Predicting functional divergence in protein evolution by site-specific rate shifts

Eric A. Gaucher, Xun Gu, Michael M. Miyamoto and Steven A. Benner

Most modern tools that analyze protein evolution allow individual sites to mutate at constant rates over the history of the protein family. However, Walter Fitch observed in the 1970s that, if a protein changes its function, the mutability of individual sites might also change. This observation is captured in the 'non-homogeneous gamma model', which extracts functional information from gene families by examining the different rates at which individual sites evolve. This model has recently been coupled with structural and molecular biology to identify sites that are likely to be involved in changing function within the gene family. Applying this to multiple gene families highlights the widespread divergence of functional behavior among proteins to generate paralogs and orthologs.

Eric A. Gaucher
Steven A. Benner*
NASA Astrobiology Institute and Dept Chemistry, University of Florida, Gainesville, FL 32611, USA.
*e-mail: benner@chem.ufl.edu

Xun Gu
Dept Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA.

Michael M. Miyamoto
Dept of Zoology, University of Florida, Gainesville, FL 32611, USA.

For 40 years, molecular biology has used evolutionary approaches as a way to connect macromolecular (DNA or amino acid) sequence to function [1]. Early work applied special mathematical analyses to specific cases [2,3]. Outside the field of molecular evolution, simple evolutionary approaches (e.g. homology searches [4]) are widely used today to infer 'function', on the assumption that homologous sequences have similar function [5]. Efforts have recently been devoted to the development of more sophisticated mathematical treatments that maximize the interpretive value of genomic sequences [5–10].

The recent completion of a draft sequence for the human genome [11,12] is almost certainly only the beginning of a new episode of post-genomic biological research, when genomic databases will be used to generate hypotheses that can be experimentally tested (interpretive genomics). For this purpose, computational methods that generate hypotheses about function from sequence evolution will be valuable. Patterns of replacement, including changes in the rate of replacement, are likely to be important to these methods. For example, the functional importance of sites is generally believed to be inversely related to the evolutionary rate of amino acid replacement [13,14]. This belief arises from one interpretation of the neutral theory [13], in which sites of greatest functional significance are under the strongest selective constraints. An organism that experiences a replacement at one of these sites is less likely to survive and reproduce. Thus, the observation that a histidine is highly conserved during the evolution of a protein family is frequently taken as an indication that the residue is in the active site of the enzyme and is directly involved in catalytic function.

In some cases, the extent to which function constrains the evolution of a protein sequence can be estimated by measuring the ratio of non-synonymous to synonymous substitution during its evolution. This ratio is also widely used to detect positive selection in coding DNA [15]. However, synonymous substitutions are often selectively neutral and therefore occur at a rapid rate [13]. Hence, synonymous substitution can be used to detect only recent functional divergence, because these sites rapidly become saturated with mutations. For a typical vertebrate nuclear-encoded gene [14], this type of analysis has been useful only as far back as ~150 million years.

To assess more broadly the functional significance of sequence evolution, particularly among more distantly related proteins, new approaches have emerged that consider amino acid replacements (or, equivalently, non-synonymous substitution) alone. These begin by analyzing how the evolutionary rates of amino acid replacement differ among sites in a protein sequence [16–18] (site-to-site rate heterogeneity; Box 1), with a statistical formalism in which the rate varies among sites according to a gamma distribution. In the conventional analysis of sequence evolution using the gamma model [17], rapidly and slowly evolving sites remain rapid or slow across the entire evolutionary tree. Because of this, the model is termed 'homogeneous'. A homogeneous evolutionary rate is expected when the functional constraints at sites are constant over the entire evolutionary history.

However, if the function of the protein is changing, some residues might be subject to altered functional constraints in various portions of the phylogenetic tree. This, in turn, implies that the evolutionary rates at these sites will be different in different branches of the tree. Such change in function is called type-I evolutionary functional divergence [6,19,20]. To capture this evolutionary phenomenon, the constraint of fixed rates per site along the phylogeny must be relaxed to allow the identities of 'fast' and 'slow' sites to change over evolutionary time; that is, to allow site-specific rate shifts. Here, we refer to this process as the 'non-homogeneous gamma model'. Conceptually, it can be tied to the covarion process; that is, lineage-specific patterns of rate variation [2]. During the past three decades, many studies have

http://tibs.trends.com 0968-0004/02/$ – see front matter PII: S0968-0004(02)02094-7

Box 1. Homogeneous and non-homogeneous gamma models

Four amino acid positions of two hypothetical descendent proteins and their common ancestor are shown in Fig. I. Each 'dot' represents one unit of evolutionary change in relation to the three other dots for a given sequence. Thus, each dot represents the relative rate and not the absolute rate. The descendent and ancestral sequence profiles include the same number of slowly, moderately and rapidly evolving positions (i.e. each consists of one 'slow', one 'moderate' and two 'fast' sites). The difference is that the identity of these slow, moderate and fast sites can change in the non-homogeneous gamma process, in contrast to their fixed status in the homogeneous alternative. Thus, a slow site can become a fast one (or vice versa) in the non-homogeneous model but not in the homogeneous one. Such rate changes at sites might reflect shifts in their selective constraints and can thereby identify positions that are probably involved in functional divergence.

