Project Summary

The number of sequenced genomes currently stands at over 1600 and will continue to grow exponentially for the foreseeable future. Each of these genomes is thought to contain a large amount of important information. Extracting and interpreting this information is a major challenge in biology. Our research will provide biologists with computational tools that will allow them to extract (a) codon specific missense and nonsense error rates and translation rates, (b) quantitative measures of the contribution these forces play in shaping the evolution of codon usage bias, and (c) reliable estimates of gene expression directly from an organism's genome.

While numerous heuristic methods for measuring codon usage bias (CUB) exist, their biological interpretation is limited. Our alternative approach to understanding CUB is based on mechanistic models of known biological processes. The proposed research will lead to "novel and significant advances in the use of biological data."[1] The model based approach we will use integrates mathematical models from multiple disciplines such as molecular biology, population genetics, and Bayesian statistics. Our previous successes in bioinformatics and the advances presented here clearly demonstrates the feasibility of our proposed research.

Intellectual Merits: The results of our research will affect "a significant segment of the biological research community supported by the NSF BIO Directorate."[1] For example, our current knowledge of error rates during protein translation is based on limited experimental data. Our research will give molecular biologists access to the additional information about these error rates held within an organism's genome. Because these error rates also have important implications for an organism's energy budget and other cellular processes, our research will provide important information to cellular and systems biologists. A better understanding of energetics of protein translation will give evolutionary biologists important insights into the selective forces driving the evolution of CUB, resolving a long standing debate in the field. Additionally, the vast majority of sequenced genomes are for non-model organisms whose biological roles are at best poorly understood. Predicting gene expression levels is an important first step in our ability to draw inferences about an organism's biology that will have implications for microbiology as well as the broader general public via medical, veterinarian, and phyto- pathology. The ability of other researchers to access this additional information will be greatly facilitated by the development "user friendly" versions of our computer code, something often missing from theoretically driven studies. By dramatically increasing the amount and nature of information we can infer from a genome, our research will have a substantial effect on the fields of molecular, cellular, evolutionary, and systems biology.

Broader Impacts: Through active mentoring by the PI and Co-PI for all participants, this project will provide postdoctoral, undergraduate and graduate students with hands on research experience. Students will present results at local, regional, and national meetings, including the annual Undergraduate Research Conference sponsored by the National Institute for Mathematical and Biological Synthesis (NIMBioS). The PI and Co-PI's ties with NIMBioS will also facilitate exposure of more advanced students to leading researchers in the field of mathematical biology. These advanced students will also be formally and informally involved in undergraduate teaching. In addition, the PI use this project as part of the NIMBioS sponsored teacher collaboration programs and their summer REU/REV program which targets women and underrepresented minorities. Finally, the PI and Co-PI will incorporate findings from this work into his undergraduate courses and other academic and public events.

[1] Advances in Biological Informatics Program Solicitation, NSF 10-567