Gene Families, Epistasis and the Amino Acid Preferences of Protein Homologs

Evandro Ferrada

Center for Genomics and Bioinformatics, Faculty of Science, Universidad Mayor, Santiago, Chile.

Evolutionary Bioinformatics
Volume 15: 1–3

DOI: 10.1177/1176934319870485

ABSTRACT: In order to preserve structure and function, proteins tend to preferentially conserve amino acids at particular sites along the sequence. Because mutations can affect structure and function, the question arises whether the preference of a protein site for a particular amino acid varies between protein homologs, and to what extent that variation depends on sequence divergence. Answering these questions can help in the development of models of sequence evolution, as well as provide insights on the dependence of the fitness effects of mutations on the genetic background of sequences, a phenomenon known as epistasis. Here, I comment on recent computational work providing a systematic analysis of the extent to which the amino acid preferences of proteins depend on the background mutations of protein homologs.

KEYWORDS: Gene families, site-specific amino acid preferences, protein biophysics, mutational trajectories, genetic background

RECEIVED: July 19, 2019. **ACCEPTED:** July 27, 2019.

TYPE: Commentary

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Universidad Mayor (PEP I-2018041).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Evandro Ferrada, Center for Genomics and Bioinformatics, Faculty of Science, Universidad Mayor, Camino La Pirámide 5750, Huechuraba, 8580745 Santiago, Chile. Email: evandro.ferrada@mayor.cl

COMMENT ON: Ferrada E. The site-specific amino acid preferences of homologous proteins depend on sequence divergence. *Genome Biol Evol.* 2019;11(1):121-135. doi:10.1093/gbe/evy261. PubMed PMID: 30496400. PubMed Central PMCID: PMC6326188. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6326188/.

Homologous genes evolve from common ancestors by processes such as speciation and duplication. As a result of imperfect replication genes accumulate mutations in a manner that is generally proportional to the time since their divergence. In the case of most genes, mutations also accumulate under the main constraint of preserving the gene's essential function and structure. Today, systematic classifications of homologous genes result on the order of 17 000 families, with a vast variation on gene length and the number of known genes per family.[1] Classifications of gene families are based on probabilistic models that capture mutational patterns characteristic of a family and are obtained from the comparison of genes through multiple sequence alignments.

The relation between patterns of mutation observed in multiple sequence alignments and the structure/function in a family of homologous genes is complex. In general, two types of forces contribute to these patterns. One is a stochastic component encompassing the contingencies of the evolutionary history of genes, such as speciation and duplication events, expansion and contractions of the number of genes in a particular family due to selection and adaptation, mutations in other genes, changes in gene expression, as well as changes that result from adaptation to new and fluctuating environments.

A second, more deterministic than stochastic component, encompasses intrinsic constraints to the gene family, which are associated with the biophysics of molecular structure and function. Structure and function are not independent properties,[2] and their degree of association arguably varies depending on the type of structure and the type of function. Furthermore, multiple functions can evolve and coexist as part of a single gene,[3] or different genes in a single family.[4] As a result of the contribution of these forces, mutational patterns in a family are complex, idiosyncratic to the family and, in particular, to subgroups of genes of recent common ancestry. Consequently, mutations observed across an entire family reflect conserved correlations mainly associated with thermostability constraints necessary to preserve structure.

Understanding the origin and variation of mutational patterns in families of homologous genes is essential to molecular evolution. One reason is that the rate at which amino acids change at a particular site varies in a manner that is inversely proportional to the amino acid conservation. Amino acid conservation at a site is often summarized by a vector that expresses the probability of a protein site to be occupied by any of the 20 amino acids, or *site-specific amino acid preferences* (SSAP) (Figure 1). Because protein sites evolve at rates proportional to their SSAP, this information has been essential to inform amino acid substitution models for phylogenetics.[6]

Traditionally, SSAP used to inform substitution models have been inferred directly from the multiple sequence alignments of protein families. Recently, however, high-throughput sequencing methods allow to evaluate simultaneously the functional performance of all single nucleotide variants of a gene.[7] The functional performance of all variants can be transformed to obtain amino acid preferences per site, enabling the estimation of SSAP profiles specific to an individual sequence of a protein family.[8]

Compared with traditional SSAP profiles derived from the multiple sequence alignment of a family, the SSAP profiles derived empirically for individual genes were shown to considerably improve the phylogenetic fit to sequence data.[8] However, a debated question arose as to what extent the SSAP profiles of individual genes of the same family are sensitive to the mutation background of each sequence.[9-12] In other words, are the

