Opinion

Genome re-annotation: a wiki solution?

Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology and Department of Computer Science, 3125 Biomolecular Sciences Building, University of Maryland, College Park, MD 20742, USA. Email: salzberg@umiacs.umd.edu

Published: 1 February 2007

Genome ***Biology*** 2007, **8:**102 (doi:10.1186/gb-2007-8-1-102)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/1/102

Abstract

The annotation of most genomes becomes outdated over time, owing in part to our ever-improving knowledge of genomes and in part to improvements in bioinformatics software. Unfortunately, annotation is rarely if ever updated and resources to support routine reannotation are scarce. Wiki software, which would allow many scientists to edit each genome's annotation, offers one possible solution.

So you think that gene you just retrieved from GenBank [1] is correct? Are you certain? If it is a eukaryotic gene, and especially if it is from an unfinished genome, there is a pretty good chance that the amino acid sequence is wrong. And depending on when the genome was sequenced and annotated, there is a chance that the description of its function is wrong too.

Large-scale genome sequencing has revolutionized biology over the past ten years, generating vast amounts of new information that has radically transformed our understanding of hundreds of species, including ourselves. Sequencing centers continue to churn out new DNA sequences for a fantastic variety of species, covering more and more of the tree of life. Along with these sequences, the centers also produce genome annotation, which includes the locations and descriptions of all identifiable genes. These gene lists are the first pictures we get of what's inside a newly sequenced genome, and they can reveal key insights into what makes an organism distinctive. Sometimes the gene lists themselves are part of the story; for example, when the human genome was published [2,3], the headline was that humans have 'only' 25,000 genes, in contrast to earlier estimates of 100,000 or more. For many microbial species, the genome helps us to understand how the organism can accomplish something particularly difficult, such as how *Deinococcus radiodurans* (to cite just one of many examples) can withstand exposure to radiation levels far in excess of what a human could tolerate [4]. With each new human pathogen, the gene list helps us determine how the organism infects humans, how it causes sickness and (sometimes) how it becomes resistant to antibiotics. For these and other reasons, the accuracy of the gene list is tremendously important.

What is genome annotation?

Before addressing the problems with annotation, I will first summarize how it is done. The process of sequencing and annotating the DNA of a bacterial species has become highly automated in recent years, but the major steps are quite similar to what was done for the very first bacterial genome, *Haemophilus influenzae*, in 1995 [5].

Figure 1 shows an outline of the main steps of the whole-genome shotgun sequencing and annotation process for a bacterial genome. Similar procedures - for both sequencing and annotation - are followed for much larger genomes, including the human genome, although the details vary. The laboratory steps have not changed greatly since *H. influenzae*: they begin with DNA purification, followed by shearing the DNA into countless small fragments (the 'shotgun' step). These fragments are then cloned and sequenced from both ends and assembled, usually resulting in a set of contiguous DNA sequences (contigs) joined together into larger scaffolds. The annotation pipeline can be applied immediately to these contigs, but in projects where the genome will be finished, the annotation software is usually run later, when the gaps between contigs have been filled in.

