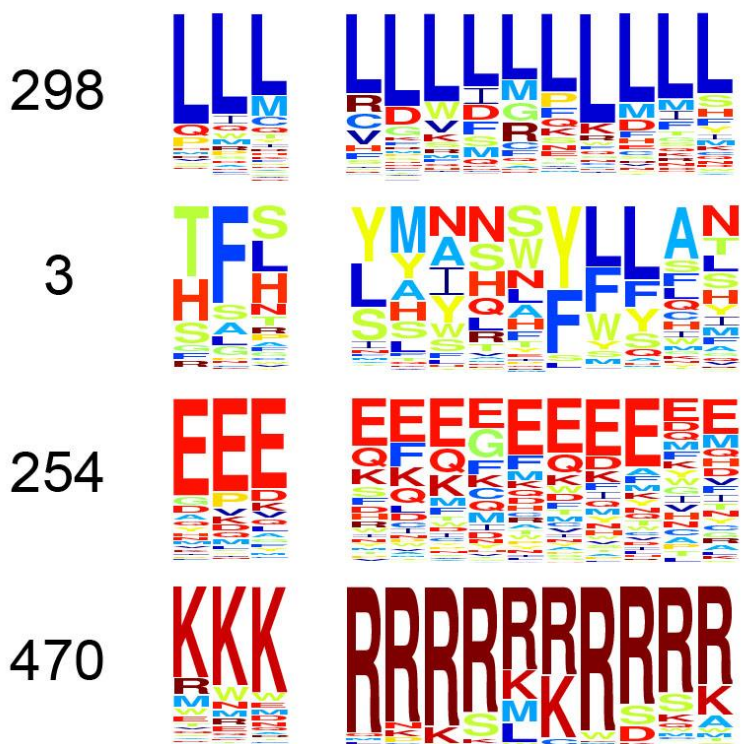


Picky proteins: amino acid preferences are all in the family

September 21, 2015

GE Zentner

Site Biological Replicates



Visual representation of amino acid preferences at selected sites in the H1N1 NP (left) and H3N2 NP (right). The height of a letter indicates the degree of preference for that amino acid.

Image from the publication

Evolution results in changes in protein sequence over time. However, these changes are not random. It has been appreciated for several decades that different sites within a protein evolve under different constraints, such that some sites can tolerate substitution to almost any amino acid while others may only be changed to a few amino acids. These constraints generally arise from interactions within a protein that shape its folding, stability, and function. An important question regarding site-specific amino acid preferences is how conserved they are during evolution. To address this, graduate student Michael Doud and postdoctoral fellow Dr. Orr Ashenberg in the laboratory of Dr. Jesse Bloom (Basic Sciences Division) performed a comprehensive characterization of the tolerance of amino acid positions to substitutions in two closely related influenza nucleoprotein (NP) homologs. They found that site-specific amino acid preferences are

highly conserved between the two proteins, enabling improved modeling of the evolution of other related proteins.

The authors chose to study two NPs from two human influenza strains, H1N1 and H3N2, which are separated by 30 years of evolution and are 94% identical in sequence. To comprehensively assess the site-specific amino acid preferences of each site in these two NPs, the authors used deep mutational scanning. This is a technique in which PCR is used to create a library of plasmids encoding single-codon mutants, which are then incorporated into influenza viruses. Cells are then infected and grown to select functional NP proteins, and high-throughput sequencing is used to quantify the occurrence of each mutation before and after infection, the principle being that more fit NP variants would be overrepresented, while less fit variants would be underrepresented.

Using deep mutational scanning, the authors were able to mutagenize 497/498 sites in each NP (the N-terminal methionine was not mutagenized) to all 19 other amino acids and measure amino-acid preferences at each site. Using these data, they then correlated the preferences between homologs. Strikingly, the correlation between the homologs' preferences was nearly as robust as the correlation between preferences for a single homolog measured in this and a previous study. Analysis of individual sites revealed that, by and large, shifts in site-specific amino acids preferences were relatively small, with a consistently preferred amino acid across multiple biological replicates. Preferred amino acids were often conserved between homologs, but a small number of cases where a strong amino acid preference was different at a given site were observed. Further analysis revealed that some of these sites were evolutionarily variable, and that others were in functionally important regions of the protein, including its RNA-binding sites.

The authors next used their experimentally determined site-specific amino acid preferences to inform models of NP substitution and trace the phylogeny of human influenza. Models trained with the deep mutational scanning data, particularly a combined dataset encompassing data for both homologs, performed much better than a non-site-specific model. Experimentally informed models also performed well in tracing influenza virus evolution in swine and equine hosts, indicating that site-specific amino acid preferences are well-conserved across influenza NP from more distantly related hosts.

This study shows that site-specific amino acid preferences are largely conserved across closely related protein homologs, which has important implications for extrapolating experimental data from one homolog to another. "Our study is important because in biology, we often perform studies of a specific gene but would like to extrapolate our conclusions to closely related homologs of that gene in other species. We therefore wanted to take advantage of new experimental techniques to make

the first measurement of how similar (or different) the effects of all possible mutations are to homologs of the same gene," said Dr. Bloom. "Our study provides a quantitative measure of how much one would expect the effects of mutations to change as a sequence diverges, and is therefore of great value in determining how to best formulate quantitative models of protein evolution."

[Doud MB, Ashenberg O, Bloom JD](#). 2015. Site-specific amino-acid preferences are mostly conserved in two closely related protein homologs. *Mol Biol Evol* [Epub ahead of print]

Funding source: This work was supported by the National Institute of General Medical Sciences of the National Institute of Health (grant number R01 GM102198). M.B.D. was supported by NIH Training Grant T32 AI083203 and a fellowship from the Seattle Chapter of the Achievement Rewards for College Scientists Foundation. O.A. was supported by a PhRMA Postdoctoral Fellowship in Informatics.