Journal of Phylogenetics & Evolutionary Biology

Lee and Brocchieri, J Phylogen Evolution Biol 2013, 1:1
DOI: 10.4172/2329-9002.1000e102

Editorial Open Access

The Evolution of Fuzzy Proteins

Hye Won Lee[1] and Luciano Brocchieri[2]*

[1]Genetics and Genomics Graduate Program, University of Florida, Gainesville, FL, USA
[2]Department of Molecular Genetics & Microbiology and Genetics Institute, University of Florida, Gainesville FL, USA

Advances in sequencing technology encourage the accumulation of molecular data and the development of phylogenetic methods that use nucleotide or amino acid sequences to study the evolution of gene and protein families, and the phylogenetic relations of species. Phylogenetic tree reconstructions are based on a choice of algorithms, and rely on the accuracy of nucleotide or amino acid substitution models in describing the process of molecular evolution. Here, we describe recent approaches to modeling protein evolution and their biological interpretation based on the concept of "fuzzy protein".

Probabilistic Approaches to Phylogenetic Tree Inference

Probabilistic approaches, including Maximum-Likelihood (ML) and Bayesian methods, are widely used and considered the most accurate in phylogenetic inference [1,2]. In probabilistic phylogenetic methods, protein evolution is modeled as a continuous Markov process, described by a matrix $Q=\{q_{ij}\}$ of amino acid *transition rates,* from amino acid type *i* to amino acid type *j* [2]. Q is derived by combining a symmetric substitutability matrix R, and a vector of amino acid equilibrium frequencies π, to obtain transition rates $q_{ij}=Cr_{ij}\pi_j$ $(i \neq j)$. The diagonal terms $q_{ii}=-\sum_{j\neq i} q_{ij}$ are normalized by choosing *C*, so that $-\pi_i \sum_i q_{ii}=1.0$, that is, rates are scaled so that one unit of relative evolutionary time corresponds on average to one substitution per site. Q establishes a relation between evolutionary distance and expected sequence similarity, by which the evolutionary distance between two sequences can be inferred based on their sequence identity. Furthermore, Q can be used to calculate the likelihood of a phylogenetic tree in ML methods, or the ratio of the posterior probabilities of two phyogenetic trees in Bayesian approaches, to identify the optimal tree(s).

The observation from sequence alignments of related proteins that different protein sites show different propensities to differentiate, however, suggests the opportunity to incorporate in evolutionary models devices to model site heterogeneity of the evolutionary process. A traditionally used way to do so is to assign different *rates* of evolution to different sites. This is generally accomplished by rescaling Q by a coefficient v_k specific to each site *k*, so that $Q^{(k)}=Qv_k$ and $\sum_k v_k = 1.0$. The value of v_k is most commonly drawn from a discretized gamma distribution $\Gamma(\alpha,\alpha)$, whose shape is optimized by the choice of α [3-6]. A second, commonly used device to fit site-dependence of evolutionary rates is to allow for a fraction *I* of *invariable* sites [7-9]. In a model including both invariable and gamma distributed sites (I+Γ), a rate-coefficient *v*=0 is assigned to a fraction *I* of sites, and gamma-distributed positive rates are assigned to the remaining fraction (1–*I*) of sites. Evolutionary rates that substantially vary across sites have a significant effect on the relation between evolutionary distance and sequence similarity (Figure 1), as substitutions that would otherwise uniformly spread across all sites, tend instead to cumulate at fewer, fast evolving sites.

While site-specific rates affect the speed of evolution, they do not affect the evolutionary pattern of each position. Different evolutionary patterns can instead be fitted to individual sites by deriving site-specific Q matrices. Remembering how Q is constructed, this can be accomplished by allowing site-specificity to R, to π, or to both. The first choice, implemented in the QMM model [10], is computationally quite challenging, requiring the optimization of 189 parameters per site-class. A relatively simpler approach is to allow for site-specific stationary

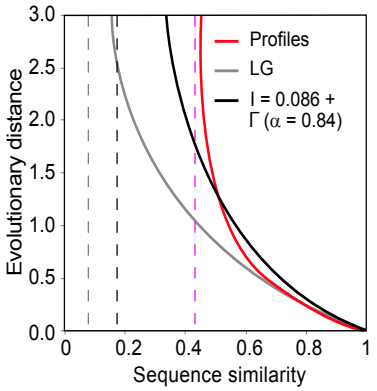

Figure 1: Expected relation between sequence similarity and evolutionary distance, as predicted by the LG model (LG) [18], by the LG model with estimated proportion of invariable sites and gamma distributed rates (I+Γ), and by profiles based on the amino acid types observed at each alignment position of bacterial DnaK sequences.

frequencies $\pi^{(k)}$. This approach also appears consistent with the observation from multiple sequence alignments that different subsets of amino acid types typically occupy at different sites. Site-specificity of equilibrium frequencies has many interesting repercussions on the features of the evolutionary process, on phylogenetic tree reconstruction, and on the relation between sequence conservation and mutational saturation.