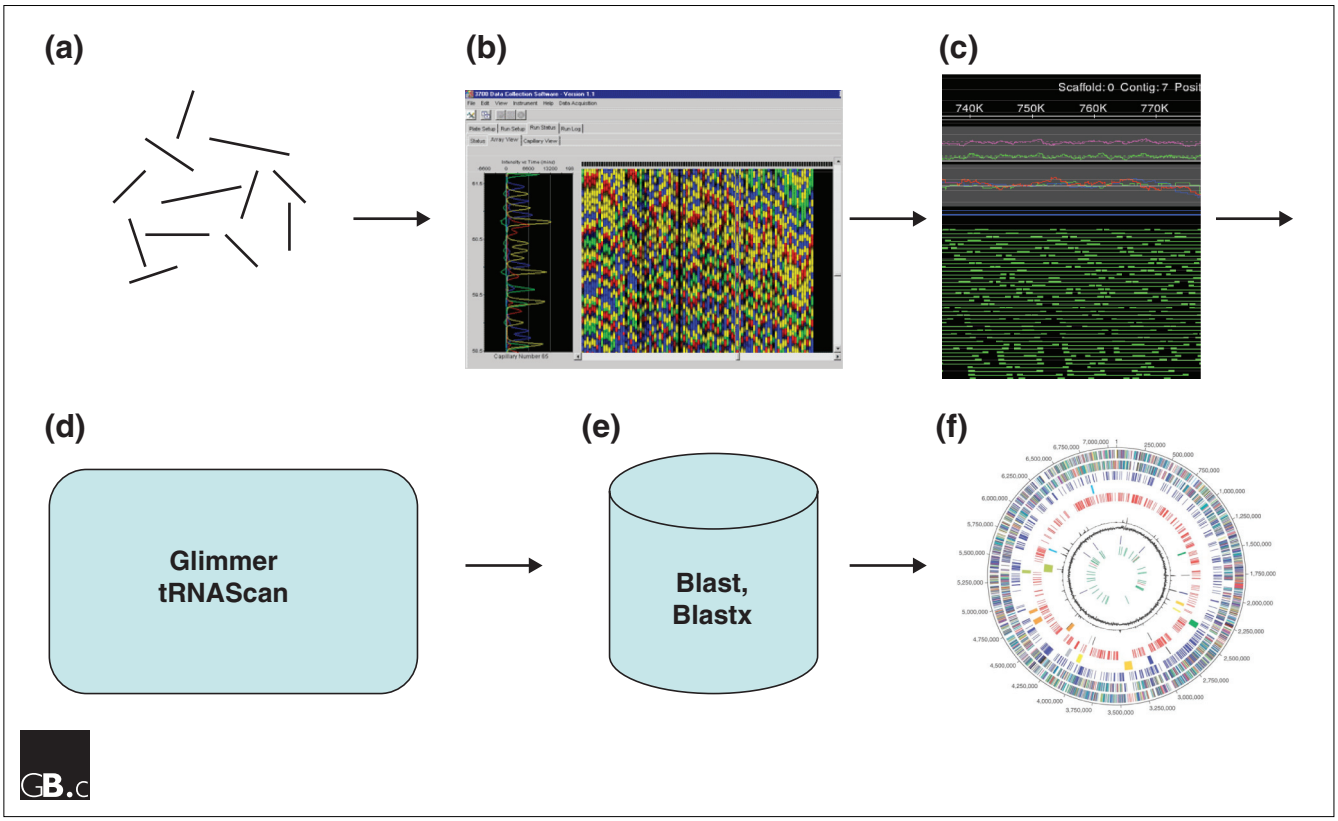

Figure 1
Overview of sequencing and annotation for a whole-genome shotgun project, for example, sequencing a bacterial genome. First **(a)**, genomic DNA is purified, broken into short fragments and cloned into *E. coli*. The cloned fragments are then sequenced from both ends on an automated sequencing machine. The resulting sequences (shown in **(b)** as they appear on the sequencing machine display) are then assembled using a complex software program that identifies overlaps into **(c)** large, contiguous sequences representing the chromosomes from the original DNA. Gaps are filled until the genome is complete. **(d)** Annotation begins with the execution of several gene-finding programs, such as Glimmer, which identifies protein-coding genes, tRNAScan, which identifies tRNAs, and other programs for other genome features. **(e)** These initial predictions are used as the basis for BLAST searches against large protein databases, which identify related proteins based on sequence similarity. Translated (Blastx) searches are then used to scan the databases to detect any proteins that match the DNA regions in between predicted genes. Customized annotation programs are used to decide what name and function to assign to each protein, leading to **(f)** the final annotated genome.

Most annotation pipelines are considerably more complex than shown in Figure 1, but they share the same outline. First a gene finder (such as, for bacteria, Glimmer [6] or GeneMark [7]) is run over the genome, producing a set of predicted protein-coding genes. These programs are very accurate, though not perfect. They are far more accurate than eukaryotic gene finders, however, primarily because the problem is far more difficult in eukaryotic genomes. In either case, the next step in the pipeline is to take the set of predictions and search them against one or more protein databases using BLAST [8], HMMer [9] or other programs. For each gene that has a significant match, the BLAST output can be used to assign a name and function to the protein. The accuracy of this step depends not only on the annotation software, but also on the quality of the annotations already in the database. For genes with no match, the pipeline might keep them and label them as 'hypothetical', or it might discard them based on criteria as simple as minimum length.

Annotation pipelines also run separate searches for tRNA and rRNA genes, and they may include other components as well. The pipeline software will usually take extra steps to find any genes missed by earlier steps; typically this involves running a translated search, aligning all six possible translations of the unannotated sections to a database.

Partial and draft genomes

Finishing a genome - sequencing every remaining nucleotide of every chromosome and creating a gap-free assembly - is considerably slower and more expensive than the high-throughput shotgun-sequencing phase. As a result, a growing number of genomes are being released in 'draft' form and will remain in this form indefinitely. These include many bacteria and the majority of eukaryotic genomes. (In fact, only a handful of eukaryotic genomes, such as those of *Saccharomyces cerevisiae* and *Caenorhabditis elegans* but not including the human, are truly finished.)

