Abstract Submitted
for the MAR17 Meeting of
The American Physical Society

Using Maximum Entropy to Find Patterns in Genomes[1] SOPHIA LIU, ADAM HOCKENBERRY, ANDREA LANCICHINETTI, MICHAEL JEWETT, LUIS AMARAL, Northwestern Univ — The existence of over- and under-represented sequence motifs in genomes provides evidence of selective evolutionary pressures on biological mechanisms such as transcription, translation, ligand-substrate binding, and host immunity. To accurately identify motifs and other genome-scale patterns of interest, it is essential to be able to generate accurate null models that are appropriate for the sequences under study. There are currently no tools available that allow users to create random coding sequences with specified amino acid composition and GC content. Using the principle of maximum entropy, we developed a method that generates unbiased random sequences with pre-specified amino acid and GC content. Our method is the simplest way to obtain maximally unbiased random sequences that are subject to GC usage and primary amino acid sequence constraints. This approach can also be easily be expanded to create unbiased random sequences that incorporate more complicated constraints such as individual nucleotide usage or even di-nucleotide frequencies. The ability to generate correctly specified null models will allow researchers to accurately identify sequence motifs which will lead to a better understanding of biological processes.

[1]National Institute of General Medical Science, Northwestern University Presidential Fellowship, National Science Foundation, David and Lucile Packard Foundation, Camille Dreyfus Teacher Scholar Award

Sophia Liu
Northwestern Univ

Date submitted: 09 Nov 2016 Electronic form version 1.4