Epistasis produces an excess of invariant sites in neutral molecular evolution

Ravi Patel[a,b] and Sudhir Kumar[a,b,1]

Substitution rate variability among sites is a common feature of protein evolution. This variability is frequently modeled by a gamma (Γ) distribution (1), which is proposed to be an emergent property of epistasis by de la Paz et al. (2). However, the conclusion of de la Paz et al. (2) is based on the analysis of highly divergent sequences, as many as 24 substitutions per site. We examined if this conclusion holds for more biologically realistic sequence divergences (0.1 to 5 substitutions per site) (3). Instead, we found that creation of a class of invariant sites is an emergent property of epistasis, and the sequence divergence level dictates whether an equal- or a variable-rate model will best fit the data.

In de la Paz et al. (2), neutral molecular evolution was simulated for protein families with a fitness function defined by a model inferred via direct coupling analysis (e.g., Fig. 1*A*). We counted the number of sites that experienced zero, one, two, three, and more substitutions directly from the generation by generation simulated sequences (e.g., Fig. 1*B*), which avoids confounding effects of estimation errors. This site frequency spectrum (SFS) will follow a Poisson distribution when the data are compatible with a single (S) evolutionary rate across sites (4). SFS will follow a negative binomial distribution if the evolutionary rates among sites are Γ distributed (1). We tested model fits to SFS at different levels of sequence divergence, allowing for a category of invariant sites (I+) with S and Γ models (5). For each dataset, the best-fit model from four candidates (S, I + S, Γ, and I + Γ models) was selected using corrected Akaike information criterion (AIC_c) because these models are not all nested (6, 7).

The S model fits best for a vast majority of datasets at sequence divergences less than 0.5 substitutions per site (Fig. 1*C*). That is, epistasis did not create significant evolutionary rate variability among sites at low sequence divergences. For divergences larger than 0.5 substitutions per site, models allowing for an invariant class of sites (I + S and I + Γ) provide the best fits (Fig. 1*C*). Similar patterns are observed for the other nine protein domain families analyzed by de la Paz et al. (2) (Fig. 2). The Γ model alone was not the dominant model except for one case (Fig. 2).

Therefore, at sufficient evolutionary distances, epistasis generates many invariant sites and substitution rate variability, likely because incompatible substitutions at coupled sites are under greater negative selection than substitutions at loosely coupled and uncoupled sites. At relatively low divergences, fewer coupled sites are mutated, likely resulting in less differential negative selection among sites. So, even a single-rate model may fit the data well. Overall, our results suggest that epistasis provides a mechanistic explanation for the abundance of invariant sites beyond what is explained by an equal (S)- or a variable (Γ)-rate model. Consequently, models incorporating a separate class of invariant sites best describe evolutionary rates of proteins.

Acknowledgments

This work was supported by NSF Grant 934848 and NIH Grant GM139540.

[a]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122; and [b]Department of Biology, Temple University, Philadelphia, PA 19122

Author contributions: S.K. designed research; R.P. and S.K. performed research; R.P. analyzed data; and R.P. and S.K. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: s.kumar@temple.edu.

Published April 19, 2021.