Rate variation among sites is summarized in the gamma distribution by its shape parameter (α). When the mean replacement rate per site is set to 1.00, α becomes equal to the inverse of the variance in rates among sites. Thus, α increases as the variation in rates among sites decreases. As has been proved statistically, combining two sets of sequences with non-homogeneous rates increases the value of α relative to its separate estimates for each group. Conceptually, this increase in α can be explained by reference to the two diagrams of descendent and ancestral sequences. In the case of homogeneous rates, each position of the two descendent sequences retains its ancestral identity of slow, moderate or fast. Thus, the same set of rate identities underlies the estimation of α, regardless of whether the descendents are analyzed separately or together. By contrast, when rates can shift, the slow sites of one group can be opposed by the fast ones of the other (and vice versa).

Consider the situation in which the sequences from D3 are combined with the sequences from D4 to generate a single multiple sequence alignment (Fig. I). The combined D3–D4 data are then used in a phylogenetic analysis to calculate the evolutionary rates for the four positions. The estimated evolutionary rates for D3–D4 will not be similar to the evolutionary rates when D3 and D4 are considered separately. Such differences introduce a dampening effect when the two groups are combined, because the rate differences among sites are reduced. This apparent reduction in the rate differences among sites results in a corresponding increase in the estimation of α. For real sequences, such increases in α, when the data are combined, can offer strong evidence that the rate variation among sites is not homogeneous.

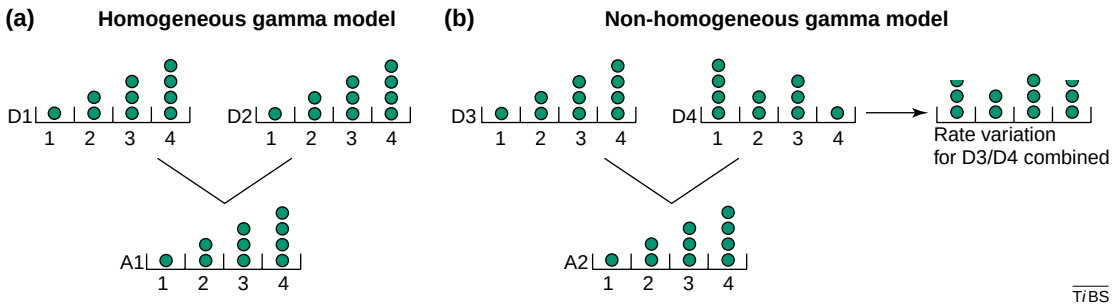

Fig. I. Differences in evolutionary-rate behavior for the homogeneous (a) and non-homogeneous (b) gamma models. Each sequence contains four positions (indicated by sequential numbers 1–4). A1, A2 and D1–D4 represent ancestral and descendent populations, respectively. By considering a population of sequences (or subtree), one can calculate the evolutionary rate at each position of the sequence. For the homogeneous gamma model, A1 gives rise to two descendent populations (D1 and D2) each of which has the same site-specific rates as the ancestral population. By contrast, for the non-homogeneous gamma model, A2 gives rise to two descendent populations (D3 and D4) in which the site-specific rates can be different from those in the ancestral population, and possibly different from each other. Notice that each descendent population D3 and D4 contains the same number of slow, moderate and fast sites. Thus, the rate variation among sites (α) is the same for each of these individual descendent populations.

found patterns of site-specific rate shifts during protein evolution [2,21–36]. For example, statistical tests have been developed and used to show that the non-homogeneous gamma model provides a better explanation of the evolutionary process than does the homogeneous gamma model [24,25,30,33,36].

Site-specific rate shifts can influence tree making, as has been demonstrated by empirical analysis [27,29] and statistical modeling [26]. A non-homogeneous gamma model is also useful in identifying sites that could be involved in the change of protein function [6,19]. Rate-shifted sites are the residues that have either enhanced or reduced selective constraints as a consequence of the change in function during evolution, according to the hypothesis of type-I evolutionary functional divergence. For functional genomics, this analysis turns out to be valuable. Indeed, these sites can be further evaluated for their roles in functional divergence by mapping them onto the available tertiary structures of their proteins [20,31,34,37–41], and be targeted in experiments. Such studies have either predicted or correlated functional divergence among gene families as diverse as caspases [20], elongation factors [31,34], globins [40], Janus kinases [35] and class-I α-mannosidases [42].

