

International Society for Bayesian Analysis, 9th World Meeting,
Hamilton Island, Australia, 2008.

IMPORTANCE SAMPLING OF WORD PATTERNS IN DNA AND PROTEIN SEQUENCES

Louis H. Y. Chen^{1*}, Hock Peng Chan¹ and Nancy Ruonan Zhang²

¹ National University of Singapore, Singapore

² Stanford University, California, USA

* imsdir@nus.edu.sg

The use of Monte Carlo evaluation to compute p-values of pattern counting test statistics is especially attractive when an asymptotic theory is absent or when the search sequence or the word pattern is too short for an asymptotic formula to be accurate. The drawback of applying Monte Carlo simulations directly is its inefficiency when p-values are small, which precisely is the situation of importance. We provide a general importance sampling algorithm for efficient Monte Carlo evaluation of small p-values of pattern counting test statistics and apply it on word patterns of biological interest, in particular, palindromes and inverted repeats, patterns arising from position specific weight matrices, as well as co-occurrences of pairs of motifs. We also show that our importance sampling technique satisfies a log efficient criterion.