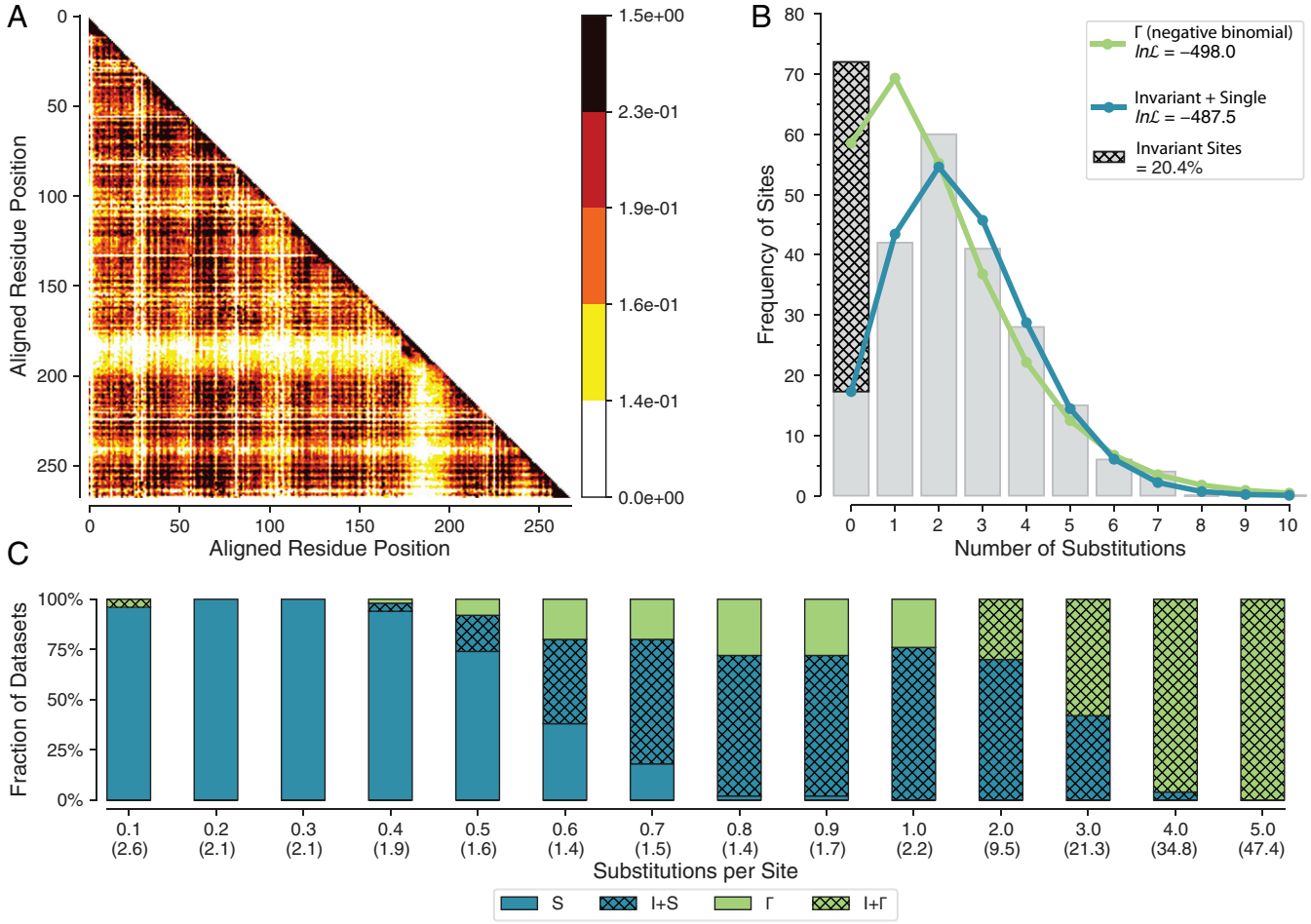

Fig. 1. Impact of epistasis on models of rates of neutral molecular evolution. (*A*) Coupling coefficients between 268 sites in the PF00001 family of protein domains. For each pair of sites, the Frobenius norm of coupling coefficients for all residue combinations is shown with darker, hotter colors representing a stronger overall coupling between a pair of sites. (*B*) Distribution of the number of substitutions for sites in one dataset of PF00001 when the expectation was two substitutions per site. The I + S model (blue) has a higher log likelihood than a Γ model (green). In the I + S model, the "invariant" class contains ∼20% of the sites (black hatched). (*C*) The fraction of PF00001 datasets in which a given model fits the best is determined using AIC_c (6, 7). The mean difference in AIC_c between the Γ model and the best-fitting model is shown in parentheses below the *x* axis. Each bar corresponds to the analysis of datasets with a given evolutionary divergence, which is in units of the expected number of substitutions per site. Each evolutionary divergence analysis is composed of 50 datasets generated from the first 10,000 generations of simulated data provided by de la Paz et al. (2), with the first dataset sampled after sequence Hamiltonians reached an equilibrium state (figure S1 in ref. 2) at generation 2,000. Subsequent datasets were sampled with the starting generations increasing by 50 (i.e., step size = 50 generations). I + S and I + Γ models were tested through the corresponding zero-inflated Poisson and negative binomial models, respectively. The unmutated sites during simulation were not considered invariant, as they were not subjected to evolutionary pressures.

