JOURNAL OF BACTERIOLOGY, Jan. 2009, p. 20–22 Vol. 191, No. 1
0021-9193/09/$08.00+0 doi:10.1128/JB.01491-08

GUEST COMMENTARY

It Is Computation Time for Bacteriology![▽]

Igor B. Zhulin*

Computer Science & Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37886, and Department of Microbiology, University of Tennessee, Knoxville, Tennessee 37996

Biology is an experimental science. In contrast to other natural sciences, physics, and chemistry, it has a very minor computational component. In order to support this statement with some data, I looked up 100 papers in the area of molecular and cellular biology published in 2008 in the journal *Science* and found that only 3 of them were purely computational. Another 15 were experimental papers with a significant computational component that was employed either to drive the experiment or to extend and interpret its results. Glancing through the *Journal of Bacteriology* (JB), one can see that the percentage of computational papers in this journal is even smaller. There are several reasons. First, this situation reflects the status quo, where experimental disciplines, such as genetics and biochemistry, play a key role in any scientific inquiry yielding worthwhile insights into the biology of the cell; hence the "if it ain't broke, don't fix it" (by bringing in some computational stuff) sentiment. Second, in contrast to the experimental approach, with its century-long history of success, its culture of careful controls, and its army of highly trained professionals, the computational approach is new, its controls (if any) are ill-defined, and it is carried out by a few "aliens" whose culture seems murky to experimentalists. Finally, there might be an increasing concern that traditional in vivo, in situ, and in vitro approaches will soon be replaced with this foreign in silico approach.

Although the worries are legitimate (even if some are clearly exaggerated), there is still something going on with biology that biologists can no longer afford to ignore or to be afraid of. This "something" appears to comprise a fast-growing paradox in biology. It is clear that biology is experiencing an assault by advanced technologies, mathematics, and computing that have already led to such neologisms as "genome biology," "systems biology," "integrative biology," "synthetic biology" and countless types of "-omics." All this novelty is based solely on amazing progress in genome sequencing. We now face a state of affairs where more (in fact, much more) biological data is amenable to computational than experimental analysis. Thousands of sequenced genomes and many more species sampled by environmental sequencing provide materials for computational research. Almost four thousand genomes of various organisms are now available for data mining—see the GOLD database at www.genomesonline.org/ (6). The Human Microbiome Project, a new major National Institutes of Health initiative (www.genome.gov/27528386), will add 1,000 microbial genomes to the collection within this year. Metagenomics is another important player in the drastic increase in the amount of sequenced DNA and the number of genes and proteins in public databases. The Sorcerer II Global Ocean Sampling expedition more than doubled the size of the nucleotide database in 2007 (14). The Human Microbiome Project will also add its large metagenomic portfolio to the cohort of environmental DNA sequencing ventures. Fantastic developments in the new sequencing approaches and technologies, such as whole-genome amplification from a single cell (9, 10) and nanopore sequencing (1), may soon lead to a situation where most genomes in our hands will be from "the unseen majority," uncultivable microorganisms that we will probably never study in experiments. The other side of the paradox is that currently 99% of microbiologists, who are experimentalists, study only a handful of species representing a tiny fraction of Earth inhabitants. A vast amount of experimental knowledge in microbiology is being generated on a few models, such as *Escherichia coli* and *Bacillus subtilis*, and human pathogens, such as *Staphylococcus aureus* and *Pseudomonas aeruginosa* (ironically, one needs computational tools implemented in Google and PubMed searches to obtain accurate ranking of microorganisms according to the frequency of their "usage"). Will this paradoxical situation change? Not dramatically. Experiment will never be fully replaced by computation as a primary tool for gathering in-depth information about the cell; however, computation will play an increasing role in (i) extrapolating the knowledge obtained on a few model organisms to the entire genomic landscape and (ii) piecing together fragmental experimental knowledge in order to obtain a more complete picture of specific functions and eventually of the entire cell. These two key areas of computation are known as computational genomics and systems biology. Both these new scientific disciplines, which comprise a core of modern computational biology, will also play an important role in guiding future experiments and in linking biological scales from molecules to ecosystems (11).

