

Abstract: Sequence alignment is one of the most important bioinformatics tools for modern molecular biology. The statistical characterization of gapped alignment scores has been a long-standing problem in sequence alignment research. Using a variant of the directed path in random media model, we investigate the score statistics of global sequence alignment considering the compositional bias of the sequences compared. Such statistics are used to distinguish accidental similarity due to compositional similarity from biologically significant similarity. To accommodate the compositional bias, we introduce an extra parameter indicating the probability for positive matching scores to occur. When is small, a high scoring alignment obviously cannot come from compositional similarity. When is large, the highest scoring point within a global alignment tends to be close to the end of both sequences, in which case we say the system percolates. By applying finite-size scaling theory on percolating probability functions of various sizes (sequence lengths), the critical at infinite size is obtained. For alignment of length t , the fact that the score fluctuation grows as is confirmed upon investigating the scaling form of the alignment score. Using the Kolmogorov-Smirnov statistics test, we show that the random variable χ , if properly scaled, follows the Tracy-Widom distributions: Gaussian orthogonal ensemble for p slightly larger than p_c and Gaussian unitary ensemble for larger p .