Position Specific Profiles of Amino Acid Usage

Possibly, the most successful implementation of the idea of site-specificity of amino acid stationary distributions is the CAT mixture model of Lartillot and collaborators [11-15]. In the CAT (category) model, amino acid equilibrium frequencies $\pi^{(k)}$ were empirically identified using a Bayesian approach [11]. To speed up computation, sets of preassembled profiles of amino acid frequencies are provided in ML and Bayesian phylogenetic reconstruction implementations [16,17]. Profiles $\pi^{(k)}$ specific to each site *k* are used in combination with a general substitutability matrix R, to construct site-specific *normalized* transition-rate matrices $Q^{(k)}$, with $q_{ij}^{(k)}=C^{(k)}r_{ij}\pi_j^{(k)}$, $q_{ii}^{(k)}=-\sum_{j\neq i} q_{ij}^{(k)}$, and $C^{(k)}$ such that $-\pi_i^{(k)}\sum_i q_{ii}^{(k)}=1.0$. In comparison to the global vector of stationary frequencies, as implemented, for example in the LG model [18], profiles of the CAT model tend to favor different subsets of amino acid types with similar physico-chemical properties

***Corresponding author:** Luciano Brocchieri, Department of Molecular Genetics & Microbiology and Genetics Institute, University of Florida, Gainesville FL, USA, Tel: +1 352 273 8131; E-mail: lucianob@ufl.edu

Received December 24, 2012; **Accepted** December 28, 2012; **Published** December 31, 2012

Citation: Lee HW, Brocchieri L (2013) The Evolution of Fuzzy Proteins. J Phylogen Evolution Biol 1: e102. doi:10.4172/2329-9002.1000e102

Copyright: © 2013 Lee HW, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

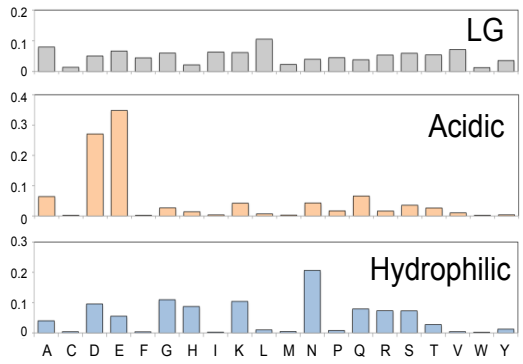

Figure 2: Amino acid equilibrium frequencies from the LG model [18], compared to those described by two profiles from the C20 set [16,17].

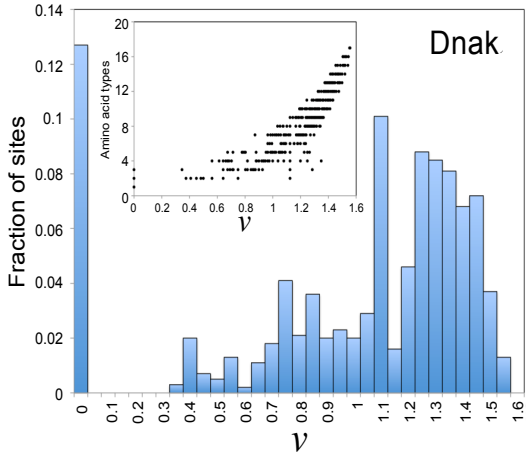

Figure 3: Distribution of site-specific relative rates of transition obtained from an alignment of about 500 DnaK proteins, as described in the main text. The insert shows the relation between these rates and the number of different amino acid types observed at corresponding positions.

