approximations and corrections, the model behind these ratios often did not capture certain features of DNA sequence evolution. For example, some approaches did not include unequal rates of transition substitutions, different transition and transversion substitution rates, and other features of the microscopic processes occurring at the level of specific DNA molecules. Further, the K_a/K_s and d_N/d_S ratios did not capture the possibility that natural selection might favour certain codons over others.

This motivated the improvement of codon models. In their first step, this improvement represented sequence divergence using 61 characters and specific parameters for the rates at which each of these interconvert to every other. This improvement had obvious advantages to evolutionary biologists.

Functional biologists, however, encountered different problems. Function can be altered in a protein by very few amino acid replacements, even as few as a single replacement (Benner and Ellington, 1988). As a typical catalytic protein contains 300 amino acids, the few replacements that change functional behaviour can easily be lost in a statistical analysis of the protein sequence as a whole. An episode of functional change can therefore be overlooked.

Further, while amino acid replacements at catalytic centres or binding sites might contribute positively to a changing function, amino acid replacements that disrupt the fold rarely do.

Accordingly, a wide range of sites will remain under purifying selection even in proteins whose functions are changing. These will include those at the core of the fold. These will cause the d_N/d_S ratios to be below those expected for neutral drift, even for proteins whose functions are changing.

Finally, replacements at some sites in a typical protein are almost certainly *not* constrained by purifying selection. These might be on the surface of the fold distant from a catalytic center, for example. These accept replacements that are not removed by natural selection, increasing the K_a/K_s and d_N/d_S ratios, even in a protein whose function is otherwise strongly conserved.

Operating together, these facts defeat naive application of K_a/K_s and d_N/d_S ratios as metrics for assessing functional change. Sequence divergence during an episode of functional change can have a ratio far less than unity, but only because the few nonsynonymous replacements that enable functional change are lost in the majority of conserved sites needed to maintain the fold of the protein. These problems cannot be solved by explicit codon models.

10.4 Heuristic approaches to improve codon analysis beyond the K_a/K_s and d_N/d_S ratios

Some functional biologists have attempted to circumvent these issues by evaluating K_a/K_s or d_N/d_S ratios first for the entire protein family, where the ratios are initially calculated for each branch of a family tree. These provide a view of the ratios found *typically* during the divergence of that family (Benner *et al.*, 1998; Yang, 1998). This approach assumes that function was typically stationary during at most of the episodes represented by branches in the tree. Then, branches where the K_a/K_s and/or d_N/d_S ratios are significantly higher than the average represent candidate episodes for functional change.

Other approaches consider the three-dimensional crystal structure of the protein. As early as 1989, it was recognized that plotting non-Markov behaviour on a model for the protein fold could add interpretive value when applied to specific sub-branches within an evolutionary tree (Benner, 1989), and this was subsequently applied as an interpretive tool for leptin (Gaucher *et al.*, 2003), G-protein coupled receptors (Soyer *et al.*, 2003), and ribonuclease (Sassi *et al.*, 2007). Liberles and his group introduced the concept of 'tertiary winnowing' based on a strategy that combines structural biology with sequence analysis (Berglund *et al.*, 2005). This is likely for theoretical reasons to be more useful than primary winnowing based on a sliding scale within a protein sequence (Fares *et al.*, 2002) or the branch-site model (Zhang *et al.*, 2005), although more empirical work will be necessary to develop all of these ideas.

This analysis reflects the fact that the evolutionary models, including codon models, used to describe divergent evolution need not be *stationary* over the course of divergent evolution. They can be different in some branches of the evolutionary than in others. This non-stationarity can carry functional information that is useful to functional biologists (Gaucher *et al.*, 2002).

For the bioinformatician, such non-stationarity creates a problem: Should we introduce multiple models for different branches of the tree? Or should we derive a model that captures as best as possible the average? And how should we limit a model to a reasonable number of parameters, so that we are modelling the signal without modeling the noise?

Such questions have been explored in single protein families. For example, bovine seminal ribonuclease (BS-RNase) constitutes 2% of the protein in bovine seminal plasma. It displays distinctive properties, including inhibition of immune cell proliferation, and a dimeric structure built from two identical polypeptide chains joined by two disulfide bonds (Sassi *et al.*, 2007). These features are all absent from digestive RNases, homologues made by the pancreas that diverged from seminal RNase about 40 million years ago. Resurrection of ancestral RNases shows that these distinctive properties were also absent from the last common ancestor of digestive and seminal RNases, which apparently was a digestive enzyme as well (Jermann *et al.* 1995).