JB has decided to welcome the "new (computational) wave" by launching a new section of the journal—Computational Biology. This is a bold move taken by the JB editors, who feel that the time is right to open the journal to high-quality papers that employ computation to gain insights into novel biological phenomena and mechanisms. Is the JB readership going to experience a cultural shock? Perhaps a minor one. Yes, there will be phylogenetic trees (which the JB readers are already

* Mailing address: Department of Microbiology, University of Tennessee, M409 Walters Life Sciences Bldg., Knoxville, TN 37996. Phone: (865) 974-7687. Fax: (865) 974-4007. E-mail: ijouline@utk.edu.
▽ Published ahead of print on 31 October 2008.

used to) and even some mathematical formulas (oh, no!). Not everything will be understood by experimentalists who do not have serious training in mathematics or physics, but let us look at the bright side—this might become a wonderful learning ground for those who are interested in hearing more about computation and the new possibilities it can offer. So, let us welcome computational "aliens" to the journal! Computational biologists have already formed a successful community of their own, and "their" journals have earned respectable impact factors. They are a force one needs to recognize.

As a former experimentalist turned computational scientist, I do not share the popular view that the new generation of biologists should be equally well versed in computation and experiment. This would require twice as much learning time and a split personality. It is a reality of today and tomorrow that there are and will be two camps, experimental and computational biologists, and they should try to learn each other's language and embrace each other's culture, at least to a certain degree. It is like a vacation in Paris: one should try asking for directions to the Louvre in French after enjoying coffee and pain-au-chocolat on the street terrace, instead of complaining that Parisians are rude and there is no Wal-Mart. Hopefully, the Computational Biology section of the journal can become such a learning ground, especially for graduate students and postdoctoral researchers in experimental labs who might spend some time digging into computational materials and methods. Similarly, computational biologists who otherwise would not read experimental papers published in JB might do so just because they are interested in a computational biology paper on the next page.

Computational biology papers in JB will be held to the same standard as experimental papers. They should provide sufficient details in methodology and approach so that results can be reproduced by others. They should also address a fundamental biological problem or mechanism. If a computational paper describes a novel computational method or approach, it should clearly demonstrate its benefit to microbiology by applying it to a concrete biological problem. Experimentalists often complain that they have to "believe" conclusions of a computational paper rather than understand them. One way to solve this problem, at least partially, is to have a balanced peer review process. Each paper will be reviewed by specialists in the paper's research area (e.g., computational genomics), which will ensure consensus on the methodology, approach, and conclusions. On the other hand, having at least one reviewer from the experimental biology area could provide an independent view on the value of presented findings to the noncomputational majority. Computational biologists also have their share of skepticism when it comes to experimental biology, especially when experimentalists use computational tools. There are examples of apparently wrong conclusions being made first by improper use of computational tools and then, surprisingly, being confirmed by experimental validation (4). Another tendency is that many graduate students in computational biology who never carried out any biological experiments hold an almost religious belief that a biological experiment is a truth in and of itself. "Why do you question the validity of this statement? They have shown this in experiments!" I have heard this from graduate students more than once. The bottom line is that we have a long way to go to see a happy marriage of computation and experiment, but mutual sympathy, flirting, and perhaps even an engagement could be our short-term goals.

The inaugural Computational Biology section in this issue of JB features several papers from leading computational laboratories that highlight some but certainly not all areas of computational biology. A minireview by Julio Collado-Vides (Universidad Nacional Autonoma de Mexico, Morelos, Mexico) (2) summarizes recent advances in the development of the popular database RegulonDB and its use by experimental microbiologists. A report from the Peer Bork laboratory (EMBL, Heidelberg, Germany) (13) is an example of how biological information can be retrieved from metagenomic datasets. This is a pioneering study in this area, which carries a trademark of one of the best comparative genomics groups in the world. Those who are interested in metagenomics should pay close attention not only to the novel and exciting findings but also to the caveats of metagenomics data mining that are well defined in this work.

What should we be sequencing now—the genomes of closely related species and strains, or representatives of distant phyla? This question was raised at several funding agency panels and workshops after the first bacterial genomes were sequenced. All sorts of arguments were presented for and against each of these suggestions. The paper from Eugene Koonin's group (National Institutes of Health, Bethesda, MD) (8) puts these arguments to rest (although it seems as we will be sequencing every piece of DNA we can get our hands on anyway). His and Peer Bork's previous work from the beginning of the genome era demonstrated the usefulness of making long-distance evolutionary comparisons to decipher novel biological phenomena. In their current paper, Koonin and his colleagues show the value of comparative analysis of closely related genomes.

