Function–structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors

Eric A. Gaucher*†, Michael M. Miyamoto‡, and Steven A. Benner*

*Department of Chemistry and Molecular Cell Biology Program, College of Medicine, University of Florida, Gainesville, FL 32611-7200; and ‡Department of Zoology, University of Florida, Gainesville, FL 32611-8525

Communicated by Walter M. Fitch, University of California, Irvine, CA, November 8, 2000 (received for review July 26, 2000)

The divergent evolution of protein sequences from genomic databases can be analyzed by the use of different mathematical models. The most common treat all sites in a protein sequence as equally variable. More sophisticated models acknowledge the fact that purifying selection generally tolerates variable amounts of amino acid replacement at different positions in a protein sequence. In their ''stationary'' versions, such models assume that the replacement rate at individual positions remains constant throughout evolutionary history. ''Nonstationary'' covarion versions, however, allow the replacement rate at a position to vary in different branches of the evolutionary tree. Recently, statistical methods have been developed that highlight this type of variation in replacement rates. Here, we show how positions that have variable rates of divergence in different regions of a tree (''covarion behavior''), coupled with analyses of experimental three-dimensional structures, can provide experimentally testable hypotheses that relate individual amino acid residues to specific functional differences in those branches. We illustrate this in the elongation factor family of proteins as a paradigm for applications of this type of analysis in functional genomics generally.

Elongation factors Tu (EF-Tu) and 1α (EF-1α) are homologous proteins essential to translation in bacteria and eukaryotes, respectively (1, 2). These GTPases catalyze the binding of aminoacyl-tRNAs to the A-site of the ribosome. As they are among the slowest evolving proteins known, EFs are commonly used to study cellular functions (2–4) and to root the universal tree of life (5, 6). This sequence stability presumably reflects enormous functional constraints on the divergent evolution of EFs, highlighting their central role in translation since the last common ancestor of the three primary domains of life (7). Nevertheless, EF-Tu and EF-1α differ in several of their specific functions (1, 2). For example, bacterial EF-Tu binds GDP $\approx$100-fold tighter than GTP. Eukaryotic EF-1α, in contrast, binds both with similar affinities. EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts. EF-1α requires the multisubunit nucleotide exchange factor EF-1$\beta\gamma\delta$. EF-1α also interacts with the eukaryotic cytoskeleton and thereby may play a role in cellular transformation and apoptosis (2, 3). EF-Tu can have no such role in bacteria.

These shifts in function must correspond at some level to changes in protein sequence. Thus, functional changes can leave signatures in the sequences of a protein family, which then can be detected with a well-constructed history of their relationships and replacements. In many cases, it appears possible to identify this record from the background noise of molecular evolution. In alcohol dehydrogenase (8) and superoxide dismutase (9), for example, previous studies have shown that variable replacement rates at specific positions can generate inferences relating changes in sequence structure to those in function. These proteins, however, have diverged far more rapidly than EFs. Furthermore, these studies have used neither the full power of a mathematical evolutionary (8) nor a crystallographic (9) analysis. We show here how this combination is of value in functional genomics, even in proteins not generally regarded as good examples of functional divergence.

From a mathematical perspective, the most common way to model rate heterogeneity among sequence positions is the gamma distribution, with its shape parameter α (10, 11). This distribution can accommodate a wide range of rapidly and slowly evolving sites. However, this model assumes a stationary substitution process, whereby positions retain their same relative rates of change throughout evolutionary history. This assumption is not expected to hold entirely true for proteins that change function. As an alternative, the covarion model proposes that the replacement rates of amino acid positions can change over time (9, 12–15). Although EFs might be expected to follow only a gamma model, given their overall functional conservation, previous studies have instead suggested that a covarion process is needed to adequately describe their evolution (5, 16, 17). This conclusion is examined more closely in this study and forms the basis of our integrated evolutionary and structural biology analyses of functional divergence between EF-Tu and EF-1α.