This example of possible functional divergence following duplication encounters paradoxes if examined more closely. In most close relatives of ox, a gene for seminal RNase is present but is not obviously expressed. Further, the seminal RNase gene is frequently damaged and cannot possibly encode an active protein. This raises the possibility that the seminal RNase was not functional for a long period of time after it diverged from the digestive paralog 40 million years ago.

To ask whether the protein was a pseudogene throughout its recent history, one might ask whether the d_N/d_S ratio was unity during the episodes represented by internal branches within the tree (Figure 10.1). Unfortunately, only a small number of sites suffered nucleotide substitution in this time (only about 20 out of 124 amino acids are 'in play'). Therefore, it is difficult to justify calculating a separate d_N/d_S ratio for each branch of the tree. An initial analysis attempting to do so gave some nearly infinite ratios for some branches (due to division by nearly zero) and others ratios that were near zero.

This indicates overparameterization. An attempt to avoid overparameterization led Sassi *et al.* (2007) to build a simpler model with fewer free d_N/d_S parameters than the number of branches. One build grouped all branches with low $d_N/d_S(< 1)$ from the initial analysis into a single group having a single ratio. The branches with d_N/d_S higher than unity were allowed to have individual d_N/d_S values unless the branches were adjacent. The models were then designed to cover all possible combinations for calculating d_N/d_S covering from the most complex, with all the branch groupings, to the simplest, with one d_N/d_S for all branches.

The Akaike Information Criterion (AIC) (Posada and Buckley, 2004) was then used to identify the model that both fit the data best and had the optimal number of parameters. The AIC uses its Delta AIC function and Akaike weights to determine where within an ordered set of models with increasing complexity the increase in complexity is no longer balanced by an increase in the closeness to 'truth'.

This combination of statistical analysis with an abbreviated codon model gave an answer to the

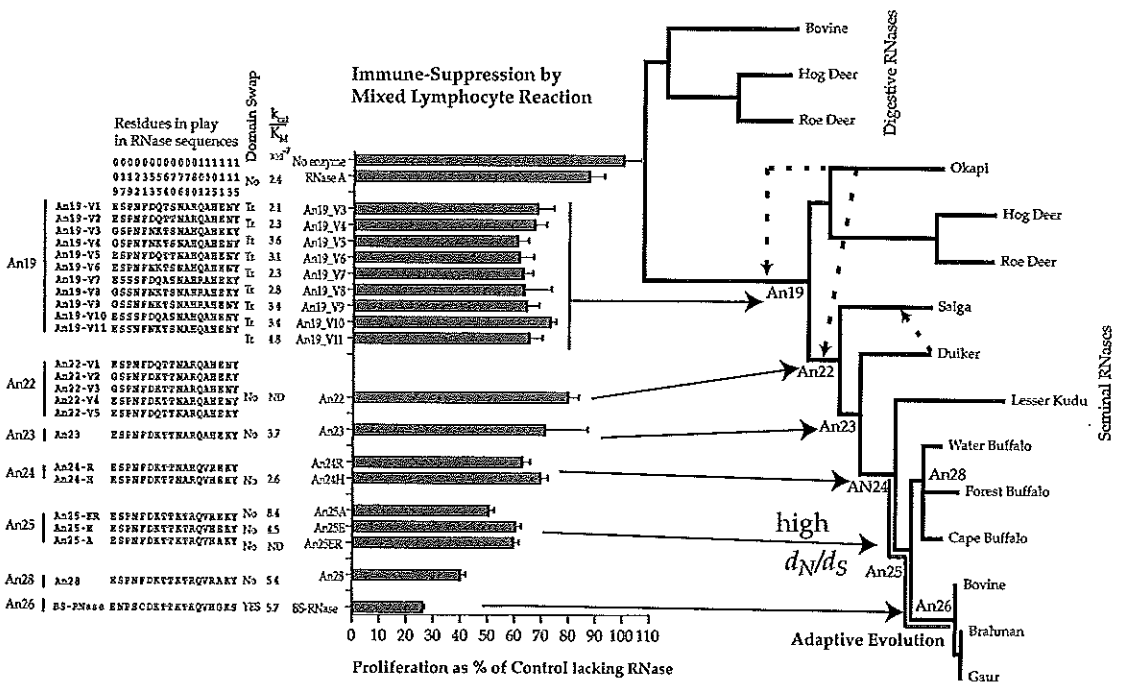


Figure 10.1 Phylogenetic tree interrelating seminal ribonuclease (RNase) sequences from various ruminants. The dashed lines represent ambiguities in the phylogenetic model. The branches given in double lines connecting ancestral proteins 24, 25, and 26 (An24, An25, An26) represent an episode of adaptive evolution, as indicated by a high K_a/K_s in the model preferred using the Akaike Information Criterion (Posada and Buckley, 2004). The bars represent immune cell proliferation in the presence of various ancestral proteins, with variable residues shown at left.

functional question. The best models indicated adaptive evolution ($1.6 < d_N/d_S < 6$, depending on the model) along the branch leading to the modern seminal RNase in the modern gaur, Brahman, and western ox, and adaptive evolution nowhere else ($d_N/d_S < 0.3$). Only the branches that extend from the ancestral node An24 to An26 (Figure 10.1) showed a high d_N/d_S indicating adaptive evolution. This answer was robust with respect to varying topologies, outgroups, and likelihood codon models.

