Ground Truth

The concept of ground truth is a well-established principle in cartography, where data collected at a distance are confirmed by measurements made on location. Those local measurements are used to calibrate remote sensing devices, verify or correct experimental inferences, and update geographic databases. Ground truth observations also provide a means of training and supervising image classification software and resolving errors of omission or commission. Cartographic methods have improved significantly because of the development of precise positioning methods (GPS), the development of interoperable data standards for rapid exchange of precise and highly interlinked information, and the development of various devices and visualizations that serve-up information on-demand to different classes of end-users.

There are strong parallels between mapping geographical and biological space. Much of what is underway in genomics, evolutionary biology and systems biology is analogous to the development of a coordinate system onto which living systems can be mapped, natural boundaries and interrelationships uncovered, and predictions of properties and behaviors based. However, it is likely that the dimensionality of any biological coordinate system will exceed the four dimensions of the geographical system. This will confound visualization and complicate "navigation" through biological space, whether it is for purely exploratory purposes or to get from one point to another.

It is a given that the volume of biological data will continue to grow super-linearly for the foreseeable future, as new computational methods are applied to answer the "big questions" in biology. In the absence of major innovation, it is likely that the gap between the cost of data analysis and the cost of data generation will continue to widen. The outcome of such analyses are highly dependent on the quality of the input data, including the underlying information and knowledge used to inform the creation of datasets, the algorithms used in analyses, and the interpretation. Errors of commission and omission appear to be more common in biological data sets than physical data set and the former are more likely to be affected by semantic ambiguity and hidden biases. What is not yet established is which of the labor-intensive curatorial and interpretive tasks can be automated and what metadata that is absent from the public databases may be located and recovered from other sources in a usable form.

The impact of semantic ambiguity in biological data has been noted previously as it pertains to identifiers [1] or biological names [2]. These problems confound accurate and complete retrieval of biological data from public and private databases and from the biological literature; especially in cases in which taxonomic information was misinterpreted or the source organisms were misidentified. Bortolus [3]provides some insights into error cascades in the biological sciences that are attributable to this problem. While his remarks were aimed at field ecologists, the observations apply to computational biologist, modelers, and system biologists as well. So too, do the consequences that such errors have on nature, our knowledge of nature and the socioeconomic costs. Laurin [4] and Hillis [5]provide some additional insight into the potential challenges that will arise as the first adherents of the Phylocode begin to apply their system of nomenclature to plants and animals and their data sets flow into the public repositories. This will add yet another layer of complexity to mining databases and the literature and will represent a source of methodological and theoretical bias that will need to be factored into interpretation of biological data in the future.

This is not an unfamiliar territory to those who have carried out "large-scale" phylogenetic, taxonomic, or ecological analyses in the past. Incorrectly labeled data and data derived from incorrectly identified samples remain common and will continue to confound naive users of public databases and the literature. Tools and techniques to detect and visualize such discrepancies could be useful as components of analytical pipelines, data submissions routines, or as value added service. Authoritatively maintained reference sets of gene and genome sequences derived from taxonomic type material are also needed, especially those that are richly linked to validated phenotypic, physical and geographic metadata and delivered in a highly structured, standards compliant form

[6,7]. *The Genomic Encyclopedia of Bacteria and Archaea* [8] represents an outstanding example of an international collaboration that aims to produce such high value data and will provide the ground truth for the next generation of phylogenetic models on which future studies of bacteria and archaea will depend.

George M. Garrity

August 19, 2009

References

1. Clark T. Identity and interoperability in bioinformatics. Brief Bioinform 2003; 4:4-6. PubMed doi:10.1093/bib/4.1.4
2. Garrity GM, Lyons C. Future-proofing biological nomenclature. *OMICS* 2003; **7**:31-33. PubMed doi:10.1089/153623103322006562
3. Bortolus A. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. *Ambio* 2008; **37**:114-118. PubMed doi:10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2
4. Laurin M. The splendid isolation of biological nomenclature. *Zool Scr* 2008; **37**:223-233. doi:10.1111/j.1463-6409.2007.00318.x
5. Hillis DM. Constraints in naming parts of the Tree of Life. *Mol Phylogenet Evol* 2007; **42**:331-338. PubMed doi:10.1016/j.ympev.2006.08.001
6. Garrity GM, Field D, Kyrpides N, Hirschman L, Sansone SA, Angiuoli S, Cole JR, Glockner FO, Kolker E, Kowalchuk G, *et al*. Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 2008; **12**:157-160. PubMed doi:10.1089/omi.2008.A2B2
7. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, *et al*. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 2008; **26**:541-547. PubMed doi:10.1038/nbt1360
8. Anonymous. US Department of Energy Joint Genome Institute: Genomic Encyclopedia of *Bacteria* and *Archaea*. http://www.jgi.doe.gov/programs/GEBA/.