SMBE 2017

Open Symposium

POB-387

Theory of measurement for site-specific evolutionary rates in amino-acid sequences

Dariya K. Sydykova [1*], Claus O. Wilke
[1]Department of Integrative Biology, The University of Texas at Austin, Austin TX, United States

Abstract

Many applications require the calculation of site-specific evolutionary rates from alignments of amino-acid sequences. For example, catalytic residues in enzymes and interface regions in protein complexes can be inferred from observed relative rates. While numerous approaches exist to calculate amino-acid rates, however, it is not entirely clear what physical quantities the inferred rates represent and how these rates relate to the underlying fitness landscape of the evolving protein. Further, amino-acid rates can be calculated in the context of different amino-acid exchangeability matrices, such as JTT, LG, or WAG, and again it is not known how the choice of the matrix influences the physical interpretation of the inferred rates. Here, we develop a theory of measurement for site-specific evolutionary rates, but analytically solving the maximum-likelihood equations for rate inference performed on sequences evolved under a mutation–selection model. We demonstrate that the measurement process can only recover the true expected rates of the mutation–selection model if rates are measured relative to a naïve exchangeability matrix, in which all exchangeabilities are equal to one. Rate measurements using other matrices are quantitatively close but not mathematically correct. Our results demonstrate that insights obtained from phylogenetic-tree inference do not necessarily apply to rate inference, and best practices for the former may be deleterious for the latter.

Expanded summary

Different sites in a protein evolve at different rates [1,2]. The heterogeneity in rates within a protein sequences is caused by the interplay of functional and structural constraints [3]. For instance, active sites are generally very conserved [4,5]. The protein core tends to be more conserved than the surface, presumably because mutations in the core are more likely to disturb the protein structure [6,7]. Because the evolutionary rates correspond to structurally and functionally important sites, having a reliable and accurate method for inferring site-wise rate of evolution is essential. Specifically, in most viral populations, proteins evolve very rapidly. In influenza, mutations in one site of a surface protein hemagglutinin allow the virus to escape host antibodies and propagate. Thus, detecting site-wise rates of evolution can be crucial to the efforts of viral surveillance and control.

Many methods to infer site-wise rate have been developed over the years. These methods employ a substitution matrix, which captures exchangeabilities between all pairs of amino acids. The substitution matrices are made by analyzing large protein data sets and even protein sequences specific to an organism or an organelle. However, it remains an open question which substitution matrix is the most suitable for site-wise rate inference. When we measure site-wise rate, rate is defined as a scalar in front of the substitution matrix. The substitution matrix serves as a ruler by which we measure the rate of evolution at a site. Depending on what ruler or substitution matrix we use, the inferred rate changes. We developed a theory that shows the effect of the substitution matrix on the inferred site-wise rate. We demonstrate that only a naïve exchangeability matrix, in which all exchangeabilities are equal to one, can recover the true expected rates. We also demonstrate that inference with the true substitution matrix yields site-wise rate of 1. Finally, we demonstrate that the current best-practice matrices (JTT, WAG, and LG) in phylogenetic inference do not recover correct site-wise rates. Along with the mentioned results, our analytical derivations allow for further insights into models of molecular evolution.

There are two ways to measure the rate of evolution in a protein coding sequence. One by calculating the rate of evolution in codon sequences, and the other by calculating the rate of evolution in amino acid sequences. We recently demonstrated that the two inference methods produce comparable rates [8]; however, our current analytical work can establish a direct mathematical link between the two frameworks. We can directly address the issue with inferring rates with amino acid models from codon sequences. Finally, our calculations allow us to incorporate mutation rates into the inference of site-wise rate. We can test different assumptions about the mutation rates and their effect on the rate of evolution at a site.

References

1. Kimura, M. and Ohta, T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 71:2848-2852

2. Perutz, M. F. et al. (1965). Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.*, 13:669-678.

3. Echave, J. et al. (2016). Causes of evolutionary rate variation among protein sites. *Nature Rev. Genet.*, 17:109-121.

4. Jack, B. R. et al (2016). Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biol.* 14:e1002452.

5. Dean, A. M. et al (2002). The pattern of amino acid replacements in alpha/beta-barrels. *Mol. Biol. Evol.* 19:1846-1864.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3002v1 | CC BY 4.0 Open Access | rec: 2 Jun 2017, publ: 2 Jun 2017

6. Franzosa, E. A. and Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at residue level. *Mol. Biol. Evol.* 26:2387-2395.

7. A. Shahmoradi, D. K. Sydykova, S. J. Spielman, E. L. Jackson, E. T. Dawson, A. G. Meyer, C. O. Wilke (2014). Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *J. Mol. Evol.* 79:130–142.

8. D. K. Sydykova, C. O. Wilke (2017). Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. P*eerJ* 5:e3391.

Disclosure of Interest: None Declared

Keywords: None