A different approach to the same problem was approached by Yang *et al.* (2000). These authors recognized that individual sites might have especially high d_N/d_S values. This is expected, for example, in influenza proteins where a few sites interact with the 'prey' in a predator–prey co-adaptation variation. Such behaviour is, of course, exceptional, but analytical tools to detect it are valuable as complements to tools that identify specific branches that have special adaptive behaviour and couple these to structural biology and paleogenetic resurrections to analyse hypotheses of adaptive evolution (Benner, 2002).

These tools aside, these examples demonstrate a real world fact: real proteins having real lengths and real histories rarely support parameterization or use of a complete codon analysis. Nevertheless, abbreviated codon models and codon models that are coupled with non-sequence analyses can generate the types of conclusions that are interesting to functional biologists.

10.5 Clocks

Functional biologists also seek bioinformatics tools to help combine sequence analyses with non-sequence databases. One of these comes from the palaeontological and geological records, which provide entirely independent insights into the historical record that a bioinformatics model for protein families represents.

Many had hoped to use silent substitutions, such as K_S or d_S, as a molecular clock to correlate the genomic record of molecular history with the geological record. Unfortunately, each encoding nucleotide, A, G, T, and C, is a specific chemical molecule. Therefore, each kind of substitution (A for C, G for A, T for C, and so on) does not occur in the same way.

In particular, the various kinds of substitution reflect different combinations of 'microscopic' processes. Substitutions may arise through chemical transformation of the nucleotides themselves, such as the deamination of cytidine to give uridine, effectively converting a C to a T. Substitutions may arise via mismatching during copying. For a substitution to appear in a database, the substitutions must survive repair, which itself can occur by various mechanisms.

Thus, the rates for each of the 12 types of substitutions are weighted aggregates of the microscopic process that create them. These rates depend on the environment, but also on features of the genetic system that are themselves under selective pressure. In particular, the accuracy of polymerases and the efficiency of repair are all determined by the sequences of polymerases and repair enzymes. Further, not all codon substitutions need to be neutral with respect to fitness.

These facts all create further problems for those attempting to use codon models to describe the divergent evolution of encoding gene sequences. A 61-character model can easily capture modestly different fitnesses of different codons for the same amino acids for a species, for example. It requires additional parameters should the bias change over time, and still more if mutation rates change over time. Nevertheless, even with genome-sized databases, the amount of data is small relative to the number of possible free parameters in the model.

Accordingly, various groups have attempted to reduce the number of free parameters by aggregating phenomena where aggregation makes chemical and biological sense. For example, we can aggregate two-fold redundant codon systems into two groups; one where synonymous codons are interconverted by purine–purine transitions (in the standard genetic code, these are codon systems for Glu, Gln, and Lys), and the other where synonymous codons are interconverted by pyrimidine–pyrimidine transitions (in the standard genetic code, Cys, Asp, Asn, Tyr, Phe, and His). Generally translated by the same tRNA (using wobble), codon usage is not likely to be biased strongly by tRNA abundances. The rates of transition are not likely to be influenced by what codon holds the transitioning nucleotide.

Further, two-fold redundant sites within codons whose encoded amino acid is conserved follow 'approach to equilibrium' kinetics. Consider the codon system for a conserved Gln (CAA or CAG), where the nucleotide at the third position alternates between A and G according to the scheme:

$$\mathrm{A} \underset{k_{\mathrm{G}\rightarrow\mathrm{A}}}{\overset{k_{\mathrm{A}\rightarrow\mathrm{G}}}{\rightleftarrows}} \mathrm{G}. \tag{10.1}$$