We begin here with a summary of different models for the study of type-I evolutionary functional divergence of proteins by site-specific rate shifts during evolution. We then focus on approaches that use the non-homogeneous gamma model (Box 1) to identify sites that might be significant for functional divergence and apply these approaches to several gene families, highlighting the widespread occurrence of functional divergence during protein evolution. We finally illustrate the value of this

Table 1. Coefficients of evolutionary functional divergence (θ) between homologous clusters of ten gene families, estimated by DIVERGE

Gene family	Member gene cluster[a]		Sites[b]	$\theta_{AB} \pm SE$[c]	Refs
	A	B			
Y-box-binding protein	YB-1a (13)	Variant A (5)	179	0.31 ± 0.16	[35]
CC chemokine receptor 2/5	CCR5 (15)	CCR2 (4)	341	0.42 ± 0.12	[35]
Sarcoplasmic reticulum Ca^{2+} ATPase	SERCA1 (13)	SERCA2 (5)	990	0.49 ± 0.09	[35]
Endothelin	END1 (6)	END2/3 (6)	130	0.56 ± 0.17	[35]
Calponin	H1 (11)	H2 (5)	182	0.61 ± 0.15	[35]
Hemoglobin	α (56)	β (56)	137	0.36 ± 0.07	[40]
Transferrin	TF (7)	LTF (5)	553	0.19 ± 0.07	[6]
Myc	c-Myc (14)	N-Myc (8)	276	0.39 ± 0.08	[6]
COX	Cox-1 (8)	Cox-2 (11)	583	0.44 ± 0.09	[19]
Caspase	CED-3 (31)	ICE (11)	198	0.29 ± 0.09	[20]

[a]Numbers in parentheses represent the number of sequences for each member gene cluster.
[b]The total number of amino acid positions (without gaps) in the multiple sequence alignment.
[c]The coefficient of evolutionary functional divergence between clusters A and B, and its standard error (SE).

approach with specific cases that combine knowledge about shifting rates with information from structural biology. Such combination is a powerful and cost-effective method for the further determination of protein function.

Evolutionary models for functional divergence

A new approach has been developed that models the site-specific rate shift under the non-homogeneous framework in such a way that the homogeneous gamma model can be treated as a special case. This provides an opportunity to determine which of the two models provides a better statistical fit to the data (Box 2). Gu developed a two-state model to this end [6]. Consider a phylogeny with at least two monophyletic clusters generated by gene duplication or speciation. It is proposed that an amino acid site has two states. In one state (S_0), the site has the same mutation rate in both clusters; in the other state (S_1), the site has different rates. In each state, the evolution rate varies among sites according to the gamma model. The prediction of functional divergence (θ) between two clusters is defined as the probability of a site being in S_1 [i.e. $\theta = P(S_1)$], which is called the coefficient of evolutionary functional divergence. According to this approach, the homogeneous gamma model is a special case in which $\theta = 0$. A fast algorithm (Poisson based with corrections) was initially developed for estimating θ values [6]. To incorporate both a more rigorous statistical analysis and more complex evolutionary models, a likelihood approach was recently developed; this was also extended to analyze three or more gene clusters simultaneously [19].

Conceptually, θ measures the degree of independence (i.e. the lack of correlation) between the relative evolutionary rates at the sites in one protein subfamily and those in another. This coefficient ranges from 0 to 1, with an estimate of 0 indicating that the same relative rates apply to the sites of one group and to those of the second. In turn, values approaching 1 reflect increasing differences between the relative rates of each site in the two subfamilies. Thus, values of θ that are significantly greater than 0 document the occurrence of rate shifts at specific sites and the insufficiency of the homogeneous gamma model. Furthermore, given the premise that shifting rates reflect changes in protein function, θ can also be interpreted as a predictor of the overall degree of functional divergence between protein subfamilies.

Type-I evolutionary functional divergence (i.e. change in function results in site-specific rate shifts) after gene duplication provides a biological basis for the non-homogeneous behavior of evolutionary rates [6,19,20]. This premise has been tested by analyzing ten vertebrate gene families (Table 1). The θ coefficients between two member gene clusters of a family range from 0.19 to 0.61, and all of these are significantly greater than 0. This implies that most duplicate genes have undergone shifted functional constraints after gene duplication. Moreover, a similar pattern of functional divergence after domain duplication (shuffling) is also observed, indicating the importance of non-homogeneous evolution in generating multiple-domain proteins. For example, Gu *et al.* [35] have studied the JAK protein family, a set of non-receptor tyrosine kinases. These tyrosine kinases have two homologous domains: a tandem kinase domain (JH1, functional) and a pseudokinase domain (JH2, function unknown). The θ coefficient between the two domains is 0.412 ± 0.049 (Fig. 1).

Under the gamma model, the rate heterogeneity among sites is characterized by its shape parameter (α). When there are rate-shifted (non-homogeneous) sites, combining two clusters increases the value of

Box 2. DIVERGE software

To study proteins under functional divergence, we have developed a new software system called DIVERGE (for 'detecting variability in evolutionary rates among genes') [a], which comes under the umbrella of PHYBA (phylogeny-based analysis) [b,c]. The emphasis in DIVERGE is on four main concerns: accuracy of results, ease of use, expandability and accessibility. DIVERGE has been developed to function identically whether the environment is Microsoft Windows 98 or NT, or Unix.

