Challenges of Information Retrieval and Evaluation in Data-Centric Biology

Yi-Kuo Yu

Dear Editor:

THE IMPORTANCE OF DATA in science can never be overstated. The three laws of planetary motion proposed by Johannes Kepler were founded on careful examination of the data collected by Tycho Brahe, Kepler's mentor, and it was these three laws, extracted from the observed data, that eventually helped Newton to develop his law of gravitation. This law has had far-reaching impact, playing a major role in the planning of moon landings. If the sun is viewed as a spherical mass, according to Newton's law, the planetary orbit is a fixed ellipse. The rate of precession of Mercury's orbit, again from observed data, was eventually shown to disagree with Newton's law after careful reanalysis of other possible causes of orbital precession using a Newtonian-based approach. This disagreement between observed data and Newtonian theory called for an explanation that was finally provided by Einstein's alternative expression of gravitation: General Relativity. This simple example illustrates the importance of starting with observational data and how the experimental data guides the development of theories that not only summarize the data but also provide powerful predictions.

Looking at the functional and organizational complexity achievable in a myriad of diverse organisms—all of which share a universal set of building blocks (water, ions, saccharides, fatty acids, amino acids, nucleotides, and other small molecules)—one must admit that our knowledge in life science remains quite limited. In fact, despite vast amounts of effort invested, our current understanding of biology, ranging from a microscopic level to a more macroscopic level, is far from complete. For example, in terms of molecular interactions, phenomenological treatments at the intra- and inter-protein level have not yet fostered an effective theory capable of predicting how proteins fold and how a protein complex organizes itself. Namely, a proper coarse-graining procedure to bring out only the relevant degrees of freedom is lacking. This may be due to our insufficient understanding of biology in terms of relevant physics and chemistry, or it may be because the search for higher organizing principles is hindered by our inability to bring out information buried in noise and/or by conflicting interpretations of data.

A natural question thus arises. What makes biological data harder to interpret than data types such as that of planetary motion? The answer to this question indicates challenges for

data intensive approaches in life science. First, due to finite life span of organisms, time translational invariance in biological systems is in general violated. This makes the task of data purification through time, as is done with planetary data, rather difficult. Second, biological responses are environment- and context-specific. Identical stimuli in different biological environments need not trigger identical responses. Third, high throughput data accumulation inevitably introduces noise that may not be well controlled into large data sets. This noise may hinder meaningful biological inferences from observed data. Fourth, there are numerous data generated by computationally processing experimental data and the noise or uncertainty within experimental data may be amplified through these approaches and then feed back to the next round of data processing. That is, there is an inherent danger of *error amplification* and *error propagation*. A simple example of error propagation already occurs in literature citations. An article may be miscited by a highly referenced paper and the incorrect citation therein may propagate into many more papers for a long time.

To clearly illustrate some of the aforementioned challenges, we describe the gaps between what is ideal and what is currently achievable, and provide some potential directions for bridging the gaps. We will use examples from mass spectrometry (MS) based proteomics here on.

In the postgenomic era, proteomics is among the most challenging and most important subjects in biology. The most important issue in MS-based proteomics is peptide identification via tandem MS (MS/MS). The ideal scenario is as follows. One uses a certain enzyme to digest, in a mixture of interest, proteins into peptides. One then separates peptides through chromatography techniques, charges the eluted peptides, and then sends them for MS/MS analysis. From each MS/MS spectrum, one identifies the underlying peptide based on the mass fragments observed. Given the list of peptides identified, one may then infer and preferably quantify the proteins present in the original mixture. What is currently achievable, however, is far from ideal. First, the percentage of spectra that can find matched peptides is quite low. That is, despite a large number of spectra generated each time, one cannot obtain many confident identifications. Second, chromatography is not yet able to separate fully all peptides present. This implies that one must consider the possibility of coeluted peptides during the data analysis. Third, the task of identifying the underlying peptide for a

given MS/MS spectrum is far from trivial. The major problem involved is a statistical one: how does one assign statistical confidence to the peptides identified?

Just as biological responses depend on the environment, the peptide fragmentation patterns also depend on the MS instrument and chromatography protocol (determining materials coeluted, or copresent, with the underlying peptide). That is, the underlying noise varies *per spectrum*. Naturally, when assigning statistical significance to candidate peptides, one should not ignore this important factor and the statistics should be *spectrum-specific*.

Further complications exist. There exist two types of variations in proteomes. The first one is a manifestation of broken time translational invariance: the proteome of an individual changes over time due to development, aging, and diseases. The second type arises from genetic variations among individuals, resulting in different proteomes. The best way to utilize proteomics data may depend on which type of variation is larger. However, one thing is certain: although existing protein databases can be used as references, one must allow variations when a specific individual is considered. To achieve personalized proteomics, leading to personalized medicine, one must take into account single amino acid polymorphisms (SAPs), different posttranslational modifications (PTMs), as well as their links to diseases. Blind inclusion of such information may cause a huge expansion of search space and a reduced sensitivity. Carefully thought through approaches with modular knowledge integration should be developed to handle this problem.

Another issue within MS-based proteomics arises when one needs to compare data analysis results from different approaches. Unlike converting length from meters to feet, there is no simple means to translate the peptide *score* assigned by one method to that of another. This fundamental problem comes from the lack of a universal statistical standard that naturally accommodates *spectrum-specificity*. Even though this point is finally recognized by the community, many practitioners mistakenly called *score-specific* statistics *spectrum-specific* statistics. It is, of course, useful if one can properly combine data analysis results from different ap-

proaches. The main problem here lies in the fact that most data analysis tools are using more or less similar characteristics to score candidate peptides. As a consequence, there is a non-negligible correlation between analysis methods. Taking into account correlations among methods while combining analysis results is thus a direction worth exploring.

Finally, we would like to reemphasize that the main challenge in massive data generation is quality control. After all, data is useful only if it contains more signal than noise. Data intensive science becomes powerful only when more information can be confidently extracted from the massive data accumulated. Although biological data is intrinsically more complicated, compared to planetary data, useful information may be extracted via carefully and correctly designed statistical methods. Furthermore, the need to dig deeper into massive data sets may inspire careful development of novel statistical methods such as how to properly combine analysis methods that are weakly but nonnegligibly correlated.

Acknowledgments

This research was supported by the Intramural Research Program of the National Library of Medicine of the National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by the NIH.

Author Disclosure Statement

The author declares that no conflicting financial interests exist.

Address correspondence to:

Yi-Kuo Yu

National Center for Biotechnology Information/NLM/NIH

9000 Rockville Pike

MSC 3829

Bethesda, MD 20894

E-mail: yyu@ncbi.nlm.nih.gov