As time increases, the third site equilibrates between A and G, where the rate constant k_{R} for the equilibration reaction is equal to the sum of the forward rate constant *and* the reverse rate constant, that is, $k_{\mathrm{R}} = k_{\mathrm{A}\rightarrow\mathrm{G}} + k_{\mathrm{G}\rightarrow\mathrm{A}}$ (Atkins and de Paula, 2002). At equilibrium, the ratio of $[\mathrm{G}]_{\mathrm{eq}}$ to $[\mathrm{A}]_{\mathrm{eq}}$, where $[\mathrm{G}]_{\mathrm{eq}}$ and $[\mathrm{A}]_{\mathrm{eq}}$ are the respective fractions of G and A at equilibrium, is equal to the ratio of the forward and reverse rate constants, that is, $[\mathrm{G}]_{\mathrm{eq}}/[\mathrm{A}]_{\mathrm{eq}} = (k_{\mathrm{A}\rightarrow\mathrm{G}})/(k_{\mathrm{G}\rightarrow\mathrm{A}})$. Thus, if the fraction of A (f_{A}) at $t = 0$ is unity ($\mathrm{A}_0 = 1$), then the fraction of A remaining after time t, expressed as $f_{\mathrm{A}} = [\mathrm{A}(t)]/\mathrm{A}_0$, is given by the equation:

$$\frac{[A(t)]}{[A_0]} = f_{Geq}e^{-(k_{A\rightarrow G}+k_{G\rightarrow A})t} + f_{Aeq}, \tag{10.2}$$

where f_{Geq} and f_{Aeq} are the fractions of G and A at equilibrium (that is $f_{\mathrm{Geq}} = [\mathrm{G}]_{\mathrm{eq}}/([\mathrm{G}]_{\mathrm{eq}} + [\mathrm{A}]_{\mathrm{eq}})$ and $f_{\mathrm{Aeq}} = [\mathrm{A}]_{\mathrm{eq}}/([\mathrm{G}]_{\mathrm{eq}} + [\mathrm{A}]_{\mathrm{eq}})$).

This gives the fraction of sites that initially held A and still hold A after a time t. It does so without any need for 'correction' to reflect the fact that as the system approaches equilibrium, any particular molecule can undergo an indefinite number of interconversions, back and forth, between the two states (Yoder and Yang, 2000).

To apply two-fold redundant exchange as a clock, we first stipulate that substitution at each site in a gene is independent of substitutions at other sites, the rate constants for substitutions are the same at all sites, and that the silent sites are at equilibrium. The last hypothesis is an approximation, but a good one as long as the rate constants are large compared to the rate of change of the rate constants.

We now consider two identical sequences, where one is given the opportunity to diverge. How will the fraction identity at sites constrained to hold purines diverge in the evolving sequences? Consider separately the sites that are occupied by A at $t = 0$ and the sites that are occupied by G at $t = 0$. For those that are originally occupied by A, the sites conserved after time t are those that have A after time t.

The conserved sites arising from A is given by:

$$(f_{\mathrm{Geq}}e^{-k_R t} + f_{Aeq})f_{\mathrm{Aeq}}, \tag{10.3}$$

where the f_{Aeq} term outside of the parentheses represents the fraction of the starting sites that are occupied by A, while the term within parentheses describes the fraction of these that remain A after time t.

The equation describing the number of conserved sites arising from G as a function of time is similarly derived:

$$(f_{\mathrm{Aeq}}e^{-k_R t} + f_{Geq})f_{\mathrm{Geq}}. \tag{10.4}$$

The fraction of all sites having the same purine after time t as they had at time zero, $f_{2\mathrm{R}}$ is the sum of these two equations:

$$f_{2\mathrm{R}} = f_{Aeq}f_{\mathrm{Geq}}e^{-k_R t} + f_{Aeq}f_{\mathrm{Aeq}} + f_{Aeq}f_{\mathrm{Geq}}e^{-k_R t} + f_{Geq}f_{\mathrm{Geq}}. \tag{10.5}$$

Since [G] + [A] is always equal to unity, we have:

$$([G] + [A])^2 = 1 \tag{10.6}$$

and:

$$[G]^2 + 2[G][A] + [A]^2 = 1. \tag{10.7}$$

For all [G] and [A], including $[\mathrm{G}]_{\mathrm{eq}}$ and $[\mathrm{A}]_{\mathrm{eq}}$, let:

$$E_R = [G_{\mathrm{eq}}]^2 + [A_{\mathrm{eq}}]^2 \tag{10.8}$$

$$P_R = 2\mathrm{f}_{Aeq}f_{Geq}, \tag{10.9}$$

therefore,

$$P_{\mathrm{R}} + E_{\mathrm{R}} = 1. \tag{10.10}$$

Equation 10.5 can therefore be rewritten as:

$$f_{2\mathrm{R}} = P_R e^{-k_R t} + E_R. \tag{10.11}$$

Thus, the fraction of conserved purine nucleotides at two-fold redundant sites follows an exponential first order approach to equilibrium towards an equilibrium end point, E_{R}, the equilibrium fractions occupied by A and G. Again, this equation correctly handles the possibility of multiple substitutions at a single site; indeed, this is why the equilibrium is approached.