The operation of DIVERGE requires the user to supply a gene tree divided into multiple subtrees (neighbor-joining trees can be generated by DIVERGE), an amino acid replacement matrix (standard ones are provided) and a multiple sequence alignment. Also, if available, a suitable protein structure in PDB format can be provided by the user. After DIVERGE has performed the statistical analysis of the sequences, the user can examine important residues for functional divergence by their posterior probability analysis, which can then be plotted onto the alignment and/or the protein structure (if available). By allowing interactions with the protein structure, new discoveries about the relationships among residues, such as the spatial clustering of those with shifted rates, can be identified. The software is freely available at http://xgu1.zool.iastate.edu/

References

a Gu, X. and Vander Velden, K. (2002) DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. *Bioinformatics* 18, 500–501
b Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674
c Gu, X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* 18, 453–464

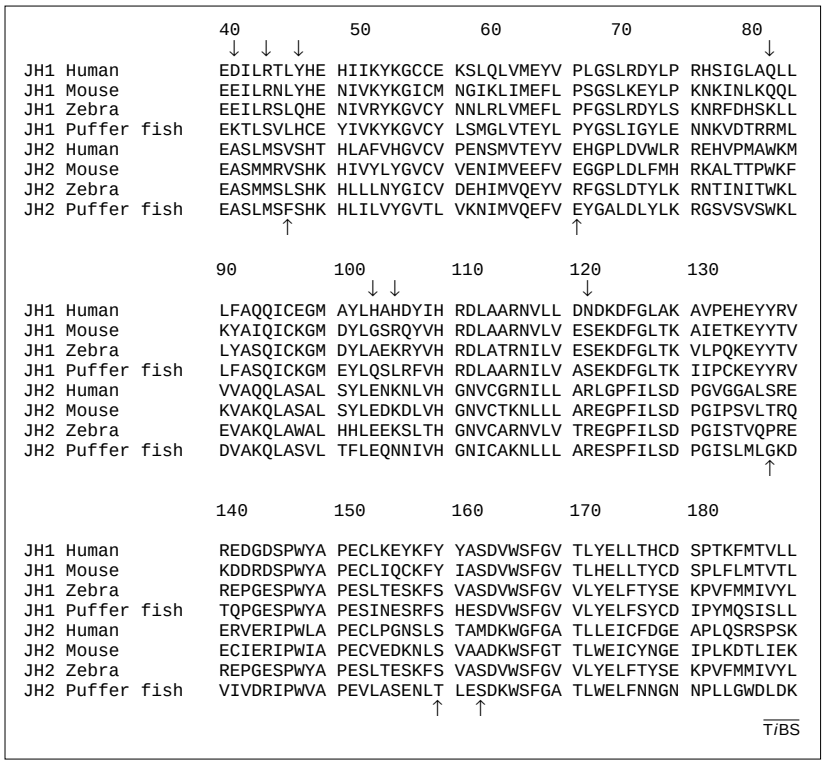

Fig. 1. Representative multiple sequence alignment (MSA) used to calculate θ (the coefficient of evolutionary functional divergence) between the JH1 and JH2 domains of Jak proteins [35]. The complete MSA contained 212 non-gapped positions (23 JH1 and 22 JH2 sequences). Positions with posterior probabilities of ≥95% are indicated with arrows: arrows above the sequences represent positions that are conserved in JH2 but variable in JH1, whereas arrows below the sequences represent positions that are conserved in JH1 but variable in JH2. Conserved amino acids at positions 103 (Glu, JH2) and 137 (Tyr, JH1) are important for the biochemical behavior of their respective domains, as demonstrated by mutagenesis experiments [35].

α relative to its separate estimates for each cluster [6]. This increase in α can be intuitively explained by reference to the previous diagrams of descendent and ancestral sequences (Box 1). For real sequences, such increases in α, when the data are combined, offer strong evidence that the gamma process is not homogeneous when adequate sample sizes are analyzed. This type of test was adopted by Gaucher *et al.* [31] to document non-homogeneous behavior in the elongation factors (EFs) of bacteria and eukaryotes.

Site-specific profile for identifying important residues

If the non-homogeneous gamma model provides statistical evidence for site-specific rate shifts after gene duplication or speciation, we next wish to identify specific sites in the protein that might have experienced a shift in their functional constraints. These sites are most likely to be relevant to our understanding of the structure–function basis of the differences between proteins. The posterior probability, denoted by $P(S_1 \mid X)$, of a site being in the S_1 state (i.e. type-I evolutionary functional divergence, or a site-specific rate shift), given the observed amino acid pattern at a particular position of the multiple sequence alignment, has been suggested as an indicator of this type of evolution [6,19]. This approach uses the empirical bayesian inference procedure, because the parameters in the prior distribution (θ) are estimated from the current data by the maximum likelihood method [19].

