The role of epistasis in protein evolution

ARISING FROM M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame & F. A. Kondrashov *Nature* **490,** 535–538 (2012)

An important question in molecular evolution is whether an amino acid that occurs at a given site makes an independent contribution to fitness, or whether its contribution depends on the state of other sites in the organism's genome, a phenomenon known as epistasis[1–5]. Breen and colleagues recently argued[6] that epistasis must be "pervasive throughout protein evolution" because the observed ratio between the per-site rates of non-synonymous and synonymous substitutions (dN/dS)[7] is much lower than would be expected in the absence of epistasis. However, when calculating the expected dN/dS ratio in the absence of epistasis, Breen *et al.*[6] assumed that all amino acids observed at a given position in a protein alignment have equal fitness. Here, we relax this unrealistic assumption and show that any dN/dS value can in principle be achieved at a site, without epistasis; furthermore, for all nuclear and chloroplast genes in the Breen *et al.* data set, we show that the observed dN/dS values and the observed patterns of amino-acid diversity at each site are jointly consistent with a non-epistatic model of protein evolution.

For a variety of proteins under purifying selection, Breen *et al.*[6] constructed alignments and recorded the amino acids observed at each position; these observed amino acids were deemed "acceptable" with respect to natural selection. They then assumed that substitutions occur at neutral rates among the acceptable amino acids in order to calculate, for each protein, an expected value for dN/dS in the absence of epistasis. Because their empirical observations of dN/dS were much lower than these expected values, Breen *et al.*[6] concluded that epistasis must be extremely prevalent.

The flaw in this reasoning is that Breen *et al.*[6] considered only a single class of fitness assignments, so that all amino acids observed at a site were assumed equally fit. A more realistic assumption is that some amino acids observed at a site are more fit than others[8,9].

To illustrate the principle that low dN/dS can arise without epistasis, we considered a non-epistatic model in which, among the acceptable amino acids at a given site, one of these is preferable to the rest. We performed the following experiment: in a hypothetical protein of length 300 amino acids, for each position we randomly designated eight amino acids as acceptable (the average number of acceptable amino acids reported by Breen *et al.*[6]), but gave one of these a selective advantage over the rest. We then calculated the equilibrium dN/dS (ref. 10) for this protein as a function of the selective advantage of the preferred amino acid, $2Ns$ (Fig. 1). Whereas dN/dS is high for the case $2Ns = 0$, corresponding to the Breen *et al.*[6] assumption, dN/dS is much lower for larger $2Ns$. Thus, a large range of dN/dS values are consistent with non-epistatic models of protein evolution.

Although non-epistatic models can in principle produce low dN/dS values (Fig. 1), can such a model account for the Breen *et al.*[6] data? To answer this question, we considered a more general non-epistatic model that assigns to each amino acid at a site a different fitness. For each gene in the Breen *et al.*[6] data set, we assigned fitnesses at each site in such a way that the resulting equilibrium distribution of amino acids under our model precisely matches the amino-acid frequencies observed for that site[11]. Furthermore, as a result of these fitness assignments, the asymptotic mean pairwise sequence divergence under our model necessarily matches the mean pairwise divergence observed in the data (Table 1 and Methods).

Using this model, for each gene in the Breen *et al.*[6] data set we repeatedly simulated the evolution of a pair of sequences from their common ancestor and computed dN/dS. For the 13 mitochondrial genes, the average simulated dN/dS values, although substantially lower than the Breen *et al.*[6] expectations, are still greater than the empirically observed values (Table 1). However, for the three nuclear and chloroplast genes in the Breen *et al.*[6] data set, the average dN/dS values under our non-epistatic model are comparable to or even lower than the empirical dN/dS values Breen and colleagues reported. Thus, the dN/dS values observed in these genes need not be attributed to epistasis, but rather can be explained by the more parsimonious assumption that the various amino acids observed at a site have different fitnesses.

It is important to note that the effects of natural selection and phylogeny are confounded in the amino-acid frequencies observed at each site, and therefore in our fitness estimates. Although methods exist to disentangle these effects when the phylogeny is small and known[12,13], there is no well-accepted phylogeny for the vast range of taxa studied by Breen *et al*[6]. Nonetheless, whatever the true phylogeny may be, under the standard assumption that molecular evolution can be modelled as an equilibrium Markov chain (see, for example, ref. 14, as used by Breen *et al.*[6]) our fitness estimates are maximum likelihood. Relaxing this assumption, or allowing more complex models (for example, allowing fitnesses or population sizes to vary across time or clade), would make it only more difficult to reject the non-epistatic null hypothesis.