Solving Eqn 10.11 gives a distance based on transition redundant exchange (TREx) kinetics:

$$k_R t = -\ln[(f_{2R} - E_R)/P_R] = \text{TREx distance}, \tag{10.12}$$

where P_R is the pre-exponential term $(= 2 f_{Aeq} f_{Geq})$ and E_R is the f_2 reached at equilibrium $(= {f_{Aeq}}^2 + {f_{Geq}}^2)$.

A value for $k_R t$ can therefore be determined from an f_{2R} value using Eqn 10.12.

In this model, f_{2R} as a function of time follows a first-order exponential decay from unity to an end point defined by the expression $({f_{Aeq}}^2 + {f_{Geq}}^2)$. If A and G appear with equal frequency, then the equilibrium end point $E_R = 0.5$. If, however, A and G appear with frequencies of (for example) 0.6 and 0.4, then the end point E_R is 0.52 $(= 0.6^2 + 0.4^2)$.

If the rate constants are assumed to be time-invariant, f_{2R} can be used as a molecular clock. To implement this clock, we identify sites in a pair of aligned DNA sequences that are constrained to mutate between A and G only. The third positions of codons for three amino acids (Glu, Gln, and Lys) are so constrained if the amino acid has not been replaced in the interval separating the two genes. In practice, as nonsynonymous substitutions are generally more infrequent than synonymous substitutions, we can ignore the possibility that two compensatory nonsynonymous substitutions have led to overall amino acid conservation. From a pair of aligned gene sequences, we count the number of Glu, Gln, and Lys codons that are conserved in the two encoded proteins, the number of those codons having the same nucleotide at their third positions, and calculate f_{2R} by dividing the second counted number by the first.

An analogous kinetic expression can be written for pyrimidine–pyrimidine transitions. The third positions of six amino acids (Cys, Asp, Phe, His, Asn, and Tyr) are constrained to have only T or C. In a pair of aligned gene sequences, the number of conserved Cys, Asp, Phe, His, Asn, and Tyr is counted, the number of those codons that have the same nucleotide at their third position is counted, and f_{2Y} (Y for pYrimidines) for the pair of genes is obtained by dividing the second counted number by the first. TREx distances can be calculated using a formula analogous to Eqn 10.12.

10.6 Calibrating the TREx clock

Li *et al.* (2006) calibrated the transition redundant exchange (TREx) clock for various vertebrates, recognizing that the accuracy of a clock is highest when dating the divergence of genes separated by a time similar to the half-life associated with the transition rate constant, $t_{1/2} = \ln 2/k$. For events occurring near the time of the divergence of the major mammalian orders *c.* 80 million years ago (Ma), for example, the optimal rate constant would be *c.*4.4×10^{-9} transitions/site/year, recognizing that 160 million years in total time separates two contemporary taxa that diverged 80 Ma (we have here doubled the time to reflect a double lineage process).

To calibrate the TREx clock, Li *et al.* (2006) began by recognizing that after two taxa arise by speciation, each gene in one taxon has a corresponding ortholog in the other. For gene i, the two genomes generate the $i_T : i_U$ pair. Subsequently, individual genes may be lost in separate lineages, removing $i_T : i_U$ pairs.

Absent lateral transfer of genes between species, orthologous proteins in an inter-taxa comparison can have diverged no more recently than the date when the two lineages themselves diverged. Therefore, no clock should date any inter-taxon pair as having diverged after the two taxa diverged; the f_{2Y} and f_{2R} values should be the same for all true orthologs, and characteristic of the date of species divergence.

It is possible, however, for an inter-taxon pair to have diverged *before* the two taxa diverged (and be so dated). This will be the case, for example, if the last common ancestor of the two taxa already contained two paralogous genes arising from gene duplication prior to the date of divergence (see Figure 10.2). These are called 'outparalogs'. Here, the f_{2Y} and f_{2R} values can be smaller; if the initial paralogization occurred a long time ago, these values will be at or near the equilibrium value.

This pattern is in fact seen with real data. For example, the f_{2Y} and f_{2R} values for rat–mouse orthologs form a cluster ($f_{2Y} = 0.88$ and

