

Biodiversity Informatics: Enabling a Macroscopic View of Biology

Indra Neil Sarkar

MBLWHOI Library, Marine Biological Laboratory, Woods Hole, MA USA

There are an estimated 1.8 million organisms that are known to us on Earth. Still, much of biological and biomedical studies remain focused on a limited number of “model” organisms. These model organisms are crucial in the understanding of topics such as basic genetics, genotype-phenotype correlations, and disease inquiries. At the same time, such studies can be complemented with data from non-model organisms. Biodiversity informatics is focused on the application and development of techniques to link data spanning the full spectrum of life.

This presentation will explore the various aspects of biodiversity informatics, ranging from molecular data to population data. First we will explore the types of structured data that are available from existing resources (e.g., molecular sequence data and occurrence data). We will particularly emphasize DNA barcode data and its analysis, and how it can lead to the development of areas where there can be immediate synergy between biomedical inquiry and biodiversity knowledge. Next, techniques for extracting biodiversity data from unstructured sources (e.g., literature) will be discussed. In particular, we will focus on the linking of information across data sources using natural language processing techniques that are honed to identify biodiversity entities (with a particular emphasis on organism name identification).

Throughout the presentation, we will ground ourselves in the discussion of how biodiversity data can complement biomedical studies. To this end, the discussion of biodiversity informatics will be done within two contexts: (1) Infectious Diseases and (2) the Biology of Aging. In so doing, we will demonstrate how a “macroscopic” view of life on Earth can be used to facilitate biomedical studies.