Several other measures have been proposed based on different criteria but for the same purpose [28,31]. Instead of the standard gamma model, Morozov *et al.* [28] adopted the technique of spectrum analysis (e.g. the Fourier transformation) to obtain a site-specific profile of evolutionary rates (or rate differences). Gaucher *et al.* [31] selected a set of important candidate sites using statistical quantiles as a measure. Technically, site-specific profiles can also be developed using Galtier's model [30], although this remains to be done. Several other research groups have developed region-specific profiles for rate differences based on the sliding-window procedure, which considers blocks instead of individual positions. Dermitzakis and Clark [43] used a method modified from Tang and Lewontin [44] to test whether the pattern of rate shifts is region specific in some vertebrate duplicate genes. Independently, Marin *et al.* [45] used a similar approach but took the biochemical properties of amino acids into account. These region-specific approaches are particularly useful when the number of available sequences is small. For instance, Dermitzakis and Clark's test is specifically designed for two gene clusters, each of which has only two sequences. In the future, it would be interesting to compare their performances.

In general, when phylogeny-based sequence analysis is coupled with information from protein crystal structures, considerable additional insight can be extracted from the evolutionary perspective. For example, Benner used this combination to identify active-site residues in the alcohol dehydrogenase family and to predict quaternary structure [3]. This approach was adopted in the 'evolutionary trace' approach [37] and later modified by Landgraf *et al.*, who used weighted replacements in their study of the heregulin gene family [39]. Most recently, Landgraf *et al.* incorporated site-specific rate shifts into the model [41]. Although an *ad hoc* scoring system is used, these approaches have been successful because they incorporate the structure–function correlation into comparative analyses.

Structural basis of site-specific rate shifts

The value of non-homogeneous gamma model analyses is perhaps best illustrated by specific case studies. Elongation factors Tu (EF-Tu) and 1A (eEF1A) are homologous proteins that are essential to translation in bacteria and eukaryotes, respectively. These GTPases catalyze the binding of aminoacyl tRNAs (aa-tRNAs) to the A site of the ribosome. Despite their similar overall roles in translation and very low rates of evolution, these proteins differ in several of their specific functions.

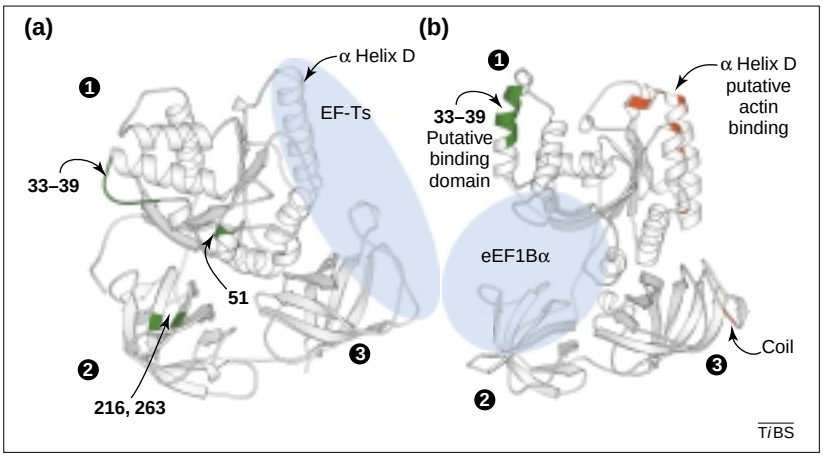

Fig. 2. Tertiary structures of (a) EF-Tu from *Thermus aquaticus* [47] and (b) eEF1A from *Saccharomyces cerevisiae* [48]. Green identifies sites from the posterior probability analysis that are evolving more slowly in eukaryotes than in bacteria, and red identifies sites that are evolving more quickly in eukaryotes than in bacteria. The ovals represent the EF-Ts- and eEF1Bα-binding domains for their respective elongation factors. The numbers 1–3 indicate the different domains of the proteins. Abbreviations: EF, elongation factor; eEF, eukaryotic elongation factor.

For example, EF-Tu regenerates its active form via the single-subunit nucleotide-exchange factor EF-Ts. By contrast, eEF1A is regenerated by the multisubunit nucleotide-exchange factor eEF1B. eEF1B is composed of the subunits α, β and γ; subunit α is responsible for eEF1A binding and for the catalytic activity of nucleotide exchange. eEF1A also interacts with eukaryotic cytoskeletal actin and might thereby play a role in tRNA channeling, cellular transformation and apoptosis. EF-Tu can have no such role in bacteria.

