

The estimation of relative site variability among aligned homologous protein sequences - (session: Comparative Genomics and Molecular Evolution)

D. S. Horner and G. Pesole

Dipartimento di Fisiologia e Biochimica Generale, Università di Milano

Maximum likelihood-based methods to estimate site by site substitution rate variability in aligned homologous protein sequences rely on the formulation of a phylogenetic tree and generally assume that the patterns of relative variability follow a pre-determined distribution. We present a phylogenetic tree-independent method to estimate the relative variability of individual sites within large datasets of homologous protein sequences. It is based upon two simple assumptions. Firstly that substitutions observed between two closely related sequences are likely, in general, to occur at the most variable sites. Secondly that non-conservative amino acid substitutions tend to occur at more variable sites. Our methodology makes no assumptions regarding the underlying pattern of relative variability between sites.

We have compared, using data simulated under a non-gamma distributed model, the performance of this approach to that of a maximum likelihood method that assumes gamma distributed rates. At low mean rates of evolution our method inferred site by site relative substitution rates more accurately than the maximum likelihood approach in the absence of prior assumptions about the relationships between sequences. Our method does not directly account for the effects of mutational saturation. However, we have incorporated an "ad-hoc" modification that allows the accurate estimation of relative site variability in fast evolving and saturated datasets.