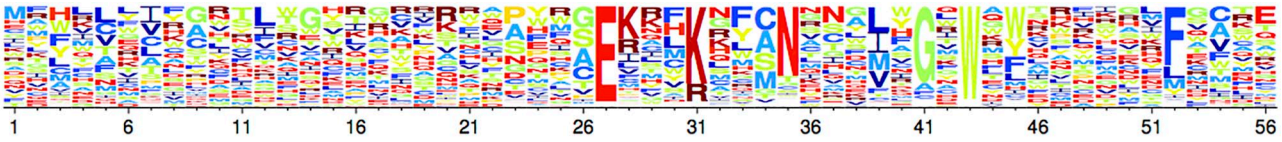

Figure 1. The site-specific amino acid preference profile of the GB1 protein.
The SSAP of all sites in a protein can be summarized as a profile. The profile represents the SSAP for each site of the 56 amino acid long domain B1 of Protein G. Data were obtained from Olson et al.[5] SSAP indicates site-specific amino acid preference.

SSAP profiles of homologous proteins expected to vary as their sequences diverge from a common ancestor, and if so, how high would this variation be? Answering this question can help clarify the limitations of using SSAP profiles as amino acid substitution models for phylogenetics. In addition, it can shed light on understanding the impact of background mutations on protein thermostability and function, a phenomenon generally known as intragenic epistasis, which remains as one of the main challenges in predicting the fitness effects of mutations.[13]

The SSAP of Protein Homologs Depend on Sequence Divergence

In order to systematically explore the dependence of SSAP profiles of protein homologs on sequence divergence, we recently implemented a computational procedure to estimate the SSAP profile of a protein structure based on changes in thermodynamic stability caused by single mutations.[14] The main advantages of our procedure, compared with previous simulations and empirical approaches, is that it allowed us to evaluate changes in SSAP that result from the same molecular trait (ie thermodynamic stability), on a large sample of structures, and at different sequence distances. Our approach matched correlations between SSAP profiles observed in experimental replicates; and it largely recapitulated the SSAP profile of the domain B1 of the Protein G (GB1), studied previously.[5]

We used our computational procedure to systematically explore differences in the SSAP of pairs of homologous proteins with known, high-quality crystal structures. Our analyses revealed a monotonic increase in the difference of SSAP as a function of sequence divergence. Comparison between SSAP generally supported conclusions from previous mutagenesis studies of closely related homologs, but also suggested that the SSAP of a significant fraction of sites is impacted by sequence divergence (ie background mutations), with divergent homologs reaching up to 30% of sites with significant differences. Notably, these observations hold under 3 different biophysical models of the effect of thermodynamic stability on fitness, and for pairs of homologs of diverse size, molecular functions, and structural classes.

Two reasons suggest that we are most likely underestimating differences in SSAP. First, our analyses were conservative. Second, and most importantly, other factors, such as selection for function, or the presence of insertions and deletions, can contribute to changes in SSAP. We observed evidence for this second effect when contrasting the SSAP of GB1 obtained experimentally, with the profile derived computationally from a crystal structure bound to its ligand. We found that at least 5% of the differences detected were mutations at positions directly involved in the binding of GB1's ligand.

Similarly, our analyses found support from recent simulation and empirical studies revealing substantial epistasis arising from the mutational background of divergent homologs.[9,15] Lunzer et al,[16] for instance, studied a pair of bacterial homologs of the enzyme isopropylmalate dehydrogenase (IMDH), which are at a sequence distance of 46% (168/365). They introduced each of the 168 individual variants of the *Pseudomonas aeruginosa* into the *Escherichia coli* homolog. Then they studied the catalytic performance of each of the 168 individual variants and showed that 18% (64/365) of sites were either advantageous or deleterious with respect to the neutral expectation. These differences are slightly above those predicted by our procedure. Another recent study revealed that approximately 80% of amino acid substitutions observed in ancestrally reconstructed sequences of the chaperone HSP90, spanning up to 30% in sequence divergence, are deleterious when introduced into the genetic background of the extant HSP90 sequence of *Saccharomyces cerevisiae*.[17] These studies support our findings based on thermodynamic stability and also suggest that differences in SSAP reveal widespread intragenic epistasis.