Methods

Thirty EF sequences were aligned by DARWIN (8) and then modified according to the secondary structures of EF-Tu for *Escherichia coli* (Protein Databank accession number 1EFC; ref. 18) and *Thermus aquaticus* (Protein Databank accession number 1TTT; ref. 19). This approach resulted in a multiple sequence alignment (MSA) with 380 aligned positions (cf. ref. 17). Maximum likelihood (ML) estimations of α and the replacement rates per site for all 380 aligned positions of EF-Tu versus EF-1α were accomplished with PAML, version 2.0, and its implementation of the Jones, Taylor, and Thorton model, with rate heterogeneity among sites according to the gamma distribution (JTT-Γ) (20). The Proportional, Poisson, and Dayhoff models for protein sequences were rejected as less appropriate for EFs on the basis of their log-likelihood ratio tests (21). The phylogeny in these ML analyses followed

Abbreviations: EF, elongation factor; MSA, multiple sequence alignment; ML, maximum likelihood.

†To whom reprint requests should be addressed at: Department of Chemistry, Leigh Hall 440, % Dr. Steve Benner's Lab, University of Florida, Gainesville, FL 32611-7200. E-mail: gaucher@ufl.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

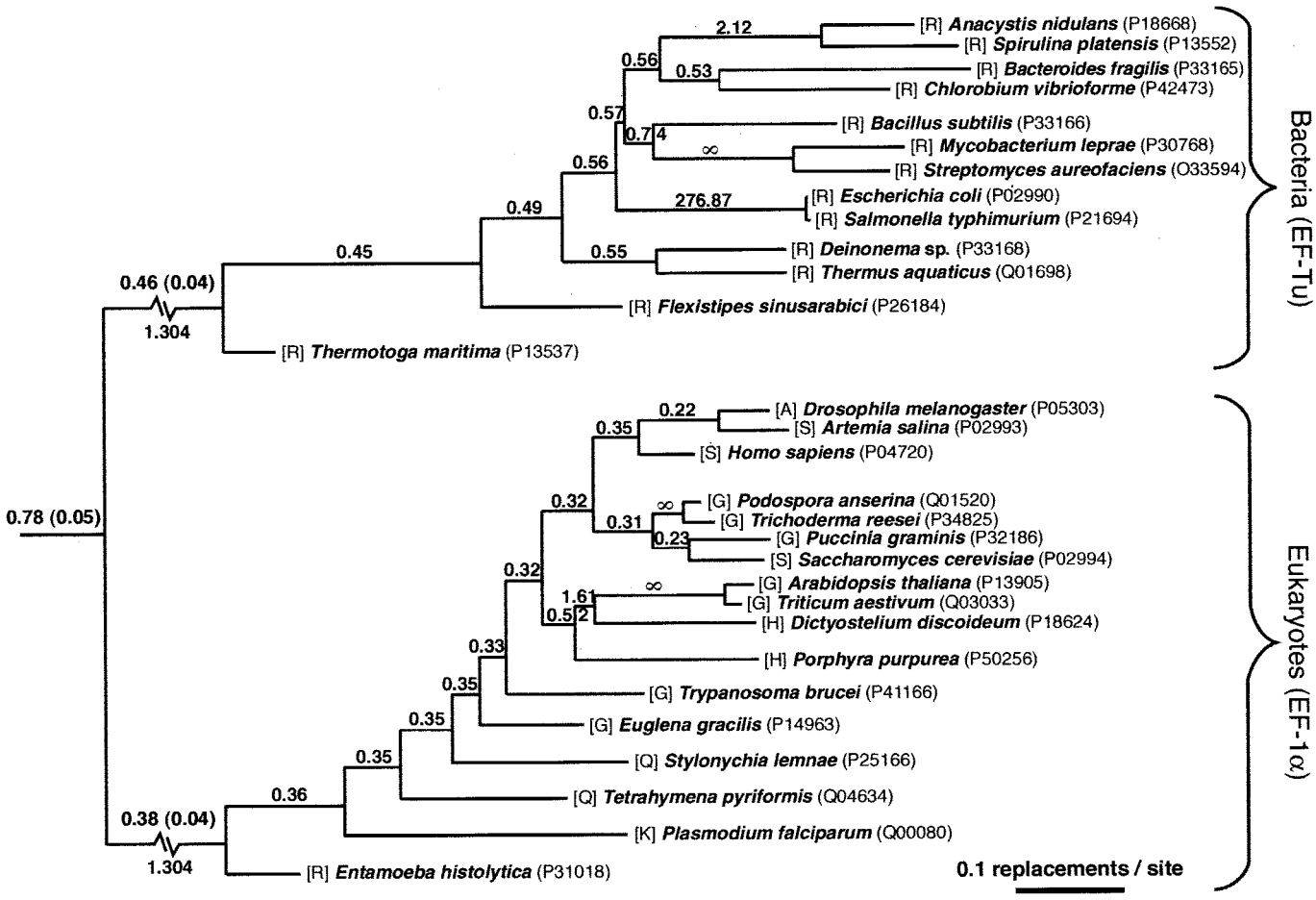


Fig. 1. Accepted phylogeny for bacteria and eukaryotes used in the ML analyses of their EF-Tu and EF-1α sequences. These sequences are from SWISS-PROT, with their accession numbers given in parentheses next to their species. Brackets refer to the amino acids of the two groups at position 305, a site illustrating a covarion pattern of sequence conservation in bacteria but considerable variation in eukaryotes. Branch lengths of this tree are drawn proportional to their ML estimates, except for the two longest internodes leading to bacteria and eukaryotes (both 1.30 replacements per site). The total tree length is 7.34 replacements per site (2.54 and 2.37 replacements per site for bacteria and eukaryotes alone, respectively). Numbers above internal branches represent the ML estimates of α for the corresponding group or subgroup of bacteria and/or eukaryotes. Standard deviations, as calculated from 20 rounds of parametric bootstrapping, are given in parentheses for the α values of bacteria, eukaryotes, and the two groups combined.