(Figure 2). As a consequence, while under a generalized Q matrix amino acid substitutions tend to wander over time across all 20 types, within a profile divergence is highly constrained within fewer amino acid types, no matter how much evolution occurs, increasing the probability of homoplasy. Furthermore, the reduced effective size of the amino acid alphabet at each site produces higher expected similarity between sequences even at high evolutionary distance. For example, the generalized LG model [18] predicts that over time sequence similarity divergences to the asymptotic value of 5.996%. Profiles in the C20 set implemented in the Phylobayes [16] and PhyML [17] methods predict instead, on average, sequence divergence to 18.37% similarity, with a range for individual profiles from 7.54% to 33.56% similarity. Thus, the CAT model estimates that generalized models under-estimate the evolutionary distances of sequences of low similarity (Figure 1), providing an explanation for the phenomenon of long branch attraction [13].

Profiles, Fuzzy Proteins, and Neutral Constrained Amino Acid Replacements

Position-specific equilibrium frequency profiles are justified by the idea that functionality and structural stability of a protein requires certain residue types at certain positions, with different degrees of stringency, depending on functional constraints. For example, a position corresponding to an active site may correspond to a profile with one amino acid type, whereas different hydrophilic amino acid types may be allowed to substitute in loops exposed at the protein surface. This suggests an interpretation of profiles based on a model of neutral constrained evolution [19]. According to this interpretation, the profile associated with a particular position defines a subset of amino acid types that can be substituted at that position, without affecting the fitness of the protein (i.e., its functionality). This model asserts that a protein can be described as a functional unit as a possibly large set of alternative sequences, each functionally equivalent to the other. Thus, from a functional perspective, a protein would be described, rather than by a sequence of amino acids, by a sequence of amino acid subsets, whose size describes different degrees of "fuzziness" of different positions. A "fuzzy protein" can evolve within the limits imposed by the sequence of amino acid subsets that describe it with no effect on functionality. With this interpretation, position-specific profiles can explain not only the evolutionary pattern, but also the speed of evolution. The reasoning is that random substitutions will be retained only if they result in substitutions allowed by the profile. Thus, if the profile is stringent, most substitutions will be rejected slowing the evolutionary process; if the profile is permissive, most substitutions will be accepted, resulting in fast evolution. To model the effect of purifying selection on evolutionary rates, we first considered a general, normalized substitution-rate matrix, whose coefficients are derived from nucleotide and codon substitution matrices. At each position *k*, the substitution matrix is filtered by a position-specific "occupancy vector" that defines the subset of amino acid types allowed at that position, so that equilibrium frequencies and transformation rates towards amino acid types not represented in the occupancy vector are set to zero. The result is a $Q^{(k)}$ matrix with a slower average transition rate $-\pi_i^{(k)}\sum_i q_{ii}^{(k)} < 1.0$. All matrices are finally renormalized to an average one substitution per site. With this model, selection against not-allowed transformations generates site-specific profiles of amino acid equilibrium frequencies, and a distribution across sites of different site-specific evolutionary rates, approximately proportional to the size of the profiles (Figure 3). The profiles that define a fuzzy protein are not likely to correspond to those identified by the CAT model, which combine substitutions *within* a profile with substitutions *between* profiles, taking into account events of profile evolution. An interesting question is how each process contributes to protein evolution.

Acknowledgements

This work is supported by NIH Grant 5R01GM87485-2.

References

1. Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA, USA.
2. Gascuel O (2007) Mathematics of evolution and phylogeny. Oxford University Press, USA.
3. Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172: 1089-1096.
4. Nei M, Chakraborty R, Fuerst PA (1976) Infinite allele model with varying mutation rate. Proc Natl Acad Sci U S A 73: 4164-4168.
5. Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7: 82-102.
6. Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10: 1396-1401.
7. Churchill GA, von Haeseler A, Navidi WC (1992) Sample size for a phylogenetic inference. Mol Biol Evol 9: 753-769.

8. Reeves JH (1992) Heterogenetity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol 35: 17-31.

9. Hillis DM, Moritz C, Mable BK (1996) Molecular systematics. (2nd edn), Sinaur Associates, Sunderland, MA, USA.

10. Wang HC, Li K, Susko E, Roger AJ (2008) A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol 8: 331.

11. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21: 1095-1109.

12. Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci 363: 1463-1472.

13. Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7: S4.

14. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25: 2286-2288.

15. Blanquart S, Lartillot N (2006) A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol 23: 2058-2071.

16. Quang le S, Gascuel O, Lartillot N (2008) Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24: 2317-2323.

17. Quang le S, Lartillot N, Gascuel O (2008) Phylogenetic mixture models for proteins. Phil Trans R Soc Lond B Biol Sci 363: 3965-3976.

18. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307-1320.

19. Brocchieri L (2001) Phylogenetic inferences from molecular sequences: review and critique. Theor Popul Biol 59: 27-40.