frontiers in GENETICS

SPECIALTY GRAND CHALLENGE ARTICLE
published: 08 June 2011
doi: 10.3389/fgene.2011.00026

Frontiers in genomic assay technologies: the grand challenges in enabling data-intensive biological research

David William Galbraith*

BIO5 Institute, University of Arizona, Tucson, AZ, USA
**Correspondence: galbraith@arizona.edu*

INTRODUCTION

We live in astonishing times, not least when we consider the recent rate of progress in science. We are experiencing a revolution, underpinned and empowered by Moore's famous law and the computational power enabled by that law (Mollick, 2006), that has largely eliminated the traditional scientific bottle-neck associated with data acquisition. Biologists now find themselves a situation familiar to any of us that has, over the last decade or so, replaced their conventional film camera with its digital counterpart. The biologist now has the ability to take technological "snapshots" of almost any attribute of a living organism, whether it be DNA sequence, protein identity, 3-, 4-, and 5-dimensional distributions of macromolecules, composition of small molecules, and ions, identity of mutations, the mapping of phenotypes to genotypes, and so-on, and this has provided an unprecedented amount of data for analysis. This situation, which extends across all fields of experimental science, has been canonized as the emergence of a "Fourth Paradigm" in scientific research, termed Data-Intensive Scientific Discovery, the first, second, and third paradigms representing conventional empirical observation, abstraction and theory, and computational simulation, respectively (Gray, 2009). As you will see below, biologists now face much of the same grand challenges as experienced by the recreational photographer.

The purpose of this article is to summarize grand challenges confronted at the frontier of genomic assay technology, and to provide some talking points as to the purpose of this new journal. Several grand challenges are relatively easy to define. At least three are directly associated with our recent and unprecedented ability to generate biological data in digital form. Given this ability, immediate challenges are: (1) How does one design the most meaningful experiments to generate the most useful data? (2) How does one handle the resultant data stream? and (3) What is the best way to archive data, and to harvest and integrate this archived data?

A fourth grand challenge naturally arises from the process of technology innovation, being associated with the identification of new ways to analyze biological organisms and systems, and the development of the associated technologies and experimental platforms.

A final grand challenge arises from the concept of stepping back and focusing on the source materials used for generating biological knowledge, the families, genera, and species, discovered and undiscovered, representing life on this planet, and asking the question as to whether we know all possible ways in which evolution can give rise to successful life-forms based on detailed understanding and study of only a very limited subset of these life-forms.

DETAILS OF THE GRAND CHALLENGES

DESIGNING MEANINGFUL EXPERIMENTS IN THE CONTEXT OF FACILE DATA PRODUCTION

Changing the rate-limiting step on biological experiments, away from the process of data production, has unexpected consequences. From a practical viewpoint, although it provides a greatly expanded range of possible experiments, it also has the effect of decreasing the focus both of the experiments and of the experimenter. At the same time, the commoditizing of experiments, and their ever-decreasing cost, runs the risk of decreased rigor in experimental design. This is particularly important from the point of view of statistics, since any experimental design should not only provide data that most directly test the questions of greatest interest, but also avoid confounding variables when it comes to the point of analysis. The context of the experiment can also be a factor: within the public sector, any dataset can be harvested to provide information for follow-up, which when published can then be defined as valuable. From this viewpoint, dealing with false-positives, such as those generated when insufficient power is associated with the experimental design, may not be a serious issue. In contrast, within the targeted research programs of the private sector, such as those aimed at the discovery of new drugs, false discovery is particularly problematic, since the downstream testing, modification, and validation of lead chemical structures requires significant investments of both time and resources. The grand challenge of optimizing both the choice of experiments and the associated experimental design will remain persistent, given that our technical abilities to perform these experiments, and therefore their scope, will continue to increase exponentially, with corresponding decreases in the associated costs, at least for the immediate future, and most probably for a further extended period.

An additional problem is associated with the fact that most genomic assay technologies now produce datasets of extremely high dimensionality. This means that the datasets provide greater insight into the biological system than can be accommodated in the summary descriptions (the associated metadata) of the experiments that are being performed. From this viewpoint, the concept of replicating an experiment, central to the scientific method as implemented in biology, comes into question. If we cannot describe the source material and the experimental treatment(s) with the same level of detail as the resultant datasets, we cannot be sure that we have identified all variables that might affect both the datasets and our understanding of their meaning (Galbraith, 2006). Many examples of this problem are found in the historical record: for example, before 1998 or thereabouts, no attention was paid to the role of small RNA species in gene expression, since these were widely considered to simply represent breakdown products of the larger, more relevant

