Late-Night Thoughts on the Sequence Annotation Problem

Sarah J. Wheelan and Mark S. Boguski

Genome Res. 1998 8: 168-169
Access the most recent version at doi:10.1101/gr.8.3.168

Late-Night Thoughts on the Sequence Annotation Problem

Sarah J. Wheelan[1,2] and Mark S. Boguski[1,3]

[1]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA; [2]Department of Molecular Biology and Genetics, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205 USA

The reader of James Joyce's *Portrait of the Artist as a Young Man* (1992) is aided by editor's notes illuminating the meaning of unfamiliar words, for example, ''greaves in his number'' means simply ''shinguards in his locker'' and ''a cod'' is a joke or prank. Such minimal explanatory notes are essential, especially for a novice Joyce reader; however, overly detailed comments can be misleading and can stifle the reader's own interpretation of the work.

During the current scale-up phase of human genome sequencing, production groups have been experimenting with various types and levels of annotation. Precedents for the biological annotation of sequence records in GenBank, however, come from qualitatively and quantitatively different types of sequences from those currently being produced. Historically, the first type of sequence record is that of the ''functionally cloned'' gene, which is the end product of often years of investigation that began with a particular biological problem in mind. There is usually a one-to-one correspondence between these records and peer-reviewed publications. The second type of sequence record might be described as the result of a population study where many isolates of a particular gene are determined for the purpose of detecting and interpreting variations. Examples include ribosomal genes used to study molecular phylogeny, HIV sequences used to study antigenic variation, and, most recently, copies of human genes from different individuals used to detect sequence polymorphisms for the development of genetic markers. For this second class of sequence data, a multiple alignment is often the most meaningful and appropriate type of annotation. Literature citations to published articles also usually accompany this type of database record. A third major class of GenBank sequences consists of single-pass expressed sequence tags (ESTs). The most important annotations on these records are organism, tissue of origin, and cloning vector used. Other features are strictly computed and there are no publications relating to the specific nature of individual sequences.

None of the traditional forms of annotation is a good model for the high throughput genomic sequence (HTGS) data now being produced. We will argue that the only annotations that are essential for HTGS data in a public database archive, apart from identification of the contributing laboratory, are the source organism from which the DNA was obtained and a confidence value or accuracy assessment of each base. Virtually everything else is computable on demand and/or quickly becomes obsolete. This is not to say that sequence exegesis is not useful or important; however, there will be many interpretations of the rich literature of the sequences of genomes and these interpretations will change over time. Resultant ''annotations'' will benefit from a publication modality that includes some version of the traditional peer review process for quality assurance.

There are two broad categories of annotation that have been applied to HTGS data: (1) the results of computations, and (2) the results of experiments. What are some of the disadvantages of applying these types of detailed annotation to the archival reference sequence? It is obvious that the results of sequence similarity searches, particularly matches against ESTs, become out of date almost immediately and can easily and efficiently be recomputed daily, or on demand, by automated systems within an individual's laboratory or from web-based facilities (Boguski and McEntyre 1994). There is also the problem that Randy Smith has referred to as ''transitive annotation'', whereby chains of inferences with weak links can lead to misleading or completely erroneous sequence interpretation (Smith 1996).

The science of gene prediction based on intrinsic sequence properties is still quite fallible in producing accurate models of complete genes (Burset and Guigo 1996). Application of these methods has already had the insidious effect of populating the protein databases with conceptual translations of protein sequences that are partially right and partially wrong. Experimental data, such as the determination of full-length cDNA sequences for ESTs matching a genomic region, has also been used to annotate HTGS; however, the main disadvantages of any experimental validation of sequence features is added cost and potentially long delays in the submission of ''finished'' sequence. Even computer-based annotation alone has these effects on the ''bottom line'' and should be subjected to cost–benefit analysis.

Because we are suggesting continual-update and compute-on-demand approaches to sequence annotation, is it realistic to expect that computational power will be sufficient to the tasks? It is illuminating in this context to look back exactly a decade ago to a 1988 article by Charles DeLisi (1988) concerned with emerging trends in computational biology. DeLisi [who, by the way, spearheaded the Human Genome Project under the auspices of the Department of Energy (Cook-Deegan 1994)], predicted that anticipated advances in computer speed would be unable to keep up with the growing sequence database and the demand for homology searches of the

[3]Corresponding author.
E-MAIL boguski@ncbi.nlm.nih.gov; FAX (301) 480-9241.