In summary, Breen *et al.*[6] provide no direct evidence of epistasis, nor do they reject the full space of non-epistatic models. They have analysed only three non-mitochondrial genes, whose evolutionary patterns, we have shown, can be explained without epistasis. Although Breen *et al.*[6] contend that epistasis is the primary factor in all of molecular evolution, further work is needed to substantiate this claim.

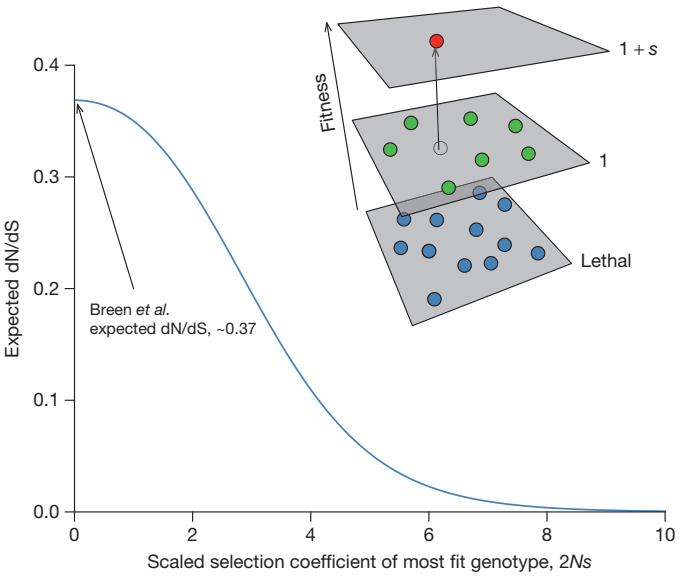

Figure 1 | Non-epistatic models of protein evolution can produce low dN/dS values. Expected dN/dS as a function of $2Ns$ for a hypothetical protein of length 300, in which eight acceptable amino acids are chosen at random for each position and one of these amino acids at random is assigned a selective advantage of size $2Ns$. The remaining 12 amino acids are lethal. The Breen *et al.*[6] expectation for dN/dS in the absence of epistasis corresponds to $2Ns = 0$.

Table 1 | Observed and expected dN/dS values

Gene	Breen *et al.*[6] expected dN/dS	Our average simulated dN/dS	Breen *et al.*[6] empirical dN/dS	Our equilibrium mean pairwise divergence	Breen *et al.*[6] empirical pairwise divergence
Mitochondrial					
ATP6	0.44	0.215	0.056	0.332	0.332
ATP8	0.56	0.624	0.224	0.615	0.615
COX1	0.28	0.078	0.015	0.188	0.188
COX2	0.43	0.140	0.025	0.348	0.348
COX3	0.32	0.144	0.036	0.290	0.290
CYTB	0.51	0.117	0.039	0.242	0.242
ND1	0.39	0.208	0.040	0.383	0.383
ND2	0.51	0.262	0.067	0.398	0.398
ND3	0.49	0.242	0.069	0.379	0.379
ND4	0.42	0.239	0.045	0.433	0.433
ND4L	0.49	0.369	0.076	0.502	0.502
ND5	0.32	0.211	0.057	0.407	0.407
ND6	0.42	0.397	0.073	0.554	0.554
Nuclear					
EEF1A1	0.11	0.031	0.020	0.080	0.080
H3.2	0.14	0.014	0.037	0.019	0.019
Chloroplast					
rbcL	0.40	0.024	0.072	0.056	0.056

Comparison of expected dN/dS values and mean pairwise divergence with the empirical values for each gene in the Breen *et al.*[6] data set. The Breen *et al.* expected dN/dS is based on the assumption that all amino acids observed at a given site are neutral relative to each other. Our expected dN/dS is based on the assumption that the various amino acids observed at a site have different fitnesses.

