A Taxonomy for Immunologists

James A. Overton, Randi Vita*, Jason A. Greenbaum, Heiko Dietze, Alessandro Sette, Bjoern Peters

Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA; Berkeley Bioinformatics Open-Source Projects, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mailstop 64-121, Berkeley, CA 94720, USA

* rvita@liai.org

The NCBI Taxonomy is a valuable resource for coordinating information about biological taxa. Its 378,802 taxonomy nodes cover the superkingdoms of Archaea, Bacteria, Eukaryotes, Viruses and Viriods, down to subspecies and strains. However, the size and scope present practical problems when using the Taxonomy in applications, specifically when using the NCBI Taxonomy in a search interface for specialized databases. Some taxa (such as specific mouse strains) are not included in the NCBI taxonomy and the naming of nodes is better suited to the field of taxonomy than to general biological usage. Therefore, many bioinformatics projects have a need to customize the Taxonomy to suit their purposes. We present a solution to this common problem.

The Immune Epitope Database (IEDB) is a free resource that facilitates search and analysis of published experimental data about the recognition of epitopes by immune adapters. Since its inception, the IEDB has used NCBI Taxonomy identifiers to annotate host organisms and epitope sources. Our data include approximately 3500 distinct NCBI taxa (mainly species), and approximately 1800 additional taxa not in the Taxonomy (mainly strains and subspecies). The iedb.org website provides a number of "finders" to help users search the nearly 100,000 epitopes and more than 600,000 assays in the IEDB. The goal of the current project was to improve the organism finder by building an easy-to-use fragment of the NCBI Taxonomy, augmented with additional strains and subspecies. The most important of these is the "organism tree", which includes all the taxa used in the IEDB.

Our technique combines ontology construction and extraction in a configurable workflow, specifically designed for manipulating the NCBI Taxonomy. The result is a smaller taxonomic tree with more familiar labels that is easier for users to navigate. Crucially, the new tree preserves the NCBI's identifiers and the truth of sub-type relations. Here we describe the construction of the IEDB organism tree, but the technique we developed could easily be adapted to suit other bioinformatics projects that utilize parts of the NCBI Taxonomy.

We have evaluated the organism tree in two ways. First, the organism tree is shallower than the NCBI Taxonomy on various measurements: the maximum, mean, median, and standard deviation of the number of ancestors are lower, and the number of ancestors is dramatically lower for frequently accessed species such as *Homo sapiens*, *Mus musculus*, and *Drosophila melanogaster*. Second, survey results show that IEDB users can classify species in the IEDB organism tree with greater accuracy and greater certainty than in the NCBI Taxonomy. We conclude that our organism tree is better suited to the needs of the IEDB than the unmodified NCBI Taxonomy. The workflow we have developed is specifically designed to manipulate the NCBI Taxonomy, but could easily be adapted by other projects that use parts of the Taxonomy.