8:168–169 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

data. In 1998, the reality is that NCBI's BLAST service efficiently processes 40,000 similarity searches of the >1 billion nucleotides in GenBank every day (including computationally intensive six-frame translation searches) on commodity compute servers. Moore's Law (which, loosely speaking, guarantees a doubling in computer power every 18 months) has held since the mid-1960s, and has been kind to us all. But how long can this trend continue? Nathan Myhrvold, the chief technology officer for Microsoft Corporation, in a March 20, 1997, speech at NIH, is counting on another 40 years, and Gordon Moore himself projects that another two human generations will enjoy this doubling rate of computer power (Leyden 1997). Hence, by 2005, we'll have ~30 times more power to search a database (the human genome sequence) that is only about three times larger than the total contents of GenBank today. This is an oversimplification, of course. Nevertheless, the last decade has shown us that advances in algorithms, data organization, and biological understanding, combined with hardware improvements, have, thankfully, acted in concert to prevent DeLisi's dire divination from coming true.

In a speech at the Cold Spring Harbor meeting on Genome Mapping and Sequencing in 1995, Maynard Olson opined that the final, irreducible products of the Human Genome Project would be three: (1) the genetic map, (2) the sequence, and (3) the peer-reviewed scientific literature (Boguski 1995). We tend to agree. But how can the literature be used most effectively to annotate a sequence? Here the responsibility lies with individual investigators and journal editors who publish information that can be related to a reference genomic sequence. Authors should be encouraged to refer experimental and computational results to specific coordinates on the reference sequence. There is ample precedent for this and every expectation of success. One already sees specific communities of investigators attempting to simplify and organize sequence-based knowledge in their own areas of expertise by agreeing on nomenclature and referring it back to specific accession numbers of GenBank records (Adachi et al. 1996; Derynck et al. 1996). This procedure can be carried out both prospectively, for new publications, as well as retrospectively, to update and refresh annotation on traditional, ''functionally cloned'' GenBank sequences.

In the past, common objections to using the literature as annotation include the fact that only abstracts have been electronically available and that even in complete articles the text is unstructured and not easily parsed by machines. However, the advent and future development of electronic publication of full-text journals on the World Wide Web will eliminate these barriers. With a little creativity and discipline on the part of editors and a little support from publishers, any relevant piece of information can be linked back to specific nucleotide coordinates on a reference sequence and vice versa.

In summary, we end with a modest proposal that the solution to the annotation problem is to eliminate most archived annotation. The only essential features of HTGSs are the contributing laboratory, the biological species, a quality measure for each base, and clone or mapping information that permits the assembly of large contigs. Interpretations of the sequence, by the producers or consumers alike, may be published separately and not necessarily as part of the reference sequence archive. In this manner, 21st-century biologists will be able to access the human blueprint, to ''see that it is that thing which it is and no other thing'' (Joyce 1992) and to interpret it as they may.

ACKNOWLEDGMENTS

We thank M.V. Olson for stimulating discussions and B.F.F. Ouellette for helpful comments and a critical reading of the manuscript. The ideas expressed here are personal and do not necessarily reflect the views of GenBank, NCBI, or *Genome Research*.

REFERENCES

Adachi, M., E.H. Fischer, J. Ihle, K. Imai, F. Jirik, B. Neel, T. Pawson, S. Shen, M. Thomas, A. Ullrich, and Z. Zhao. 1996. *Cell* 85: 15.

Boguski, M.S. 1995. *Trends Biochem. Sci.* 20: 295–296.

Boguski, M. and J. McEntyre. 1994. *Trends Biochem. Sci.* 19: 71.

Burset, M. and R. Guigo. 1996. *Genomics* 34: 353–367.

Cook-Deegan, R. 1994. *The gene wars: Science, politics, and the human genome*, pp. 92–106. W.W. Norton and Co., New York, NY.

DeLisi, C. 1988. *Science* 240: 47–52.

Derynck, R., W.M. Gelbart, R.M. Harland, C.H. Heldin, S.E. Kern, J. Massagué, D.A. Melton, M. Mlodzik, R.W. Padgett, A.B. Roberts et al. 1996. *Cell* 87: 173.

Joyce, J. 1992 (orig. 1916). *A portrait of the artist as a young man*. Bantam Books, New York, NY.

Leyden, P. 1997. Moore's Law repealed, sort of. *Wired* (May, p. 166).

Smith, R.F. 1996. *Genome Res.* 6: 653–660.