Abstract Submitted
for the MAR06 Meeting of
The American Physical Society

Score Statistics of Global Sequence Alignment and a Modified Directed Polymer Problem[1] MIHAELA SARDIU, Stowers Institute, GELIO ALVES, YI-KUO YU, NCBI/NLM/NIH — Sequence alignment is one of the most important bioinformatics tools for modern molecular biology. Using a variant of the directed path in random media model, we investigate the score statistics of global sequence alignment taking into account the compositional bias of the sequences compared. To accommodate the compositional bias, we introduce an extra parameter p indicating the probability for positive matching scores to occur. When p 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 p at infinite size is obtained. For alignment of length t, the score fluctuation $\sim \chi t^{1/3}$ is confirmed via scaling of the alignment score. Using the Kolmogorov-Smirnov statistics test, we show that χ follows the Tracy-Widom distributions: Gaussian Orthogonal Ensemble for p slightly larger than p_c and Gaussian Unitary Ensemble for larger p. The possibility of characterizing score statistics for modest system size (sequence lengths), via proper reparametrization of alignment scores, is illustrated.

[1] This work was supported by the Intramural Research Program of the NLM at the NIH/DHHS

Yi-Kuo Yu
NCBI/NLM/NIH

Date submitted: 30 Nov 2005 Electronic form version 1.4