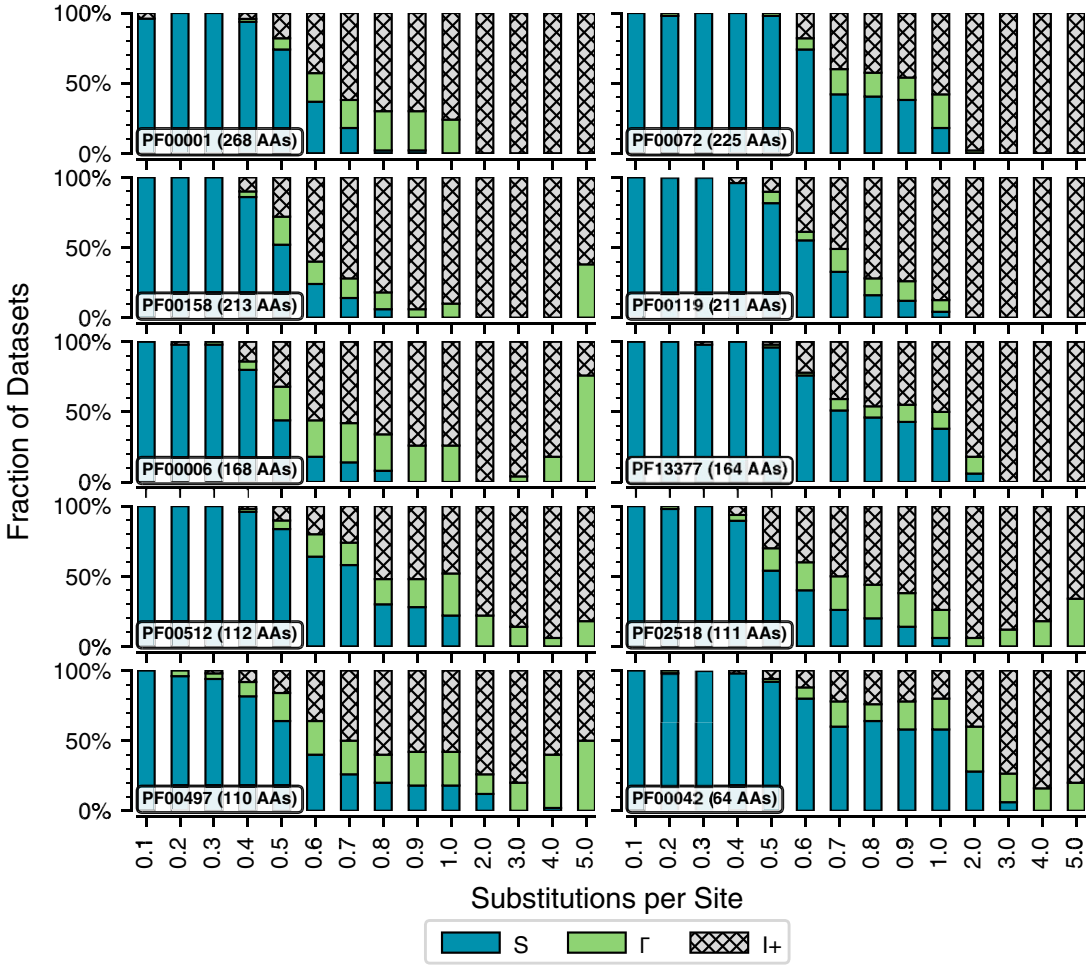

Fig. 2. Best-fitting models by sequence divergence for all protein families. The fraction of datasets where a single-rate model (S), variable-rate model (Γ), and models with an invariant class ("I+"; I + S and I + Γ) best fit the observed SFS (e.g., Fig. 1*B*). The numbers of amino acid (AA) sites simulated and analyzed in each protein family are shown.

1 T. Uzzell, K. W. Corbin, Fitting discrete probability distributions to evolutionary events. *Science* **172**, 1089–1096 (1971).

2 J. A. de la Paz, C. M. Nartey, M. Yuvaraj, F. Morcos, Epistatic contributions promote the unification of incompatible models of neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5873–5882 (2020).

3 H. Luz, M. Vingron, Family specific rates of protein evolution. *Bioinformatics* **22**, 1166–1171 (2006).

4 E. Zuckerkandl, L. Pauling, "Evolutionary divergence and convergence in proteins" in *Evolving Genes and Proteins*, V. Bryson, H. J. Vogel, Eds. (Academic Press, 1965), pp. 97–166.

5 X. Gu, Y. X. Fu, W. H. Li, Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12**, 546–557 (1995).

6 J. E. Cavanaugh, Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probab. Lett.* **33**, 201–208 (1997).

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