METHODS

We assume that each codon evolves according to an independent Markov chain, the rate matrix of which is determined by the scaled selection coefficient assigned to each amino acid[15]. The equilibrium frequency of each amino acid is then proportional to $u_i e^{2Ns(i)}$ (ref. 15), in which u_i is the number of codons that code for amino acid i, and $2Ns(i)$ is its scaled selection coefficient. After assigning site-specific fitnesses to amino acids, 1,000 simulations were conducted for each protein, as follows. For each site represented in at least half the sequences from the Breen *et al.*[6] alignment, an ancestral codon was drawn from the equilibrium distribution of our Markov chain. Two copies of this ancestral sequence were then evolved independently until dS = 0.25, which is within the range of dS = 0.05 to 0.5 used by Breen *et al*[6]. We then estimated dN/dS for each pair using PAML[14], again following the procedure of Breen *et al.* Mean pairwise divergence (Table 1) was calculated using the formula $\frac{1}{L}\left(\sum_{j=1}^{L}\left(1-\sum_{i=1}^{20} f_{i,j}^2\right)\right)$, in which $f_{i,j}$ denotes the frequency of amino acid i at site j, and L the number of majority non-gapped sites in the protein. All computer code is available on request.

David M. McCandlish[1], Etienne Rajon[1], Premal Shah[1], Yang Ding[1] & Joshua B. Plotkin[1]

[1]Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
e-mail: jplotkin@sas.upenn.edu

Received 20 December 2012; accepted 18 April 2013.

1. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
2. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
3. Kryazhimskiy, S., Dushoff, J., Brazykin, G. A. & Plotkin, J. B. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet.* **7,** e1001301 (2011).
4. Salverda, M. L. M. *et al.* Initial mutations direct alternative pathways of protein evolution. *PLoS Genet.* **7,** e1001321 (2011).
5. Hansen, T. F. & Wagner, G. P. Modeling genetic architecture: a multilinear theory of gene interaction. *Theor. Popul. Biol.* **59,** 61–86 (2001).
6. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490,** 535–538 (2012).
7. Li, W. H. *Molecular Evolution* (Sinauer, 1997).
8. da Silva, J. Site-specific amino acid frequency, fitness and the mutational landscape model of adaptation in HIV-1. *Genetics* **174,** 1689–1694 (2006).
9. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7,** 741–746 (2010).
10. Yang, Z. & Nielsen, R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46,** 409–418 (1998).
11. Choi, S. C., Redelings, B. D. & Thorne, J. L. Basing population genetic inferences and models of molecular evolution upon desired stationary distributions of DNA or protein sequences. *Phil. Trans. R. Soc. B* **363,** 3931–3939 (2008).
12. Rodrigue, N., Phillippe, H. & Lartillot, N. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl Acad. Sci. USA* **107,** 4629–4634 (2010).
13. Tamuri, A. U., dos Reis, M. & Goldstein, R. A. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* **190,** 1101–1115 (2012).
14. Yang, Z. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).
15. Halpern, A. L. & Bruno, W. J. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.* **15,** 910–917 (1998).

Competing Financial Interests Declared none.

doi:10.1038/nature12219

Breen *et al.* reply

REPLYING TO D. M. McCandlish, E. Rajon, P. Shah, Y. Ding & J. B. Plotkin *Nature* **497,** http://dx.doi.org/10.1038/nature12219 (2013)

Understanding fitness landscapes, a conceptual depiction of the genotype-to-phenotype relationship, is crucial to many areas of biology. Two aspects of fitness landscapes are the focus of contemporary studies of molecular evolution. First, the local shape of the fitness landscape defined by the contribution of individual alleles to fitness that is independent of all genetic interactions. Second, the global, multidimensional fitness landscape[1] shape determined by how interactions between alleles at different loci change each other's fitness impact, or epistasis. In explaining the high amino-acid usage (u), we focused on the global shape of the fitness landscape[2], ignoring the perturbations at individual sites[3].

u is overly sensitive to the presence of rare amino acids[2], including fixed slightly deleterious amino acid—a factor we did not consider. McCandlish *et al.*[3] apply a non-epistatic model with fitness landscape

ruggedness on the local scale of individual sites. As a null model for testing the hypothesis of an epistasis-free fitness landscape, it directly confirms our conclusion that epistasis affects most amino-acid substitutions for 13 out of 16 genes considered[3]. The other three genes are extremely conservative with low density and frequency of emerged amino-acid states in the multiple sequence alignment. In the non-epistatic model[3] such amino-acid states seem to be substantially deleterious, leading to the low predicted dN/dS values in these three genes, with the largest effect in *rbcL*.

