New algorithms for genome skimming and its applications

Project Summary

Anthropogenic pressure and other natural causes have resulted in severe disruption of the global ecosystems in recent years, including climate change with extreme weather events, change (loss) of biodiversity, and invasion of non-native flora and fauna. The ability to quickly and inexpensively sample the taxonomic diversity in an environment is critical in this era of rapid climate and biodiversity changes. In place of the currently popular technique of isolating and sequencing specific phylogenetically informative regions, the PIs propose a low-pass whole genome sequencing (*genome skims*) and alignment free methods for barcoding. To enable this approach, the PIs will develop algorithms and tools to (a) *estimate the hamming distance given genome-skims of two organisms, and search a library of genome-skims with a query.*; (b) *use genome-skims for phylogenetic reconstruction, including placement of a query skim onto a taxonomy/phylogeny of the reference organisms.*; and, (c) *extend these techniques to 'meta-barcoding' queries (genome-skims of a mix of organisms).*

Intellectual Merit. If successful, the proposed activities will allow the estimation of genomic bio-diversity for a fraction of the current costs of labor and genome sequencing. The proposal uses a number of innovative and novel algorithmic and statistical techniques, and describes the first systematic study of the feasibility of computing the genomic distance using only a small, random fraction of the genome. The investigators have a strong history of prior research in related fields, but have complementary expertise, in evolution and phylogenetic reconstruction (Mirarab), and computational population genomics (Bafna).

Broader Impacts. Much of the planet's biodiversity is in the least developed and poorest places on the planet. Rapid environmental change and anthropogenic activities is degrading this biodiversity, and has potentially severe and lasting impact on all people, including in the U.S. The proposed tools enable the cataloging and measurement of bio-diversity through inexpensive sequencing, and easy-to-adopt laboratory protocols. While the proposal focuses on computational tools, the experiments will demonstrate proof of concept and the viability of larger sampling studies with a view towards deploying them where they are most needed. The proposal will also help train scientists who are better aware of the impact of rapid environmental changes on ecology, and the role of genomics and computation in investigating and alleviating these impacts.

Keywords: bioinformatics; computational genomics; computational evolution.