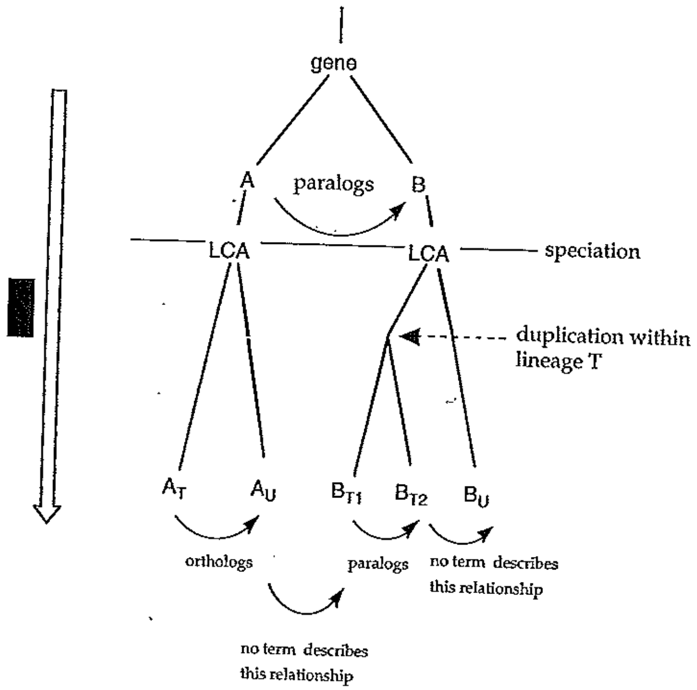

gene
A
paralogs
B
LCA
LCA
speciation
duplication within lineage T
A_T
A_U
B_{T1}
B_{T2}
B_U
orthologs
paralogs
no term describes this relationship
no term describes this relationship

Figure 10.2 While two genes in two organisms descendent from a single gene in the last common ancestor are clearly named 'orthologs', and two genes in one organism descendent from a single gene in an ancestor of that organism are clearly named 'paralogs', neither term captures the complexity of pairwise relationships between genes in two organisms that suffered duplication ('paralogization') following the divergence of those organisms in their last common ancestor (e.g. A_U versus B_{T1}). Nor does either capture the complexity of pairwise relationships when, in one taxon, one of two paralogs has suffered duplication (e.g. B_U versus B_{T1}). Various terms have been suggested (e.g. 'outparalogs'), but in any case, depending on the amount of divergence overall, extremely reliable tools are often needed to sort out this complexity.

f_{2R} = 0.90). The values for outparalogs were lower (Figure 10.3). Indeed, a substantial number of pairs of outparalogs have f_{2Y} or f_{2R} values ≈ 0.59, not far from the values of 0.52–0.54 that are expected for silent sites that have equilibrated.

The f_{2Y} and f_{2R} values of 0.88 and 0.90, with an end point of 0.51 and 0.54, correspond to TREx distances of 0.281 and 0.245. As rat and mouse diverged 16 million years ago (Li, 1977) (32 million years separate rat and mouse, if one wishes to calculate a single lineage rate constant), the pyrimidine–pyrimidine and purine–purine observed transition rate constants are estimated to be $k_{obsY} = 8.8 \times 10^{-9}$ changes/site/year and $k_{obsR} = 7.7 \times 10^{-9}$ changes/site/year.

This analysis assumes that the codon bias is time-invariant within this subset of rodents. To assess the plausibility of this assumption, the codon bias of rat and mouse was considered (Li *et al.*, 2006). The fraction of A at the two-fold redundant sites for Glu, Gln, and Lys (f_{eqA}) is 0.37 and 0.36 in rat and mouse, respectively; the f_{eqT} is 0.45 and 0.43 in rat and mouse, respectively, at the two-fold redundant sites involving Cys, Asp, Phe, His, Asn, and Tyr. From these biases, we calculate expected equilibrium end-points for f_{2R} of 0.53 and 0.54 for rat and mouse, respectively, and end points for f_{2Y} of 0.52 and 0.51 for rat and mouse, respectively. The similar codon bias at two-fold redundant sites for mouse and rat suggests that the assumption that codon bias was invariant in the time separating the two taxa is serviceable.

The f_{2Y} and f_{2R} values for true orthologous pairs are expected to be binomially distributed. This distribution can be approximated using a Gaussian. To the extent that the assumptions within the model are incorrect, the distribution should be overdispersed. We may assume, as a null hypothesis, that