that of Bauldauf *et al.* (6), except for the topological positions of *Chlorobium* and *Salmonella*. As Bauldauf *et al.* (6) did not consider these two species, their topological positions were based on our follow-up ML analyses with MOLPHY, version 2.3 (22).

Parametric bootstrapping (evolutionary simulations) was conducted with PAML to calculate the SD of the α estimates for bacteria alone, eukaryotes alone, and the two groups combined (23). These simulations (20 per group) relied on the accepted tree and subtrees of bacteria and eukaryotes, their ML estimates of branch lengths and α, and the JTT-Γ model. In turn, subsampling experiments with bacteria alone, eukaryotes alone, and the two groups combined were completed to test for sample-size effects on their estimations of α (24). In these experiments, 20 random subsets apiece were generated for all odd-numbered subsamples from 5 to 11, 13, and 27 for bacteria, eukaryotes, and both groups, respectively. The α parameter then was reestimated for each random subsample for the same ML conditions as before. In recognition of their greater numbers, the subsampling trials with the two groups combined were stratified such that an extra eukaryotic sequence was selected relative to bacteria.

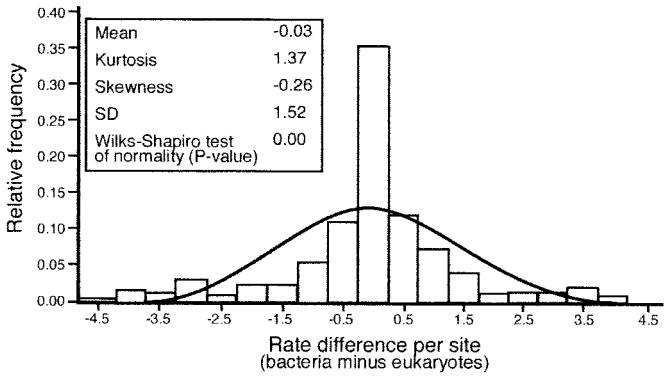
Normal distributions, sample kurtosis, skewness, and normality tests all were determined with SAS/GRAPH, release 6.03 (25). Visualization of protein structures was accomplished with CHEMSCAPE CHIME, release 2.0.3 (www.mdli.com) and PROTEIN EXPLORER, release 1.46 (www.umass.edu/microbio/chime/explorer).

Results and Discussion

Covarion Analyses, Structural Biology, and Hypothesis Generation. Our ML analyses of EF-Tu and EF-1α revealed a nonstationary α for different regions of the tree (Fig. 1). An α of 0.78 was calculated for the entire tree, with a SD of 0.05 from parametric bootstrapping. In contrast, the α values for both the bacterial and eukaryotic subtrees were significantly lower [$\alpha = 0.46$ (0.04) and $\alpha = 0.38$ (0.04), respectively]. Thus, a more uniform distribution of rates among sites was suggested when the two groups were considered together, rather than separately. Gu (14) statistically proved that such an increase in α is expected when the variable positions of one group are not the same as those of another (i.e., when the sequences are evolving under a nonstationary covarion process).

