

In vitro evolution goes deep

Alan M. Moses^a and Alan R. Davidson^{b,c,1}

Departments of ^aCell and Systems Biology, ^bMolecular Genetics, and ^cBiochemistry, University of Toronto, Toronto, ON, Canada M5S 1A8

Approximately 20 years ago when one of the authors of this commentary was a postdoctoral fellow in Bob Sauer's laboratory at Massachusetts Institute of Technology, an extremely energetic individual (the author in question was not one of these) could run three sequencing gels in one day, and thus potentially obtain 72 sequence reads of a few hundred bases each (in practice, unfortunately, one never got good sequence from every reaction). This was the cutting edge of high-throughput DNA sequencing technology at the time (we used multichannel pipettes and 96-well plates), and this approach combined with sequence randomization by cassette mutagenesis yielded what seemed like a huge amount of data pertaining to the tolerance of proteins to amino acid substitutions (1). These studies and other site-directed mutagenesis experiments (2, 3) led to the surprising (at the time) conclusion that proteins could tolerate extensive sequence variation and still function in vivo. A paper by Hietpas et al. (4) in PNAS exposes the feebleness of previous efforts in this area by using "deep sequencing" to obtain millions of mutant sequences and thereby empirically determine the fitness of all possible single base mutants in a defined region of a gene. This work provides a definitive picture of the tolerance of a protein to single amino acid substitutions within one region.

The method developed by Hietpas et al. (4) has been dubbed extremely methodical and parallel investigation of randomized individual codon (EMPIRIC) fitness. Each codon within a particular region of a plasmid-expressed gene is randomized one codon at a time to create a mutant library in which every possible codon is represented at every position. A pool of strains carrying all these mutant plasmids is then combined into one flask and grown under conditions in which activity of the mutated gene is required for cell survival. The frequency of each codon at each position is followed over a time course using Illumina technology to sequence the mutagenized regions of hundreds of thousands of plasmids. Codons encoding amino acids that are more deleterious to protein function are more quickly eliminated from the mutant pool, whereas those encoding more tolerated amino acid changes persist longer. Thus, a direct measure of the fitness of each codon is produced. There are several significant

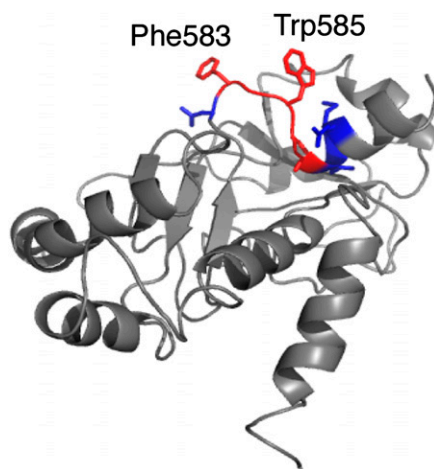


Fig. 1. Mutated region of yeast Hsp90. The nine residues investigated by Hietpas et al. (4) are shown. The positions that tolerated few substitutions are shown in red, and the others are shown in blue. This figure was generated from Protein Data Bank [PDB ID code 2CG9 (residues A445–A677)] using Pymol (<http://www.pymol.org/>).

advantages of this methodology. First, the fitness of all possible single mutations within a given region is assayed simultaneously under the same experimental conditions. Second, mutants contain base changes at only one codon, so compensating or exacerbating effects that could be caused by multiple substitutions are eliminated. Finally, because fitness is precisely gauged based on the change in frequency of each codon over time, variations in the frequency of codons in the starting mutant pool do not affect the result. In previous mutagenesis studies, skewed codon frequencies in the starting libraries could significantly bias the results. The quantitative fitness measurements contrast with previous studies, where fitness was usually assessed using an all or nothing threshold phenotypic value (e.g., whether cells bearing the mutant could form a colony after overnight growth). To demonstrate the utility of the EMPIRIC approach, Hietpas et al. (4) focused on a nine-amino acid region of yeast Hsp90. Their strategy allowed the parallel analysis of all 180 possible amino acid substitutions and over 500 different codon variants, a task that would never be attempted using traditional approaches. Mutations were constructed on an Hsp90-expressing plasmid, and the mutant pool was tested in a yeast strain carrying a temperature-sensitive mutation in its endogenous

Hsp90 gene. Once this strain was transformed with the mutant pool, the fitness of each mutant plasmid was measured by assessing its prevalence at increasing time periods after shifting to the nonpermissive temperature.

Although a very conserved region was tested in the study by Hietpas et al. (4) and Hsp90 is an unusually conserved essential protein, ~15% of all amino acid substitutions were found to be essentially neutral with respect to fitness (i.e., their fitness value was within the range observed for base changes synonymous to WT). The authors note that this is consistent with the nearly neutral theory of evolution (5). This theory was developed to explain the preponderance of genetic differences that were observed between species and individuals (6). Rather than these evolutionary changes being the result of natural selection, Kimura (6) argued that they were the result of genetic drift fixing neutral (or nearly neutral) mutations in finite populations. A common misconception of the neutral theory is that it posits all amino acid positions are free to evolve "neutrally" or in the absence of selection. On the contrary, it is well appreciated that most amino acids are under strong purifying selection, and are therefore preserved over evolution. This was also observed in the study by Hietpas et al. (4), wherein the majority of mutations were highly deleterious. The relative abundance of mutations of different types (i.e., highly deleterious, nearly neutral mutations, beneficial) is of great interest and debate in evolutionary biology. This so-called "distribution of fitness effects" is of both practical and theoretical importance. Previously, the distribution of fitness effects has been studied by introducing mutations into the entire genomes of model organisms and measuring their effects in the laboratory or through population genetic inferences based on the patterns of allele frequency distributions, polymorphism, and substitution (7). Remarkably, the bimodal shape of the distribution of fitness effects observed within this nine-amino acid region of Hsp90 is consistent with

Author contributions: A.M.M. and A.R.D. wrote the paper. The authors declare no conflict of interest.