These more subtle changes in function between EF-Tu and eEF1A might correspond to differences in the evolutionary rates of their sites. To investigate this, EF sequences were analyzed [34] using the DIVERGE program [46] to identify positions that have undergone site-specific rate shifts ($\theta = 0.71 \pm 0.04$). A total of 24 sites were highlighted as evolving significantly faster in eukaryotes, whereas 25 sites were evolving significantly faster in bacteria, given their posterior probabilities of ≥95%. These sites were mapped onto the tertiary structures of bacterial [47] and eukaryotic [48] EFs, and correlated to cellular and biochemical data [34]. In all, 39 of the 49 sites were predicted to be associated with known binding domains on the EF structures. In fact, most of the 49 sites fall into multiple clusters and so are not randomly distributed across the protein structure.

Functional explanations for these non-homogeneous sites were based on differences in nucleotide, aminoacyl-tRNA, ribosome, actin and nucleotide-exchange-factor binding for the respective EF lineages. For example, although the mechanism of nucleotide exchange is conserved, the binding of their respective exchange factors is markedly distinct in the two lineages (Fig. 2). EF-Tu binds to EF-Ts through α-helix D and coils at the surfaces of domains 1 and 3, respectively. Eight positions within these EF-Tu–EF-Ts binding regions are evolving more rapidly in eukaryotes than in bacteria. Although they are present in eEF1A, these secondary structural elements are not involved in its binding to eEF1Bα. Instead, eEF1A binds to its nucleotide-exchange factor via contacts on the surfaces of domains 1 and 2. Three positions directly involved in this interaction (51, 216 and 263) are conserved in eukaryotes but have diverged in bacteria. Thus, these eight and three positions are evolving faster in eukaryotes and bacteria, respectively, because they are under few functional constraints for the binding of EFs to their respective exchange factors.

The structure of α-helix D is conserved by both EF-Tu and eEF1A, even though it is not involved in binding of the latter to eEF1Bα. Alternatively, α-helix D might be responsible for the binding of eEF1A to the actin of the eukaryotic cytoskeleton. This possibility is supported by the sequence similarity between α-helix D of eEF1A and the actin-binding region of depatin [49]. eEF1A occurs in both the nucleus and ribosomes, and binds both charged and uncharged tRNAs. Taken together, these arguments raise the intriguing corollary that these sites might, in eukaryotes, be responsible for the ability of eEF1A to channel tRNAs between the nucleus and ribosomes [50].

Seven positions with lower rates in eukaryotes than bacteria (residues 33–39) were predicted to form a unique α-helix at the surface of eEF1A, in combination with or separate from an adjacent insertion [31,34]. Given its charged and hydrophobic residues, this unique α helix was assigned a putative binding function in eEF1A. In EF-Tu, no such binding interactions were assigned to these sites, because they were neither conserved nor part of a rigid secondary structure. The posterior probabilities and the recently determined tertiary structure for eukaryotes confirm the status of these seven sites and document the existence of this α helix between positions 33 and 39 in eEF1A (Fig. 2). These results support the hypothesis that this unique secondary structure confers distinct binding properties on eEF1A. This review of changing rates in EFs highlights the power of combining the results of evolutionary rate analyses with knowledge from protein structure and function [20,31,40].

In contrast to the EF example, structure–function relationships based on shifts in evolutionary rates can be apparent even when θ is not high. The caspase (cysteine aspartyl protease) cascade is the key component in the apoptotic machinery of programmed cell death [51]. In vertebrates, 14 caspases are classified into the CED-3 and ICE subfamilies. The CED-3-type caspases are essential for most apoptotic pathways, and the major function of the ICE-type caspases is to mediate immune responses. Site-specific rate shifts between the CED-3 and ICE subfamilies predict altered functional constraints with statistical significance ($\theta = 0.29 \pm 0.05$) [20]. Moreover, the three-dimensional

structure has been used to correlate sites that display shifted evolutionary rates with structure–function differences between the protein subfamilies [52]. For example, site 161 has a posterior probability of ≥95% and holds a conserved tryptophan in all 22 sequences from the CED-3 subfamily. By contrast, a range of amino acid residues are present at this site in the ICE subfamily. This site is near the surface loop that is present in members of the CED-3 subfamily but absent in almost all ICE-type caspases. The high variability at this position within the ICE subfamily has been interpreted as possibly being caused by this tertiary-structure difference, which is responsible for the functional divergence between CED-3- and ICE-type caspases.

In spite of the utility of DIVERGE, it is difficult at this early stage to evaluate the false-positive (or false-negative) rate of prediction by DIVERGE, owing to the lack of sufficient experimental data and the diverse nature of functional divergence (e.g. biochemical, structural or phenotypic property). However, DIVERGE could have significant potential if its results are used appropriately as a guide for experimental design.