The distribution of rate differences per site between bacterial and eukaryotic EFs was leptokurtotic, i.e., over- and under-

a



b

Eukaryotes>bacteria (i.e., left tail)	Secondary structure	Properties, function
4, 69, 160, 290	β, β, L, β	Surface, no known function
117	α	Surface, in proximity to EF-Ts and kirromycin binding
102-103, 133, 135, 138, 141, 336	L, α, L α, α, L	Surface, all residues bind EF-Ts
189	L	On loop connecting domains 1 and 2
325-326	β	Surface, 2-3 residues away from aa-tRNA binding
253, 277, 305, 322	L, L, L, β	Surface, all residues bind aa-tRNA

Bacteria>eukaryotes (i.e., right tail)	Secondary structure	Properties, function
32-36	L	Surface, possible localization sites or ribosome binding
131	α	310 Helix binds GTP/GDP, faces away from nucleotides
153, 163	α, β	Interior, no known function
269	α	310 Helix, in proximity to aa-tRNA binding
263, 327, 329	β, β, β	Surface, 3-4 residues away from aa-tRNA binding sites
67, 123, 176, 311, 351	L, L, α, L, L	Surface, possible localization sites

Fig. 2. Rate differences per site between bacteria and eukaryotes. (*a*) Histogram of the site-by-site rate differences for the 380 aligned positions of bacteria minus eukaryotes. Sample kurtosis and skewness measure the ''peakedness'' and asymmetry of the histogram relative to the superimposed normal distribution, respectively (25). (*b*) Amino acid positions in the left and right tails of the histogram (i.e., those with rate differences of >2 SD between the two groups). Numbering refers to positions in the MSA. α, β, and L refer to α-helices, β-strands, and loops, respectively, following the three-dimensional structure of EF-Tu (Fig. 3).

represented in the mean and tails versus ''shoulders,'' respectively, relative to the expectations of a normal distribution (Fig. 2*a*). Nearly 50% of the positions had essentially the same rate in the two groups (rate differences of <0.5 replacements per site per unit evolutionary distance), as expected under a stationary gamma process. However, 17 sites were evolving >2 SD faster in bacteria than eukaryotes, whereas 19 were changing >2 SD faster in eukaryotes than in bacteria (Fig. 2*b*). These sites representing 10% of the MSA are suggestive of a covarion process in the EF-Tu/EF-1α family.

By integrating structural data with these ML rate differences, this initial pool of 36 sites can be further reduced to a subset of those positions that are most likely involved in the functional shifts between EF-Tu and EF-1α. For example, 10 sites in and around the region binding tRNAs are evolving >2 SD faster in either bacteria or eukaryotes (Figs. 2*b* and 3). These rate changes can be correlated with a difference in biochemical function between EF-Tu and EF-1α. EF-1α/GDP binds charged and uncharged tRNAs, whereas EF-Tu/GDP does not. Crystallographic data for EF-Tu reveals a major conformational shift between the GDP- and GTP-bound states, whereby the tRNA-binding site of the former is disrupted (Fig. 3*b*). In contrast, available data for EF-1α suggest that this conformational shift does not occur (see ref. 2 for a review). This correlation between rate differences and protein structure–function leads to the hypothesis that at least some of these 10 positions are responsible for the different interactions of EF-Tu and EF-1α with tRNA. This hypothesis can now be tested by introducing into EF-1α the residues of EF-Tu at these positions (26). The prediction is that these introductions will result in a variant of EF-1α/GDP that does not bind uncharged tRNA.

Similarly, eight sites in and around the region where nucleotide exchange factors bind are evolving >2 SD faster in eukaryotes than in bacteria (Figs. 2*b* and 3). EF-Tu regenerates its active form by binding to the single-subunit nucleotide exchange factor EF-Ts, whereas EF-1α depends on the multisubunit EF-1βγδ. The rate differences for these eight sites lead to the hypothesis that the surface area of EF-1α in contact with its nucleotide exchange complex is different from that for EF-Tu. This difference is consistent with the divergent structures of their respective nucleotide exchange factors (1, 2).