See companion article on page 7896 in issue 19 of volume 108.

¹To whom correspondence should be addressed. E-mail: alan.davidson@utoronto.ca.

these previous studies, suggesting that the patterns observed by Hietpas et al. (4) might reflect the expectation for mutations over the entire genome.

The empirical fitness effects observed by Hietpas et al. (4) can be directly compared with the patterns of evolution observed in an alignment of Hsp90 homologs from diverse species. A paradoxical result is that most of the amino acid substitutions that were found to be nearly neutral in this experiment were never observed at the corresponding position in the alignment of homologs. If these substitutions were truly neutral, they should have occurred often during the hundreds of millions of years of evolution spanned by species represented in the alignment. The resolution of this paradox is likely that the conditions sampled in this experiment are less stringent than those sampled in nature over millions of years. Interestingly, despite these less stringent conditions, for a contiguous set of residues in the center of the region studied (positions 583–587, Fig. 1), very good agreement between the fitness measurements and the pattern of evolution was seen. These residues are exposed in a loop in the protein structure that may be directly involved in substrate recognition; thus, substitutions at these positions would directly affect biological activity. Substitutions at the other positions tested may only affect function indirectly by reducing protein stability. We speculate that such small reductions in stability would not reduce fitness over the short time span of this experiment but could become a liability in the long term or as the environment varies. Excitingly, the EMPIRIC method can be applied under a wide variety of conditions and genetic backgrounds to allow a direct test of this hypothesis and, in general, assess the rela-

tionship between environmental conditions and fitness.

Because adaptation must involve the accumulation of beneficial mutations and “directed evolution” experiments have shown that the activity of numerous proteins can be vastly increased through rounds of mutagenesis and selection (8), some might be surprised that none of the 171 mutants assayed by Hietpas et al. (4) display a significant increase in fitness.

None of the 171 mutants assayed by Hietpas et al. display a significant increase in fitness.

Although beneficial single amino acid substitutions are likely to be rare, it is possible that beneficial changes might have emerged if a different region of Hsp90 had been tested. Neutral mutations, as were frequently observed here, have been found, both in nature and in the laboratory, to be crucial for opening new adaptive pathways (9, 10). Such mutations can increase protein stability, and thus increase the tolerance for subsequent beneficial but destabilizing substitutions. Alternately, neutral mutations may lead to so-called “standing variation,” differences in function that are not under selective pressure currently but become important when environmental conditions change. Future EMPIRIC-based studies could systematically map beneficial mutations and provide insight into the process of protein adaptation, an area in which relatively few empirical data are available (11).

Aside from its advantages for evolutionary studies, the EMPIRIC method

also provides a rapid method to obtain useful mutants for use in subsequent studies. For example, 56 different amino acid substitutions were found to have significantly reduced fitness that was still considerably higher than the fitness of null mutants. Even positions in this experiment that allowed the fewest substitutions still exhibited partial fitness when many other amino acids were substituted (e.g., 7 different substitutions at position 584 exhibited partial activity). Cells carrying these partially active mutations would likely display “conditional phenotypes,” growing under some conditions and not growing under others. Such mutants can be extremely useful for genetic experiments, including high-throughput applications (12).

Recently, deep-sequencing technologies have been applied to obtain complete genome sequences for populations evolved in the laboratory, either under selection (13–15) or in its absence (16). These studies have allowed unprecedented views of the process of evolution because they allow systematic mapping of variation and direct measurements of the changes in allele frequencies in populations over time. The work of Hietpas et al. (4) provides a demonstration that combining deep sequencing and site-directed mutagenesis can address additional fundamental evolutionary questions. We are confident that further application of the EMPIRIC approach will answer many questions pertaining both to evolution and to protein structure and function.

ACKNOWLEDGMENTS. The research of A.R.D. related to this topic is funded by the Canadian Institutes of Health Research (MOP-13609) and that of A.M.M. is funded by the National Sciences and Engineering Research Council of Canada.

- Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT (1990) Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science* 247:1306–1310.
- Milla ME, Brown BM, Sauer RT (1994) Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nat Struct Biol* 1:518–523.
- Rennell D, Bouvier SE, Hardy LW, Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* 222:67–88.
- Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. *Proc Natl Acad Sci USA* 108:7896–7901.
- Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.
- Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
- Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8:610–618.
- Dougherty MJ, Arnold FH (2009) Directed evolution: New parts and optimized function. *Curr Opin Biotechnol* 20:486–491.
- Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10:866–876.
- Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Comput Biol* 4:e1000002.
- Orr HA (2005) The genetic theory of adaptation: A brief history. *Nat Rev Genet* 6:119–127.
- Li Z, et al. (2011) Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat Biotechnol* 29:361–367.
- Barrick JE, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461:1243–1247.
- Anderson JB, et al. (2010) Determinants of divergent adaptation and Dobzhansky-Muller interaction in experimental yeast populations. *Curr Biol* 20:1383–1388.
- Parts L, et al. (March 21, 2011) Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res*, 10.1101/gr.116731.110.
- Keightley PD, et al. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* 19:1195–1201.