Comparative genomics is also represented in this issue by the paper from the laboratories of Mikhail Gelfand and Adam Arkin (5) revealing the extraordinary variability of the regulatory systems associated with amino acid utilization in proteobacteria. In a remarkable example of constructive collaboration between computational and experimental biologists, Dmitry Rodionov and his European and American colleagues present their findings of a novel class of transporters in prokaryotes (12). A paper from the Marie-Agnès Petit laboratory (INRA, Jouy en Josas, France) (3) explores discontinuity between bacterial species and genera using bioinformatics methodology; systems biology is represented by a paper from Daniel Segrè's group (Boston University, Boston, MA) (7) describing a metabolic network model of a human oral pathogen. Thus, the inaugural Computational Biology section of JB presents a fairly broad view of the subject. Computational Biology is here, and it is time.

ACKNOWLEDGMENTS

I thank my numerous colleagues in both the experimental and computational communities for sharing their views on the issue.

Computational genomics work in my laboratory is funded by National Institutes of Health and the U.S. Department of Energy.

REFERENCES

1. **Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V.**

Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. 2008. The potential and challenges of nanopore sequencing. Nat. Biotechnol. **26:**1146–1153.

2. **Collado-Vides, J., H. Salgado, E. Morett, S. Gama-Castro, V. Jiménez-Jacinto, I. Martínez-Flores, A. Medina-Rivera, L. Muñiz-Rascado, M. Peralta-Gil, and A. Santos-Zavaleta.** 2009. Bioinformatics resources for the study of gene regulation in bacteria. J. Bacteriol. **191:**23–31.
3. **Deloger, M., M. El Karoui, and M.-A. Petit.** 2009. A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. J. Bacteriol. **191:**91–99.
4. **Iyer, L. M., L. Aravind, P. Bork, K. Hofmann, A. R. Mushegian, I. B. Zhulin, and E. V. Koonin.** 2001. Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol. *2:*RESEARCH0051.
5. **Kazakov, A. E., D. A. Rodionov, E. Alm, A. P. Arkin, I. Dubchak, and M. S. Gelfand.** 2009. Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. J. Bacteriol. **191:** 52–64.
6. **Liolios, K., K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides.** 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. **36:**D475–D479.
7. **Mazumdar, V., E. S. Snitkin, S. Amar, and D. Segrè.** 2009. Metabolic network model of a human oral pathogen. J. Bacteriol. **191:**74–90.
8. **Novichkov, P. S., Y. I. Wolf, I. Dubchak, and E. V. Koonin.** 2009. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J. Bacteriol. **191:**65–73.
9. **Pan, X., A. E. Urban, D. Palejev, V. Schulz, F. Grubert, Y. Hu, M. Snyder, and S. M. Weissman.** 2008. A procedure for highly specific, sensitive, and unbiased whole-genome amplification. Proc. Natl. Acad. Sci. USA **105:** 15499–15504.
10. **Podar, M., C. B. Abulencia, M. Walcher, D. Hutchison, K. Zengler, J. A. Garcia, T. Holland, D. Cotton, L. Hauser, and M. Keller.** 2007. Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl. Environ. Microbiol. **73:**3205–3214.
11. **Raes, J., and P. Bork.** 2008. Molecular eco-systems biology: towards an understanding of community function. Nat. Rev. Microbiol. **6:**693–699.
12. **Rodionov, D. A., P. Hebbeln, A. Eudes, J. ter Beek, I. A.. Rodionova, G. B. Erkens, D. J. Slotboom, M. S. Gelfand, A. L. Osterman, A. D. Hanson, and T. Eitinger.** 2009. A novel class of modular transporters for vitamins in prokaryotes. J. Bacteriol. **191:**42–51.
13. **Singh, A. H., T. Doerks, I. Letunic, J. Raes, and P. Bork.** 2009. Discovering functional novelty in metagenomes: examples from light-mediated processes. J. Bacteriol. **191:**32–41.
14. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. **5:**e16.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.