Perhaps the most intriguing functional difference between the two EFs is the ability of EF-1α to bind to actin, the main component of the eukaryotic cytoskeleton. This function, together with the ability of EF-1α (but not EF-Tu) to bind to uncharged tRNAs, may be important as a mechanism for tRNA channeling from the ribosome back to the nucleus (2, 27). Bacteria, of course, do not require channeling, thereby obviating the need for binding of uncharged tRNAs by either the GDP or the GTP state of EF-Tu. Relatively rapid sequence evolution is a general characteristic of surface residues that are not involved in protein–ligand interactions (28). Nine surface residues to which other contacts cannot be definitively assigned from biochemical and structural data were evolving >2 SD faster in EF-Tu than in EF-1α (Figs. 2*b* and 3*b*). These rate differences suggest the hypothesis that at least some of these residues in EF-1α are in contact with the actin cytoskeleton.

Positions 32–36 are conserved in EF-1α but variable in EF-Tu (Figs. 2*b* and 3*b*). In EF-Tu, biochemical and three-dimensional structural data show that this region is in proximity to the ribosome (29, 30). In EF-1α, positions 32–36 are followed by an insertion that is suggestive of a binding site with its characteristic charged amino acids and hydrophobic residues. In combination with its conserved residues 32–36, this insertion is predicted to introduce a regular secondary structural element of an α-helix (8, 31) that may reflect a difference in ribosomal structure and binding between bacteria and eukaryotes. Thus, another testable hypothesis is suggested by the integration of rate differences with protein structure and function.

How robust are our hypotheses with respect to the current sample of sequences? This question follows from the recent demonstration by Sullivan *et al.* (24) that ML estimates of rate variation among sites may be sensitive to taxon sampling. In