In the absence of epistasis, strong selection against non-optimal states markedly decreases the equilibrium sequence divergence[3,4] and the expected time to reach the equilibrium divergence[4]. The model[3] for *rbcL* simulates an equilibrium sequence divergence of ~0.06, which must be independent of phylogenetic distance beyond closely related clades[4]. Both of these predictions are easily falsified. Orthologous *rbcL* sequence divergence shows no sign of reaching a true equilibrium even between phylogeneticaly distant clades (Fig. 1), whereas a BLAST search reveals that sequence divergence between *Arabidopsis thaliana* and cyanobacterial orthologues reaches values greater than 0.16.

Generally, the non-epistatic model has a trade-off between the strength of selection against suboptimal alleles and the expected sequence divergence, which rapidly reaches its equilibrium value[4,5]. For extremely conservative genes, such as the selected three non-mitochondrial genes considered[2], the non-episatatic model can give the appearance of avoiding this trade-off, which breaks down when the long-term evolutionary predictions of the model are considered in detail (Fig. 1).

Two aspects of protein evolution are revealed by sequence similarity searches. First, protein sequence divergence occurs slowly, slower than neutral divergence. Second, sequence divergence is proportional to phylogenetic distance and is usually substantial for sequences from distantly related species. Non-epistatic models[3,4] that consider only local fitness landscape ruggedness are inconsistent with both of these basic and universal features of protein evolution. By contrast, our claim that epistasis—the global, multidimensional shape of the fitness landscape—is the primary factor of protein evolution explains the high amino-acid usage[2] and how slow long-term sequence divergence leads to highly dissimilar sequences[5]. Models that take into account both local and global aspects of fitness landscapes could lead to better quantification of factors shaping molecular evolution, although their development may be hampered by inherent complexity of multi-dimensional fitness landscapes[1] and subtle local confounding factors[6–8].

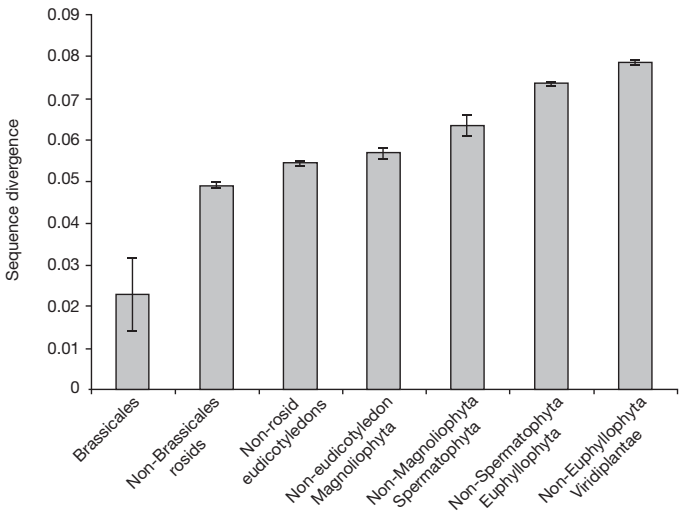

Figure 1 | Sequence divergence as a function of phylogenetic distance. Average sequence divergence for pairwise comparisons of *A. thaliana* RbcL protein sequence and other orthologous sequences used in ref. 2.

Michael S. Breen[1], Carsten Kemena[2,3], Peter K. Vlasov[2,3], Cedric Notredame[2,3] & Fyodor A. Kondrashov[2,3,4]

[1]Department of Medicine, University of California San Diego, La Jolla, California 92093, USA.
[2]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG) 88 Dr. Aiguader, 08003 Barcelona, Spain.
[3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.
[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, 08010 Barcelona, Spain.
e-mail: fyodor.kondrashov@crg.es

1. Kondrashov, F. A. & Kondrashov, A. S. Multidimensional epistasis and the disadvantage of sex. *Proc. Natl Acad. Sci. USA* **98,** 12089–12092 (2001).
2. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490,** 535–538 (2012).
3. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. *Nature* **497,** E1–E2 (2013).
4. Kondrashov, A. S., Povolotskaya, I. S., Ivankov, D. N. & Kondrashov, F. A. Rate of sequence divergence under constant selection. *Biol. Direct* **5,** 5 (2010).
5. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
6. McVean, G. & Charlesworth, B. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. *Genet. Res.* **74,** 145–158 (1999).
7. Kondrashov, F. A., Ogurtsov, A. Y. & Kondrashov, A. S. Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J. Theor. Biol.* **240,** 616–626 (2006).
8. Kimura, M. The role of compensatory neutral mutations in molecular evolution. *J. Genet.* **64,** 7–19 (1985).

doi:10.1038/nature12220