mRNA species. The discovery of the crucial importance of small RNA molecules in multiple cellular regulatory processes dramatically shifted our understanding of the process of gene expression. In another example, increasing evidence is emerging of transgenerational inheritance of epigenetic states (see, for example, Ng et al., 2010). This observation, which tints Mendel's laws with shades of ambiguity, implies that experiments involving analysis of individuals cannot provide a complete description without the availability of a history of their provenance over generations, and it becomes unclear as to what limits (if any) should be applied to this retrospection, since it rapidly becomes cost-prohibitive and is usually impractical. Bioinformaticians increasingly worry that the imposition of prescriptive structures (such as standards, schemas, controlled vocabularies, and ontologies) simply defines a shared consensus of the world, which therefore cannot represent the frontier of research (Barga et al., 2011). Solving these issues represents a grand challenge that again will not be easy.

HANDLING INCREASED DATA STREAMS

In 2007, for the first time, global digital data production, about 260 EB, exceeded our ability to store this data (Gantz et al., 2008). The size of the digital universe estimated for that year is larger than the number of stars in the universe and, given conservative predictions of growth rate, in 15 years will surpass Avogadro's number. Another context for viewing this problem involves the 2007 prediction that by 2011 (i.e., the present day), almost 50% of digital information that we create can and will have no permanent home. The grand challenge here involves addressing the question as to how best to analyze data streams with the understanding that much will not be permanently stored.

In terms of analysis of the data streams, increases in global computation power have followed an exponential course (Hilbert and Lopez, 2011), although this may not be sufficient for adequate analysis. Gray (2009) notes that data analysis typically can be classified either as a search for specific anomalies, which can be readily implemented on grids of computer clusters, or as exercises in correlation analysis, to identify global patterns underlying the data. The latter, which involves algorithms that are typically cubic in N, are much less easy to implement, and in many cases impossible. Part of this grand challenge therefore calls for the design and implementation of new, more-rapid algorithms, which necessarily will provide approximate answers, but which will nevertheless be very useful.

DATA ARCHIVING

An immediate question, given our technical inability to store all of our current data streams, is as to which data should be stored for future analysis. This question is complicated by the observation that we cannot know what methods of data analysis and what novel insights will emerge in the future that would guide the types of data analysis that should then be done on previously archived data. Further challenges in data storage concern maintenance of the data in digital archives in a state that is both readable and comprehensible; Curry (2011) provides a cautionary tale from high energy physics.

It will also be a grand challenge to determine which currently stored data should remain in storage and which discarded. Certainly, improvements in accuracy already render much stored data obsolete, sequence information in GenBank being a prime example. The cost of careful curation of the stored data may well exceed the cost of simply replacing it. This conundrum will affect all high throughput data streams, and will be a persistent problem to be confronted, given the fact that data replacement will generally be associated with exponentially decreasing costs.

NEW TECHNOLOGIES

We can confidently expect increases in accuracy, throughput, and scale, and decreases in cost, for existing platforms and technologies. A grand challenge will be to push these technologies to their technical limits. However, one the most intellectually exciting aspects of the frontier is the development of new and transformative technologies for examining biological organisms, their genomes, and the ways that these genomes function. From a philosophical viewpoint, evolution, acting over millennia, simply serves to codify the ways in which living organisms can survive and reproduce. This codification occurs within, and employs, all physical and chemical laws and mechanisms that are experienced by the evolving organism. It appears likely, therefore, that we will uncover physical and chemical mechanisms whose involvement in the activities of living systems was previously unsuspected. The associated grand challenge will involve the development of accurate technologies and platforms to detect and chart these mechanisms. A recent example includes observations that implicate the vibrational modes of molecules, rather than simply their space-filling structures, as involved in biological detection and function (Franco et al., 2011). If this is true, the development of instrumentation to readily measure these vibrational modes will become essential.

EXPLORING GLOBAL BIODIVERSITY

Implicit in the previous grand challenge is the assumption that our sampling of global biodiversity may have been insufficient to identify examples of organisms that represent all possible modes of biological evolution. Certainly, our funding mechanisms focus heavily on few species, and these species have, in many cases, been subjected to stringent selection for "appropriate behavior." For example, of the ~500,000 species of angiosperms (flowering plants), agricultural crops represent a very small subset (about 2,500 species) and, of these, 103 supply over 90% of the calories for human consumption. These crops have, without exception, been engineered to be amenable to crop improvement, requiring predictable behavior in crossing, in production of seeds, in cultivation under controlled conditions, and so-on. The implication is that unusual mechanisms of organismal development and reproduction would not survive this process of selection. The fact that descriptions do exist of organisms having unconventional genetic and biosynthetic mechanisms, with particularly good examples coming from marine biota (cf. diatoms and dinoflagellates), gives us confidence that additional, peculiar biodiversity exists and is yet to be described. Of course, this can be controversial, as indicated by the recent vigorous debate as to whether or not arsenic can replace phosphorus in living organisms (Wolfe-Simon et al., 2010). As a final note, urgency in pursuing this grand challenge is particularly emphasized by the increasing rate of anthropogenic change, since we risk losing species before they are characterized (Barnosky et al., 2011).