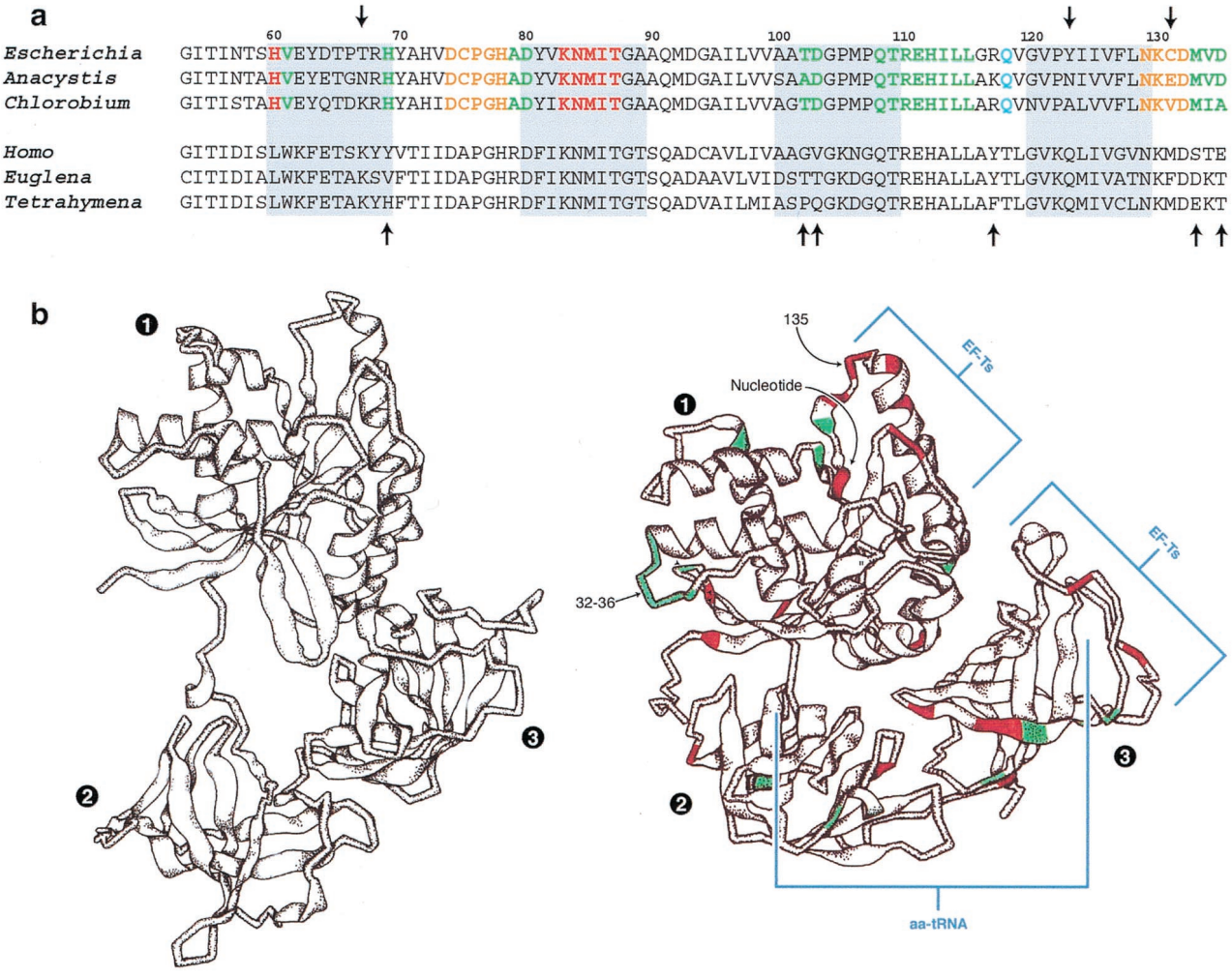


Fig. 3. MSA for EFs and tertiary structures for EF-Tu. (*a*) MSA for the ligand-binding region at the NH_2 terminus of three representative bacteria and three eukaryotes (*Upper* and *Lower*, respectively). This MSA highlights the key residues for aminoacyl-tRNA (red), EF-Ts (green), and nucleotide (yellow) binding and for kirromycin resistance (cyan), as determined for bacterial EF-Tu (1, 18, 19). Arrows, above and below the MSA, correspond to those sites that are evolving >2 SD faster in bacteria than in eukaryotes, and vice versa, respectively (positions 67, 69, 102–103, 117, 123, 131, 133, and 135) (Fig. 2). (*b*) Tertiary structures of the GDP- and GTP-bound states for EF-Tu from *E. coli* and *T. aquaticus*, respectively (18, 19). Here, green and red in the GTP confirmation highlight those sites that are evolving >2 SD faster in bacteria than in eukaryotes, and vice versa, respectively (Fig. 2).

our subsampling experiments, estimates of α were found to be upwardly biased for the smaller samples of all three groups (Fig. 4). Nevertheless, the same major difference between bacteria and eukaryotes alone versus combined was evident, regardless of the sample size. Also, α remained largely unchanged (within the range of statistical error) with the inclusion of 40 and 15 additional sequences from SWISS-PROT for bacteria (0.48) and eukaryotes (0.35), respectively. Given our initial focus on the fluctuating estimates of α for bacteria and eukaryotes, our study did not consider Archaebacteria. However, our more recent investigations of EFs document that this group is defined by an α (0.88) that is more similar to the combined estimate for bacteria and eukaryotes than to their separate values. Collectively, these various results argue against sampling error as an explanation for the nonstationary behavior of α for EF-Tu versus EF-1α.

Covarion Approaches and Functional Genomics. Functional genomics is the bridge between computational and experimental biology (32, 33). The field combines sequence data with general knowledge to generate testable hypotheses about the biological functions of genes and proteins. Today, most hypotheses in the field are generated from sequence similarity searches with BLAST (34) or FASTA (35). The function of the probe sequence is assumed to equal that of the best annotated hit that is recovered in these similarity searches.

Functional genomics is actively seeking tools to detect changes in protein function from their sequences and estimated history (14, 36). The best-known approach for this purpose uses the ratio of nonsynonymous to synonymous substitutions to identify potential cases of functional change (36–38). This approach, however, suffers as a signature of functional change among distant branches, because silent sites quickly lose their signal as they become saturated with substitutions. Shifts in protein function can also be deduced from instances of convergent or parallel evolution (39). In turn, functional constraints can be detected as compensatory covariation, whereby different residues in contact are sequentially replaced in a way that conserves some overall physical property (40).