A Mechanism for the Cumulative Changes in SSAP

What are the biophysical mechanisms behind the differences in SSAP between protein homologs? Classic comparative studies established that homologs can accumulate substantial structural deviations. Chothia and Lesk showed that deviations between structures depend exponentially on sequence distance, with homologs at sequence identities of 30% reaching on average root-mean squared deviations of 2.0Å.[18] Such deviations should affect the structural environment of equivalent residues in homologous structures in a cumulative manner. The structural context of protein sites is known to have an impact on the evolution of protein sequences. It is well known, for instance, that the solvent accessibility of residues (RSA) is an important determinant of the evolutionary rate of protein sites.[19] More recently, however, it was found that compared with RSA, residue packing, measured as the average normalized number of contacts per residue, can explain a larger fraction of the variance of the evolutionary rate of sites.[20]

To study the effect of changes in residue environment on the SSAP, we performed analyses of the atomic context of sites that were either substituted or conserved between pairs of structures in our dataset. On the one hand, substituted sites rewire on average 30% to 40% of their surrounding contacts, and this fraction is relatively independent of the sequence distance between the homologs under comparison. On the other hand, sites that preserve the same amino acid accumulate changes monotonically and approximately linearly, such that at a sequence distance of approximately 70%, differences between conserved versus substituted sites are indistinguishable. Our results resonate with a quantitative model of the effect of a residue's structural context on its own evolutionary rate, showing that the rate of evolution of a site depends linearly on the local mutational stress experience at the site.[21] Taken together, these analyses suggest that changes in residue packing caused by structural deviations of divergent sequences are a key determinant of the differences observed between the amino acid preference of protein homologs.

Several questions remain to be answered. For instance, it is still unclear how differences in SSAP, or factors specific to a protein family, would impact the performance of substitution models for phylogenetics. An experimental study of a pair of closely related homologs showed that differences in SSAP of 3% to 15% can impact the performance of SSAP profiles as amino acid substitution models for phylogenetics.[22] Another important challenge is to integrate quantitative models of thermostability with constraints on structure and function. Some promising recent advances in this direction[23,24] might help in better understanding the evolution of function, as well as the cooperative effect of background mutation on fitness.

Author Contributions

EF wrote the manuscript.

ORCID iD

Evandro Ferrada https://orcid.org/0000-0003-3242-1726

REFERENCES

1. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. *Nucleic Acids Res*. 2004;32:D138-D141.
2. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new functions trade off. *PLoS Computat Biol*. 2008;4:e1000002.
3. Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. *Trends Genet*. 2003;19:415-417.
4. Ferrada E, Wagner A. Evolutionary innovations and the organization of protein functions in genotype space. *PLoS ONE*. 2010;5:e14172.
5. Olson CA, Wu NC, Sun R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol*. 2014;24:2643-2651.
6. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molec Biol Evol*. 2004;21:1095-1109.
7. Fowler DM, Araya CL, Fleishman SJ, et al. High-resolution mapping of protein sequence-function relationships. *Nat Methods*. 2010;7:741-746.
8. Bloom JD. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Molec Biol Evol*. 2014;31:1956-1978.
9. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. *Proc Nat Acad Sci*. 2012;109:E1352-E1359.
10. Ashenberg O, Gong I, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. *Proc Nat Acad Sci*. 2013;110:21071-21076.
11. Pollock DD, Goldstein RA. Strong evidence for protein epistasis, weak evidence against it. *Proc Nat Acad Sci*. 2014;111:E1450-E1450.
12. Risso VA, Manssour-Triedo F, Delgado-Delgado A, et al. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Molec Biol Evol*. 2014;32:440-455.
13. Starr TN, Thornton JW. Epistasis in protein evolution. *Protein Sci*. 2016;25:1204-1218.
14. Ferrada E. The site-specific amino acid preferences of homologous proteins depend on sequence divergence. *Genome Biol Evol*. 2019;11:121-135.
15. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. *Proc Nat Acad Sci*. 2015;112:E3226-E3235.
16. Lunzer M, Golding GB, Dean AM. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet*. 2010;6:e1001162.
17. Starr TN, Flynn JM, Mishra P, Bolon DNA, Thornton JW. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc Nat Acad Sci*. 2018;115:4453-4458.
18. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*. 1986;5:823-826.
19. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Molec Biol Evol*. 2009;26:2387-2395.
20. Yeh SW, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. *Molec Biol Evol*. 2013;31:135-139.
21. Huang T, del Valle Marcos ML, Hwang JK, Echave J. A mechanistic stress model of protein evolution accounts for site-specific evolutionary rates and their relationship with packing density and flexibility. *BMC Evol Biol*. 2014;14:78.
22. Doud MB, Ashenberg O, Bloom JD. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Molec Biol Evol*. 2015;32:2944-2960.
23. Echave J. Beyond stability constraints: a biophysical model of enzyme evolution with selection on stability and activity. *Molec Biol Evol*. 2018;36:613-620.
24. Otwinowski J. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Molec Biol Evol*. 2018;35:2345-2354.