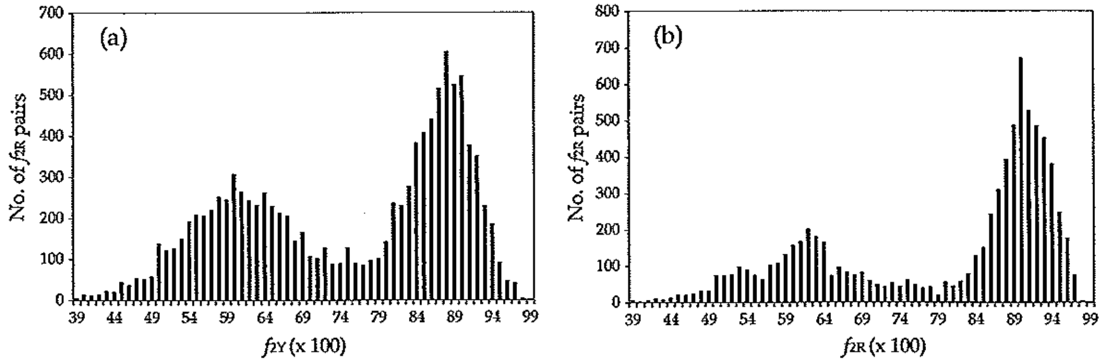


Figure 10.3 Histogram showing the f_{2Y} (a) and f_{2R} (b) values of all mouse:rat inter-taxa homologue pairs containing 50 or more characters. The peak centered at c.0.88 (a) and c.0.90 (b) reflect true orthologs. Pairs with f_2 values near 0.53 diverged so long ago that the silent sites have equilibrated.

all of the genes represented in the inter-taxon pairs have diverged with the same rate constants. To test for overdispersion, Li *et al.* (2006) extracted in f_{2Y} and f_{2R} centred around 0.88 and 0.90 to obtain a set of putative orthologs. The histograms of n for f_{2Y} and f_{2R} were fit to Poisson distributions. In both cases, only modest (but significant) overdispersion was observed. This defined a limit to the assumption that the rate constant for transitions is the same at all sites in all genes.

The relatively similar f_{2Y} and f_{2R} values for the mouse:rat inter-taxon pairs, and the relatively small (14%) difference in the estimated pyrimidine–pyrimidine and purine–purine transition exchange rate constants suggested that f_{2Y} and f_{2R} might be combined to give an f_2 metric without creating an undesirably large variance. As shown in Figure 10.4, the greater number of characters used to calculate f_2 gave a sharper distribution, balanced by a slightly larger overdispersion expected for the different pyrimidine–pyrimidine and purine–purine transition rate constants. The ratio (R_{mv}) between σ and μ of f_{2R}, f_{2Y} and f_2 was compared (Table 10.1). The R_{mv} value of f_2 is smaller than those of f_{2R} and f_{2Y}, demonstrating that f_2 metric is as good as that of f_{2R} and f_{2Y}.

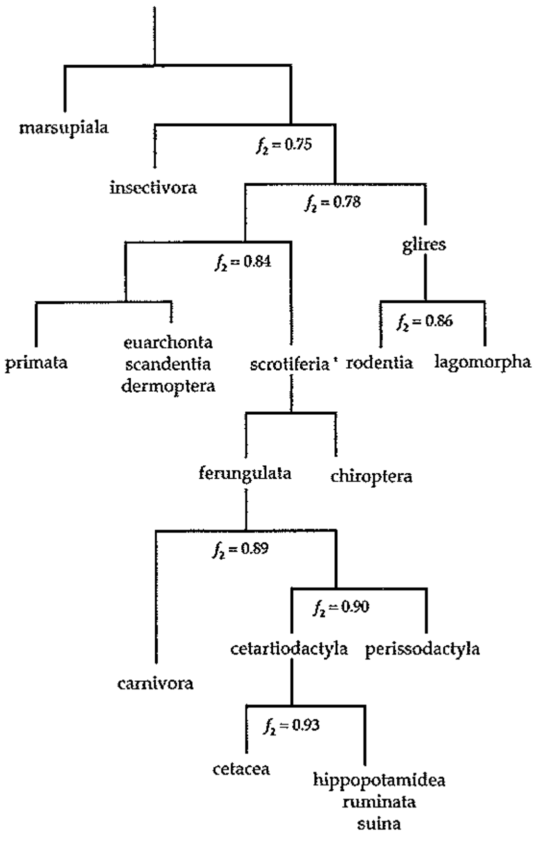


Figure 10.4 Values for f_2 for nodes in the divergence of various orders of mammals.

The TREx clock has been applied throughout vertebrate evolution. Interestingly, it has not fully equilibrated in the time separating birds and mammals. Further, it has been used to detect gene duplications that created the new metabolic pathway in yeast that allowed the fermentation of glucose to create ethanol (Thomson *et al.*, 2005). This occurred on the planet at the time that fermentable fruits arose in the Cretaceous. Thus, it serves its role of correlating the historical record caught within

Table 10.1 Comparison of f_{2R}, f_{2Y}, f_2 and f_4

	μ	σ	$R_{m/}$
f2r	0.9	0.034	0.0378
f2y	0.88	0.04	0.0455
f2	0.89	0.029	0.0326
f4	0.84	0.05	0.0595

the genomic record with that caught in Earth's rocks.