The covarion approach now offers another tool for studying the evolution of protein function (14). Variability is a feature of a position that reflects its relation to selected function. Thus, changes across groups in the variability of their sites offer insights into which positions of a protein may be most responsible

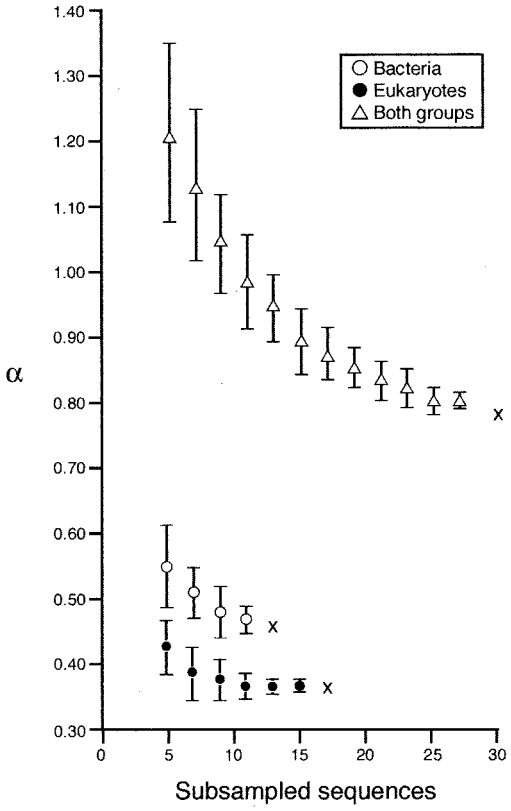


Fig. 4. The effect of sequence sample size on the ML estimation of α for bacteria alone, eukaryotes alone, and the two groups combined. Xs correspond to the final estimates of α for each group. Twenty subsampling experiments were completed for each sample size of a group, with the results summarized as means and their SD.

for its functional shifts. If the variability of many positions changes, then the inference can be made that the protein has acquired a new function (or lost its function). However, this study with EFs illustrates how much our concept of function is contingent on one's perspective and how subtle such shifts can be. In detail, EF-Tu and EF-1α function in different ways, even though their overall role in translation has remained the same. These more subtle but nevertheless significant functional differences involve on the order of 10% of the sites according to our covarion analysis (Fig. 2).

Our approach integrates structural data with a covarion-based evolutionary analysis to improve the identification of those relatively few sites that are largely responsible for the functional differences between EF-Tu and EF-1α. Together, these two sources of information allow us to target specific positions and residues for the direct experimentation of their effects on the function of EFs. Of particular interest are the surface residues that are evolving >2 SD slower in eukaryotes than in bacteria (Fig. 2). If confirmed by direct testing, the involvement of at least some of these sites in binding EF-1α to actin would constitute one of the only examples where metabolic channeling, long an issue in central pathways, has left a signature in the sequences themselves (41). It is as a tool for hypothesis generation and experimental design that covarion-based evolutionary studies, coupled with structural biology, will make their greatest contributions to functional genomics.

We thank J. Sullivan for providing us with a preprint of his article; M. D. Caraco, D. Schreiber, and S. Govindarajan for their assistance with the computer analyses; J. Piascik for drafting our figures; A. S. Edison, J. Nyborg, D. L. Swofford, M. R. Tennant, and Z. Yang for their valuable suggestions about our research and manuscript; and the National Institutes of Health, National Aeronautics and Space Administration Astrobiology Institute, and Department of Zoology, University of Florida for their financial support.

