Spatial Analysis and Visualization of Genetic Biodiversity

Robert G. Beiko, Jacqueline Whalley, Suwen Wang, Harman Clair, Greg Smolyn, Sylvia Churcher, Mike Porter, Christian Blouin, and Stephen Brooks

Spatial analysis has been vital in understanding the species-species and species-environment interactions that underpin biodiversity. Examples of such analyses include the role of physical barriers in the speciation process and the spread of pathogens during an epidemic. An important challenge in such studies, and indeed through all of biology, is the definition of class or type, expressed for instance through various concepts of biological species. However, practical definitions of species can be made from physical attributes, behavioural patterns, or their impact (e.g. symptoms of a disease), and such definitions can therefore be imprecise and inconsistent. Microorganisms in particular are small and can be difficult to grow in isolation, making them difficult to identify and assess.

With the advent of ever-cheaper genome sequencing technologies, a wide range of organisms can now be examined at the genetic level, and biodiversity can be expressed in quantitative terms of genetic similarity and shared function. High-throughput DNA sequencing techniques are now being applied to samples collected from the environment without the intermediate step of laboratory culture. Microbial species can be assessed by collecting marker gene sequences or ‘DNA barcodes’, which can reveal the identity and diversity of organisms present in an environmental sample. Beyond species identification, the sequencing of random genetic material can reveal genes with important functions such as the degradation of pollutants, energy production, and nutrient cycling. However, the majority of such large data sets are less than five years old, and the genomics research community is still coming to grips with how to analyze (not to mention store and share) these data. There is an acute need for systems that can exploit the potential of genomic data sets containing billions of ‘letters’ of DNA.

Spatial analysis of such environmental genomic data can potentially reveal the key factors influencing genetic diversity in contaminated or pristine sites, and show the evolutionary course followed by a pathogen during an outbreak of infectious disease. We have developed Genome Space, an open source, three-dimensional geospatial information system to examine and test relationships between biodiversity, geography, and the environment. The goal of Genome Space is to acquire free cartographic and environmental data from many sources, and overlay user-provided data about sample site characteristics and genetic sequences. The program includes OpenGL-based map rendering, navigation controls and a library of tools to visually compare the properties of data collected from different sample sites, as well as a command interface based on the R statistical language.

We have prototyped the data integration and analysis capacity of Genome Space using several environmental genomic datasets. Important questions we have addressed include analyses of the relationship between microbial populations, geography, and time in a set of lakes in Wisconsin, USA. We also tackle a question of energy production in the world’s oceans by examining the distribution and diversity of the light-harvesting rhodopsin gene that has been observed at many sites around the world.