10.7 Conclusions

Functional and evolutionary biologists applying codon models to specific questions provide guidance to bioinformaticians seeking to build increasingly sophisticated models to describe the divergence of gene and protein sequences. First, despite the value of analyses based on protein sequences alone, it is a shame to discard the information in the encoding DNA sequences. At the same time, naive analyses of DNA sequences of coding regions without any codon model at all are largely useless.

This drives the search for codon models that can capture the needed information in a biologically sensible way. Unfortunately, a full 61-character codon model is difficult to parameterize in anything but the largest databases. Further, the details of real sequence evolution, including its non-stationary features, drives the need for even larger parameter sets for single protein families that have still fewer characters upon which to ground parameterization. Accordingly, biologists are seeking expedients that aggregate codons, where the aggregation 'makes sense' in terms of chemistry, enzymology, or biology.

Several of these aggregative models have been benchmarked and are being used. Coupled with statistical tools like the Akiaike Information Criterion and detailed analysis of overdispersion, they achieve the compromise between completeness and parameterizability needed for utility.

Acknowledgement

Parts of this work were supported by the National Institute of General Medical Science and the National Institute of Alcohol Abuse and Alcoholism.

References

Atkins, P. and de Paula, J. (2002). *Elements of physical chemistry with applications in biology*. New York, Freeman.

Benner, S.A. (1989). Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enz. Regul.* 28, 219–236.

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. (2003). Interpretive proteomics. Finding biological meaning in genome and proteome databases. *Adv. Enzyme Regul.* 43, 271–359.

Benner, S.A. (2007). The early days of paleogenetics: connecting molecules to the planet. In: D.A. Liberles (editor): *Ancestral sequence reconstruction*, Oxford University Press, Oxford, UK, pp. 3–19.

Benner, S.A. and Ellington, A.D. (1988). Interpreting the behavior of enzymes. Purpose or pedigree? *CRC Crit. Rev. Biochem.* 23, 369–426.

Benner, S.A., Cannarozzi, G., Chelvanayagam, G., and Turcotte, M. (1997). *Bona fide* predictions of protein secondary structure using transparent analyses of multiple sequence alignments. *Chem. Rev.* 97, 2725–2843.

Benner, S.A., Trabesinger-Ruef, N., and Schreiber, D.R. (1998). Post-genomic science. Converting primary structure into physiological function. *Adv. Enzyme Regul.* 38, 155–180.

Berglund, A.-C., Wallner, B., Elofsson, A., and Liberles, D.A. (2005). Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol.* 60, 499–504.

Brenner, S. (1988). The molecular evolution of genes and proteins. A tale of two serines. *Nature* 334, 528–530.

Fares, M.A., Elena, S.F., Ortiz, J., Moya, A., and Barrio, E. (2002). A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.* 55, 509–521.

Gaucher, E.A., Gu, X., Miyamoto, M.M., and Benner, S.A. (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem. Sci.* 27, 315–321.

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

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

Hey, J. (1999). The neutralist, the fly and the selectionist. *Trends Ecol. Evol.* 14, 35–38.

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.

Li, W.H. (1977). Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* 85, 331–337.

Li, T., Chamberlin, S.G., Caraco, M.D., Liberles, D.A., Gaucher, E.A., and Benner, S.A. (2006). Analysis of transitions at two-fold redundant sites in mammalian genomes. Transition redundant approach-to-equilibrium (TREx) distance metrics. BMC *Evol. Biol.* 6, 25.241.

Liberles, D.A., Schreiber, D.R., Govindarajan, S., Chamberlin, S.G., and Benner, S.A. (2001). The adaptive evolution database (TAED). *Genome Biol.* 2, 0003.1–0003.18.

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

Posada, D. and Buckley, T.R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol* 53, 793–808.

Sassi, S.O., Braun, E.L., and Benner, S.A. (2007). The evolution of seminal ribonuclease: Pseudogene reactivation or multiple gene inactivation events? *Mol. Biol. Evol.* 24, 1012–1024.

Soyer, O.S., Matthew, W., Dimmic, M.W., Richard, R., Neubig, R.R., Richard, A., *et al.* (2003). Dimerization in aminergic G-protein-coupled receptors: Application of a hidden-site class model of evolution. *Biochemistry* 42, 14522–14531.

Thomson, J.M., Gaucher, E.A., Burgan, M.F., Aris, J.P., and Benner, S.A. (2005). Resurrecting extinct proteins from ancient yeast at the origin of fermentation. *Nature Genetics* 37, 630–635.

Trabesinger-Ruef, N., Jermann, T.M., Zankel, T.R., Durrant, B., Frank, G., and Benner, S.A. (1996). Pseudogenes in ribonuclease evolution. A source of new biomacromolecular function? *FEBS Lett.* 382, 319–322.

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

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43.

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

Yoder, A.D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* 17, 1081–1090.

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.