1. Krab, I. M. & Parmeggiani, A. (1998) *Biochim. Biophys. Acta* **1443,** 1–22.
2. Negrutskii, B. S. & El'skaya, A. V. (1998) *Prog. Nucleic Acid Res. Mol. Biol.* **60,** 47–78.
3. Yang, F., Demma, M., Warren, V., Dharmawardhane, S. & Condeelis, J. (1990) *Nature (London)* **347,** 494–496.
4. Duttaroy, A., Bourbeau, D., Wang, X. L. & Wang, E. (1998) *Exp. Cell Res.* **238,** 168–176.
5. Lopez, P., Forterre, P. & Philippe, H. (1999) *J. Mol. Evol.* **49,** 496–508.
6. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7749–7754.
7. Benner, S. A., Ellington, A. D. & Tauer, A. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7054–7058.
8. Benner, S. A., Trabesinger, N. & Schreiber, D. (1998) *Adv. Enzyme Regul.* **38,** 155–180.
9. Miyamoto, M. M. & Fitch, W. M. (1995) *Mol. Biol. Evol.* **12,** 503–513.
10. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996) in *Molecular Systematics*, eds. Hillis, D. M., Moritz, C. & Mable, B. K (Sinauer, Sunderland, MA), 2nd Ed., pp. 407–514.
11. Yang, Z. (1996) *Trends Ecol. Evol.* **11,** 367–372.
12. Fitch, W. M. & Markowitz, E. (1970) *Biochem. Genet.* **4,** 579–593.
13. Tuffley, C. & Steel, M. (1998) *Math. Biosci.* **147,** 62–91.
14. Gu, X. (1999) *Mol. Biol. Evol.* **16,** 1664–1674.
15. Morozov, P., Sitnikova, T., Churchill, G., Ayala, F. J. & Rzhetsky, A. (2000) *Genetics* **154,** 381–395.
16. Lockhart, P. J., Steel, M. A., Barbrook, A. C., Huson, D. H., Charleston, M. A. & Howe, C. J. (1998) *Mol. Biol. Evol.* **15,** 1183–1188.
17. Moreira, D., Le Guyader, H. & Philippe, H. (1999) *Mol. Biol. Evol.* **16,** 234–245.
18. Song, H., Parsons, M. R., Rowsell, S., Leonard, G. & Philips, S. E. V. (1999) *J. Mol. Biol.* **285,** 1245–1256.
19. Nissen, P., Kjeldgaard, M., Thirup, S., Polekhina, G., Reshetnikova, L., Clark, B. F. & Nyborg, J. (1995) *Science* **270,** 1464–1472.
20. Yang, Z. (1997) *Comput. Appl. Biosci.* **15,** 555–556.
21. Huelsenbeck, J. P. & Rannala, B. (1997) *Science* **276,** 227–232.
22. Adachi, J. & Hasegawa, M. (1996) *Comput. Sci. Monogr.* **28,** 1–150.
23. Huelsenbeck, J. P., Hillis, D. M. & Jones, R. (1995) in *Molecular Zoology: Advances, Strategies, and Protocols*, eds. Ferraris, J. & Palumbi, S. (Wiley, New York), pp. 19–45.
24. Sullivan, J., Swofford, D. L. & Naylor, G. J. P. (1999) *Mol. Biol. Evol.* **16,** 1347–1356.
25. SAS Institute (1988) *SAS/Graph User's Guide* (SAS Inst., Cary, NC), Release 6, 03., Ed. 549.
26. Golding, G. B. & Dean, A. M. (1998) *Mol. Biol. Evol.* **15,** 355–369.
27. Grosshans, H., Simos, G. & Hurt, E. (2000) *J. Struct. Biol.* **129,** 288–294.
28. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996) *J. Mol. Biol.* **257,** 342–358.
29. Peter, M. E., Schirmer, N. K., Reiser, C. O. A. & Sprinzl, M. (1990) *Biochemistry* **29,** 2876–2884.
30. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. & Steitz, T. A. (1999) *Nature (London)* **400,** 841–847.
31. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
32. Bork, P. & Koonin, E. V. (1998) *Nat. Genet.* **18,** 313–318.
33. Benner, S. A., Chamberlin, S. G., Liberles, D. A., Govindarajan, S. & Knecht, L. (2000) *Res. Microbiol.* **151,** 97–106.
34. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227,** 1435–1441.
35. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
36. Yang, Z. H., Nielsen, R., Goldman, N. & Pedersen, A. M. K. (2000) *Genetics* **155,** 431–449.
37. Li, W.-H., Wu, C.-I. & Luo, C.-C. *(1985) Mol. Biol. Evol.* **2,** 150–174.
38. Messier, W. & Stewart, C. B. (1997) *Nature (London)* **385,** 151–154.
39. Stewart, C. B., Schilling, J. W. & Wilson, A. C. (1987) *Nature (London)* **330,** 401–404.
40. Chelvanayagam, G., Eggenschwiler, A., Knecht, L., Gonnet, G. H. & Benner, S. A. (1997) *Protein Eng.* **10,** 307–316.
41. Reddy, G. P. V. & Pardee, A. B. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 3312–3316.