Mean-Field Analysis of Recursive Entropic Segmentation of Biological Sequences SIEW-ANN CHEONG, Cornell Theory Center, Cornell University, PAUL STODGHILL, DAVID SCHNEIDER, USDA/ARS, Ithaca, CHRISTOPHER MYERS, Cornell Theory Center, Cornell University — Horizontal gene transfer in bacteria results in genomic sequences which are mosaic in nature. An important first step in the analysis of a bacterial genome would thus be to model the statistically nonstationary nucleotide or protein sequence with a collection of P stationary Markov chains, and partition the sequence of length N into M statistically stationary segments/domains. This can be done for Markov chains of order $K = 0$ using a recursive segmentation scheme based on the Jensen-Shannon divergence, where the unknown parameters P and M are estimated from a hypothesis testing/model selection process. In this talk, we describe how the Jensen-Shannon divergence can be generalized to Markov chains of order $K > 0$, as well as an algorithm optimizing the positions of a fixed number of domain walls. We then describe a mean field analysis of the generalized recursive Jensen-Shannon segmentation scheme, and show how most domain walls appear as local maxima in the divergence spectrum of the sequence, before highlighting the main problem associated with the recursive segmentation scheme, i.e. the strengths of the domain walls selected recursively do not decrease monotonically. This problem is especially severe in repetitive sequences, whose statistical signatures we will also discuss.

Siew-Ann Cheong
Cornell Theory Center, Cornell University