Accepted Manuscript: Asia-Pacific Biotech News. 11(1): 15-18.

Grand Challenges in Biodiversity Informatics

Indra Neil Sarkar, PhD

The exponentially growing array of biological data has necessitated the development of a new information management domain, *biodiversity informatics*. It is one of the newest members of the 'informatics' sub-disciplines, which all generally focus on the management of information through the application of advanced technologies. Like other informatics sub-disciplines, biodiversity informatics depends on fundamental computer science and information science principles to facilitate the management of heterogeneous data. Biodiversity informatics distinguishes itself as being the most focused on biological knowledge dating back to the earliest dates of recorded history – while most biological or biomedical informatics studies focus on organizing and studying information spanning less than 100 years, the scope of biodiversity informatics spans the age of the Earth. Biodiversity informatics is also concerned with the widest range of disparate data types – including climatology, epidemiology, geography, and taxonomy. To this end, many informatics principles can readily be incorporated into biodiversity informatics; however, there are equally as many challenges that will require creative solutions. Here, several such challenges are presented in an effort to lay a framework for the types of issues that will define the future of biodiversity informatics and, in turn, the future of biology and biomedicine.

Linking the Past to the Future

Many contemporary studies emphasize the molecular aspects of life. It is important to note, however, that such molecular studies are only associated with less than 1% of the 10% of extant life that has yet been described (assuming ~10 million extant organisms). Nonetheless, population-level information, including scientific description and distribution information, may be available in archival form for species that are not yet associated with molecular information. This archival data might consist of either actual specimen data or literature descriptions in collection institutions or even hobbyist archives that may not be part of formal national collections. In some cases they may be the only available information for some species (e.g., if the species is extinct or cryptic). These data, in addition to the eventual data associated with organisms not yet discovered (i.e., at least 90% of the Earth's biota), may help elucidate relationships between and among organisms dating back to the earliest found fossils. The first part of our archival knowledge will be made possible through the Biodiversity Heritage Library [see: bhl.si.edu]. This consortium of ten institutions across the United States of America and the United Kingdom are joining forces to digitize much of the worlds 'heritage' natural science literature and make it publicly available. The international consortium for the Barcode of Life [see: barcoding.si.edu] will complement this heritage literature through its development of methods and algorithms to facilitate automated identification of known species. Add to this the exponentially growing volume of new literature and molecular sequence data that will continue to be produced and the number of new inquiries that can be made is simply amazing. There will thus be a significant need to develop methods and interfaces that can help elucidate linkages between literature, specimen, and molecular data.

Names and Places

Biology, by definition, is a discipline associated with organisms and tracking of where they might have been, where they are now, and where they might go into the future. Significant

research and investment has been made into the development of information retrieval methods to identify and assess relevant information among singular data types (e.g., molecular sequence can be retrieved from sequence databases using a program called BLAST). By linking relevant information associated with an organism and mapping it to geographic information (called "georeferencing"), scientists can combine information from a range of data types (e.g., climatology and epidemiology). The name of an organism and its location information (e.g., where it was collected) are perhaps few elements out of the possible elements that are associated with almost all biological data. Names of either organisms or places as identifiers present a problematic situation. Common ("vernacular") names will differ according to culture, language, and context. For example, the fish *Pomatomus saltator* might be referred to as "bluefish" or "greenfish" depending on the context. Scientific names might also change over time – e.g., *Pomatomus saltatrix* is a synonym for *Pomatomus saltator*. Similarly, names of places change with time and culture (e.g., Istanbul was formerly known as Constantinople and Byzantium). There are many contemporary tools that can take geographic information and plot it graphically (e.g., Google Earth). However, there are very few biological specimens that have their data in a form that can readily be used by Google Earth or more complex Geographic Information Systems (GIS). Two significant projects have embarked on trying to organize information according to organism name or geographic locations, respectively the uBio project [see: www.ubio.org] and the BioGeoMancer project [see: www.biogeomancer.org]. While there are many lists of organisms (often called "checklists") and geographic places (often called "gazetteers"), both of these projects have tackled the grand challenge of creating singular, harmonized, centralized indices. These indices can then be used to identify relevant data across a whole range of resources that might contain relevant biological data. This may lead to applications that can analyze trends of where past, current, and future distributions of species are and can thus be analyzed in combination with organism and/or spatial data from other disciplines.

Biodiversity Knowledge in Practice

There are billions of specimen records and observational data that exist in natural history collections worldwide, and continue to grow thanks to significant collection efforts. These data are predominantly in collection institutions, especially natural history museums and herbaria. The biodiversity community is actively engaged with establishing international frameworks to reliably access and share biodiversity data. Most notable is the Global Biodiversity Information Facility (GBIF). GBIF is an international organization that strives to develop a federated network of biodiversity information [see: www.gbif.org]. GBIF strives to make data that has never before been electronically accessible such that automated protocols can be designed to directly retrieve information from data providers, such as collection institutions worldwide. The success of the GBIF infrastructure and the ability to incorporate intelligent tools for identification of information might lead to valuable resources that can serve the entire spectrum of society – e.g., from lay people, to school children, to hobbyists, to researchers, to forensic experts, and to medical practitioners. To this end, the biodiversity informatics community has recently begun an endeavor to create the "Encyclopedia of Life," as first envisaged by E.O. Wilson. This resource would enable anyone to dynamically retrieve digital information objects (e.g., digitized text and images along with molecular data) from any accessible relevant resource. Prototype applications have been developed to demonstrate how this might work – e.g., Rod Page's *ispecies.org.* As society continues to produce exponential amounts of data, traditional mechanisms for organizing

and navigating knowledge will be challenged. In light of this, the integration of "Web 2.0" technologies (e.g., Blogs and Wikis) with the Encyclopedia of Life promises to bring together communities in ways that might not have yet been conceived. As modern society embarks towards the development of a global economy, unprecedented stresses are being placed on its ecologies. This is especially apparent through the rapid emergence of infectious diseases around the globe. Through a free and publicly available Encyclopedia of Life, society might be able to better track incidences of emerging infectious diseases that are associated with organisms, including vector-borne diseases, increasingly antibiotic-resistant bacterial infections, and rapidly mutating viral syndromes. To this end, the incorporation of biodiversity knowledge in the context of biomedical inquiries is crucial to the fundamental understanding of disease.

Conclusion

Information management is certainly not a new phenomenon in society. Since the first libraries, efficient methods have been core to identifying and synthesizing information across a wide range of resources. With the emergence of computers and the subsequent emergence of informatics, this capacity was greatly increased. Biology is also not a new discipline. Since the first recorded studies of Linnaeus, organisms have been considered essential in the study of life on Earth and one of the underpinning sources for medical breakthroughs, either by being a source of ailment or a source of healing. Biodiversity informatics strives to bring together our heritage knowledge of life with the current and future understanding of life. Insights gained from biodiversity informatics studies might thus well inform and provide breakthrough developments across the entire spectrum of biology and biomedicine.

Contact Details:
Indra Neil Sarkar, PhD
Marine Biological Laboratory
7 MBL Street
Woods Hole, MA 02543 USA
Email: sarkar@mbl.edu