Concerns and future directions

Similar to almost all statistical models of sequence evolution, those that we have reviewed for detecting non-homogeneous rate behavior between groups share the assumption that individual sites are evolving independently. This assumption might not hold for sites that are close together in the sequence [53] and for more distant ones that interact functionally or structurally [3,54]. Indeed, the issue of site dependence was raised in the original covarion studies [24], in which replacements at one position were postulated to result in rate shifts at other positions. The study of co-evolving sites could be enhanced by using replacement matrices that are based on the physiochemical properties of the amino acids [55]. A maximum likelihood model has been described that accommodates the potential for co-evolution among sites [56]. Clearly, there is a need for further studies of site-to-site dependencies within a protein, and the incorporation of these constraints into more sophisticated models of sequence evolution.

In the future, it will be desirable to combine the current model of evolutionary functional divergence with site-dependent matrices of amino acid replacements [57,58]. This variable usage of replacement matrices across sites will acknowledge that variable sets of amino acids are permitted by selective and functional constraints at different positions. For example, at one site, only acidic amino acids might be permitted (i.e. Asp and Glu), whereas at another, only aromatic residues might be allowed (i.e. Phe, Trp and Tyr). Rather than relying on a general replacement matrix for all positions (e.g. the JTT [59] or PAM matrix), the use of site-specific ones could greatly enhance our ability to detect dependence, as well as rate shifts, across sites.

Given the nearly neutral theory of molecular evolution [60], population geneticists have argued that the evolutionary rate of a protein will be greater in a small population than in a large one. This rate increase is expected in small populations because more slightly deleterious alleles escape the purging effects of negative selection. Instead, these slightly deleterious alleles can drift and even become fixed in the population, thereby elevating the evolutionary rate of the protein. One might then argue that rate shifts at individual positions reflect the changes in population size rather than varying selective constraints. However, this concern about changing population size is not an issue for the rate-shift method described here because, in effect, θ relies on the degree of correlation between the rates per site of one group and the rates per site of another. Thus, θ compares the relative patterns of the rates per site rather than their absolute rates. Population size differences affect the overall rates of proteins but not the relative patterns of the rates among individual sites. As such, population factors (such as size) will not affect θ, and this can be demonstrated mathematically (such a mathematical argument is available from X. Gu upon request). In the same fashion, other widespread factors that could affect the overall rate for a protein (e.g. a change in the overall efficiency of a DNA repair system) are not a concern.

Conclusions

Whether referred to as site-specific rate shifts [32], heterotachy [36] or covarion-like behavior [23–27,29–31,33,34], the power of the non-homogeneous gamma model approach can be sufficient for it to address more than the simplest hypothesis. However, at the very least, the biologist needs a better understanding of what non-homogeneous pattern of evolutionary rate means. Although currently not available, databases that collect rate-shifted sites would be useful in this regard, especially ones that integrate protein sequences, phylogenetic trees and experimental information. These could be combined with existing high-throughput techniques for site-directed mutagenesis to test hypotheses about the biological meaning of site-specific rate differences [38].

Once these databases are available, we will be able to ask whether still more sophisticated models are necessary to capture biological information from genomic databases [61]. The non-homogeneous gamma model still does not capture the reality of protein sequence divergence in many ways (e.g. with regard to site independence and site-to-site differences in the replacement matrices). Considerable work is needed for us to decide whether more advanced statistical models are needed for the biologist.

Acknowledgements
We thank H. Stern for his comments about the statistics. Our work was supported in part by NIH grant RO1 GM62118 to X.G.; by a post-doctoral fellowship from the National Research Council and NASA Astrobiology Institute to E.A.G.; by the Dept Zoology, University of Florida and by NASA Exobiology grant NAG5-9030 to S.A.B. We also thank U.K. Das and J. Piasck for their assistance with figures.

