OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Biological Data Extinction

PETER LI

As BIOINFORMATICIANS, we think of biological data in terms of an experiment and as a snapshot of some biological process. As such, we store and process large quantity of this data, and we seek to preserve the "sanctity of this data in perpetuity." However, our users, the biologists, never look back at old data once they have reconciled the data by an appropriate biological model. In truth, data in biology become extinct and replaced by an understanding of the biological processes that generated them. The data of interest to our biologists are always the newer data that expand upon or challenge that understanding. This generation and expiration of data, that is, a life cycle, runs counter to the bioinformatician's philosophy of data existence and permanence.

What is this life cycle, and why is it important? To answer that, we need to follow the lifetime of a typical piece of biological datum. It is first born from an experiment conducted by a researcher. That experiment, in turn, is based on a hypothesis generated from a model for some biological processes. This data is stored, processed, retrieved many times during the course of research. It is accumulated until, collectively, they confirm or refute the hypothesis. However, once they achieved their purpose, their particular contributions will be integrated into the biological models that spawned them, that is, "replaced" by a better understanding of biology. Afterwards, the data is forgotten as the researcher progresses to the next hypothesis and the next experiment.

If the generation, utilization, and expiration of biological data happen in a controlled, predictable, or designed fashion, we might have call this a "life cycle." However, these events in the real world often reflect the Darwinian competition of scientific thought itself, that is, only the fittest survives. In this case, only the most perplexing data stay to challenge the next generation of models, while the rest unceremoniously become extinct as they are absorbed by refinements of the scientific process. Without recognition of this fact, long-running bioinformatics systems will suffer the same fate, as they are clogged with extinct data, they ultimately will be replaced by newer systems without such baggage.

How do bioinformatics systems avoid the fate of extinction? Because of the inherent nature of the data, the data must be stored and manipulated with its context of the experiment, the hypothesis, and the model. As data, experiments, hypotheses, and models are reconciled, they are updated, archived or removed from the system appropriately. This is easier said than done. This expansion of information about a piece of data increases the complexity of our bioinformatics systems (databases, algorithms, UIs). Consequently, we don't consistently capture this information: sometimes partially, sometimes implicitly, and sometimes not at all. Without complete background information, we are unable to reconcile and remove data from our systems, despite that much of it have become obsolete in the minds of our users through the course of scientific achievements.

One might counter that, in a "hypotheses from data mining" paradigm, the experiments and the data are considered permanent so that new hypothesis can be continuously mined from the increasing stockpile of data. However, experimental data will change as the scientific process improves: newer instrumentation, techniques, and procedures expand and refine the experimental data. The newer and "cleaner" versions should deprecate older copies, but we rarely do that. In part, the replicated experiment is not really the same

Celera Genomics, Rockville, Maryland.

as the earlier one. Therefore, the data warehouse in support of this approach becomes clogged with extinct data and will suffer the same fate unless it undergoes periodic cleansing.

The same phenomenon of extinction applies for bioinformatics tools. They represent a snapshot of our current biological models, codified to produce "predictive" results. The use of these tools depends on the quality of the results. As better tools come to being, older ones are replaced and forgotten. The reason is that better tools are based on better models of biology, that is, the models that survived the "natural selection process" in the domain of science. This problem is also manifested at a higher level, because the integration of many tools to form a seamless system is often what the users want. This integration is at risk for upheaval whenever a component tool is replaced. Indeed, in most cases, new tools are added, old tools are never deleted. This ever-increasing system "mass" slows its own evolution until it stops adapting altogether and then eventually be replaced by a more nimble system.

As bioinformaticians, how do we avoid the mass effect of extinct data and tools in our systems? We need to make explicit plans for a "life cycle" early in development, so that we can safely retire components when the time comes. Such life cycle planning is not a new discipline: planned obsolescence is a common fact in our lives. The cost of not making such plans is to eventually see our investments suffer an unglorified extinction.

DISCLAIMER

The views and opinions expressed in this essay do not necessarily state or reflect those of the author's affiliation, Celera Genomics.

Address reprint requests to:
Dr. Peter Li
Celera Genomics
45 West Gude Drive
Rockville, MD 20850

E-mail: lipw@celera.com