CONCLUSION

A few words in conclusion are worth on the subject of the philosophy of science. A colleague recently commented to me about differences between the way in which scientific discovery is pursued in the physical and biological sciences. In the physical sciences, astrophysics and particle physics being good examples, the accepted scientific approach involves the design and construction of instrumentation, to the highest possible design tolerances and measurement accuracies, to provide datasets that are as complete as can be achieved concerning the objects of interest; these objects range from galaxies at the limits of detection, to more hypothetical objects, such as the Higgs boson, other novel elementary particles, dark matter, and dark energy. The resultant datasets are then analyzed in a more-or-less hypothesis neutral fashion, in order to detect the increasingly subtle indications of interactions, representing underlying natural order from which speculation as to general mechanisms, rules, and laws can be based. If these speculations have predictive value, which can subsequently be tested, so much the better.

In contrast, particularly in the period covering the latter part of the twentieth century, biologists have concentrated on hypothesis-driven scientific testing. Given that all hypotheses proposed for biological systems likely are incomplete, since we know little about the systems under study, and furthermore most probably will turn out to be incorrect, based on the historical record, one wonders if the time is right for a critical reevaluation of the predominance of hypothesis-testing within biological research. In other words, it may now be time for biologists to think more like physicists. The irony of this statement has not escaped our notice!

In summary, therefore, the role to be played by frontiers in genomic assay technology will include showcasing of new methods, technologies and platforms for probing the functions of living organisms, the critical analysis of methods for data acquisition, analysis and archiving, the exploration of new and uncharted organisms and life-forms, and an encouragement of healthy debate as to the philosophy of science underpinning advances in biological research. We live in exciting times, and I look forward to being astonished by these advances.

REFERENCES

Barga, R., Howe, B., Beck, D., Bowers, S., Dobyns, W., Haynes, W., Higdon, R., Howard, C., Roth, C., Stewart, E., Welch, D., and Kolker, E. (2011). Bioinformatics and data-intensive scientific discovery in the beginning of the 21st century. *OMICS* 15, 199–201.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B., and Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature* 471, 51–57.

Curry, A. (2011). Rescue of old data offers lesson for particle physicists. *Science* 331, 694–695.

Franco, M. I., Turin, L., Mershin, A., and Skoulakis, E. M. C. (2011). Molecular vibration-sensing component in *Drosophila melanogaster* olfaction. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3797–3802.

Galbraith, D. W. (2006). The daunting process of MIAME. *Nature* 444, 31.

Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., and Toncheva, A. (2008). *The diverse and exploding digital universe: an updated forecast of worldwide information growth through 2011.* IDC White Paper (EMC Corporation). Available at: http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Gray, J. (2009). "Jim Gray on eScience: a transformed scientific method," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, eds T. Hey, S. Tansley, and K. Tolle (Redmond, WA: Microsoft Research), xvii–xxxi.

Hilbert, M., and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science* 332, 60–65.

Mollick, E. (2006). Establishing Moore's law. *IEEE Ann. Hist. Comput.* 28, 62–75.

Ng, S.-F., Lin, R. C. Y., Laybutt, D. R., Barres, R., Owens, J. A., and Morris, M. J. (2010). Chronic high-fat diet in fathers programs β-cell dysfunction in female rat offspring. *Nature* 467, 963–966.

Wolfe-Simon, F., Switzer Blum, J., Kulp, T. R., Gordon, G. W., Hoeft, S. E., Pett-Ridge, J., Stolz, J. F., Webb, S. M., Weber, P. K., Davies, P. C. W., Anbar, A. D., and Oremland, R. S. (2010). A bacterium that can grow by using arsenic instead of phosphorus. *Science.* doi: 10.1126/science.1197258

Received: 04 May 2011; accepted: 20 May 2011; published online: 08 June 2011.

*Citation: Galbraith DW (2011) Frontiers in genomic assay technologies: the grand challenges in enabling data-intensive biological research. Front. Gene. **2**:26. doi: 10.3389/fgene.2011.00026*

This article was submitted to Frontiers in Genomic Assay Technology, a specialty of Frontiers in Genetics.