References

1 Pauling, L. and Zuckerkandl, E. (1962) Molecular paleontology. *Acta Chem. Scand.* 17, S9–S16
2 Fitch, W.M. and Markowitz, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4, 579–593
3 Benner, S.A. (1989) Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv. Enzymol. Regul.* 28, 219–236
4 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
5 Benner, S.A. and Gaucher, E.A. (2001) Evolution, language and analogy in functional genomics. *Trends Genet.* 17, 414–418
6 Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674
7 O'Brien, S.J. *et al.* (1999) The promise of comparative genomics in mammals. *Science* 286, 458–481
8 Eisenberg, D. *et al.* (2000) Protein function in the post-genomic era. *Nature* 405, 823–826
9 Galas, D.J. (2001) Making sense of the sequence. *Science* 291, 1257–1260
10 Lewis, P.O. (2001) Phylogenetics systematics turns over a new leaf. *Trends Ecol. Evol.* 16, 30–37
11 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
12 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
13 Kimura, M. *The Neutral Theory of Molecular Evolution*, Cambridge University Press (in press)
14 Li, W-H. and Graur, D. (1991) *Fundamentals of Molecular Evolution*, Sinauer, Sunderland, MA, USA
15 Yang, Z.H. and Bielawski, J.P. (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503
16 Uzzell, T. and Corbin, K.W. (1971) Fitting discrete probability distributions to evolutionary events. *Science* 172, 1089–1096
17 Yang, Z.H. (1996) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372
18 Gu, X. *et al.* (1995) Maximum-likelihood-estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* 12, 546–557
19 Gu, X. (2001) Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* 18, 453–464
20 Wang, Y.F. and Gu, X. (2001) Functional divergence in caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics* 158, 1311–1320
21 Fitch, W.M. (1971) Toward defining course of evolution – minimum change for a specific tree topology. *Syst. Zool.* 20, 406–416
22 Shoemaker, J.S. and Fitch, W.M. (1989) Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Mol. Biol. Evol.* 6, 270–289
23 Fitch, W.M. and Ayala, F.J. (1994) The superoxide-dismutase molecular clock revisited. *Proc. Natl. Acad. Sci. U. S. A.* 91, 6802–6807
24 Miyamoto, M.M. and Fitch, W.M. (1995) Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* 12, 503–513
25 Lockhart, P.J. *et al.* (1998) A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* 15, 1183–1188
26 Tuffley, C. and Steel, M. (1998) Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147, 63–91
27 Lopez, P. *et al.* (1999) The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* 49, 496–508
28 Morozov, P. *et al.* (2000) A new method for characterizing replacement rate variation in molecular sequences: application of the Fourier and wavelet models to *Drosophila* and mammalian proteins. *Genetics* 154, 381–395
29 Philippe, H. *et al.* (2000) Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. London B Biol. Sci.* 267, 1213–1221
30 Galtier, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* 18, 866–873
31 Gaucher, E.A. *et al.* (2001) Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. *Proc. Natl. Acad. Sci. U. S. A.* 98, 548–552
32 Knudsen, B. and Miyamoto, M.M. (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14512–14517
33 Penny, D. *et al.* (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* 53, 711–723
34 Gaucher, E.A. *et al.* (2002) The crystal structure of eEF1A supports the functional predictions of an evolutionary analysis of rate changes among elongation factors. *Mol. Biol. Evol.* 19, 569–573
35 Gu, J. *et al.* Evolutionary analysis of functional divergence of Jak protein kinase domains and tissue-specific genes. *J. Mol. Evol.* (in press)
36 Lopez, P. *et al.* (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* 19, 1–7
37 Lichtarge, O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* 257, 342–358
38 Golding, G.B. and Dean, A.M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15, 355–369
39 Landgraf, R. *et al.* (1999) Analysis of heregulin symmetry by weighted evolutionary tracing. *Protein Eng.* 12, 943–951
40 Naylor, G.J.P. and Gerstein, M. (2000) Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. *J. Mol. Evol.* 51, 223–233
41 Landgraf, R. *et al.* (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* 307, 1487–1502
42 Jordan, I.K. *et al.* (2001) Sequence and structural aspects of functional diversification in class I α-mannosidase evolution. *Bioinformatics* 17, 965–976
43 Dermitzakis, E.T. and Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* 18, 557–562
44 Tang, H. and Lewontin, R.C. (1999) Locating regions of differential variability in DNA and protein sequences. *Genetics* 153, 485–495
45 Marin, I. *et al.* (2001) Detecting changes in the functional constraints of paralogous genes. *J. Mol. Evol.* 52, 17–28
46 Gu, X. and Vander Velden, K. (2002) DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. *Bioinformatics* 18, 500–501
47 Nissen, P. *et al.* (1995) Crystal structure of the ternary complex of Phe-tRNA(Phe), EF-Tu, and a GTP analog. *Science* 270, 1464–1472
48 Andersen, G.R. *et al.* (2000) Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A, eEF1Bα. *Mol. Cell* 6, 1261–1266
49 Yang, F. *et al.* (1990) Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. *Nature* 347, 494–496
50 Duttaroy, A. *et al.* (1998) Apoptosis rate can be accelerated or decelerated by overexpression or reduction of the level of elongation factor-1 α. *Exp. Cell Res.* 238, 168–176
51 Nicholson, D.W. and Thornberry, N.A. (1997) Caspases: killer proteases. *Trends Biochem. Sci.* 22, 299–306
52 Rotonda, J. *et al.* (1996) The three-dimensional structure of apopain/CPP32, a key mediator of apoptosis. *Nat. Struct. Biol.* 7, 619–625
53 Cohen, M.A. *et al.* (1994) Analysis of mutation during divergent evolution. The 400 by 400 dipeptide mutation matrix. *Biochem. Biophys. Res. Commun.* 199, 489–496
54 Atchley, W.R. *et al.* (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* 17, 164–178
55 Tuff, P. and Darlu, P. (2000) Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol. Biol. Evol.* 17, 1753–1759
56 Pollock, D.D. *et al.* (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* 287, 187–198
57 Thorne, J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* 10, 602–605
58 Fornasari, M.S. *et al.* (2002) Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. *Mol. Biol. Evol.* 19, 352–356
59 Jones, D.T. *et al.* (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8, 275–282
60 Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286
61 Felsenstein, J. (2001) Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J. Mol. Evol.* 53, 447–455