The effect of draft genomes upon annotation is considerable: many genes will 'run off' the end of contigs or appear on two or more separate contigs. This in turn complicates the subsequent steps of annotation and is likely to lead to additional errors in assigning gene function. For example, a gene fragment is liable to match a small protein domain, and functions based on a single domain hit are not reliable. A gene that is split across two contigs might be annotated twice. Draft sequences also have much higher sequencing error rates, which can introduce erroneous stop codons in the middle of genes or improperly merge adjacent but distinct genes.

The role of GenBank

Once a genome - draft or complete - is annotated, the DNA sequence along with the annotation is normally deposited in GenBank. Countless researchers rely on GenBank [1], EMBL [10] and DDBJ [11] (which mirror one another) as their primary source for genome annotation, and for a good reason: these databases are the world's largest public repositories of genome information. GenBank now contains over 65 billion base pairs (Gbp) of sequence, up from just 2 Gbp in 1998 and 10 Gbp in 2000, and it continues to grow at an astonishing rate. If you want to find a gene, GenBank should definitely be your first stop. Yet I frequently hear claims within the bioinformatics community that the 'GenBank annotation' of a particular genome is fraught with problems, and that the speaker can fix them.

Is the GenBank annotation perfect? Of course not. How good it is, though, depends on many variables, and the consumer of GenBank data would be wise to be aware of them (in other words, *caveat emptor*). The first and most important point to understand is that GenBank is not simply a database; it is also a library. A scientist who submits a sequence to GenBank is the owner of that sequence and is listed on the 'author' line in the GenBank entry. Just as with any article published in a journal, the author (and only the author) has the right to submit an erratum. Because GenBank is an electronic library, an erratum is really an update: new sequences or annotations replace the old ones, although GenBank keeps a record of the changes so that the original entry can still be retrieved if necessary. This notion of GenBank as a library (or an electronic journal) is frequently misunderstood, especially when a scientist discovers an annotation error. Even if the error is overwhelmingly obvious, the custodians of GenBank cannot simply fix it, any more than the editor of a journal can correct one of the papers published in that journal. Another way to think of this is to recognize that a 'GenBank annotation' is not 'GenBank's' annotation, but rather the annotation of whoever deposited the sequence in the first place.

When confronted with this problem, some scientists react by suggesting that GenBank (and DDBJ and EMBL) should allow scientists to fix errors that they find. But this would quickly destroy the archival function of GenBank, as original entries would be erased over time. It would also violate the agreement that GenBank has with all its submitters that their entries belong to them and can only be changed by them. This agreement has been crucial in GenBank's near-universal acceptance by the genomics community as the central resource for DNA sequences. The idea of allowing others to alter GenBank annotation also immediately begs the question of who should be permitted to make such alterations.

This leaves us with a problem: users go to GenBank expecting to find the authoritative annotation for a genome, and what they find might be far less than that. Most genome annotation deposited in GenBank remains static for years, and many annotations have never been changed since their initial publication. Nonetheless, many scientists assume that GenBank annotation is kept up to date, and they are surprised to hear that it is not.

For example, 479 genes in the *H. influenzae* Rd genome are currently listed as hypothetical proteins. Of these, 217 have at least one extremely strong BLAST hit to another species (*E*-value < 10-100), which means they should at least be called 'conserved hypothetical' proteins. And 40 of these have matches to a gene with an assigned function, meaning that a re-annotation would result in these genes having a more meaningful name than 'hypothetical protein'.

Some inconvenient truths

Even considering all of the issues above, one might reasonably expect that as protein databases have grown, annotation has improved and that recently annotated genomes (at least) will be of the highest quality. This is not quite true. What is true is that a BLAST search of a protein that is run today will yield far more results than it would have five or ten years ago, and these results in turn should lead to better annotation. Not all software is equally good, however, and the annotation pipelines vary considerably in their quality. There is also wide variation in the skills and experience of those operating the pipelines. Further complicating matters, some genomes are subjected to careful curation and review, whereas others receive only automated annotation. In the early days of sequencing, the sequencing teams included experts on the biology of each genome, and their manual curation dramatically improved the annotation of those species. Today that is no longer true: high-throughput sequencing centers are large, efficient factories with unique expertise in the methods necessary for sequencing, but they sometimes have very little expertise on the biology of the species they are sequencing. The inconvenient truth is that, as a result of these factors and others, some genomes are poorly annotated even today.

There are several ways in which genome annotation can be erroneous. The first and most fundamental is simply that the

gene models may be wrong. Although bacterial gene-finding systems [6,7] are highly accurate, finding 98-99% of protein-coding genes in most species, they still occasionally miss genes. Their accuracy at placing the start site is a bit lower, probably closer to 90%, which is excellent but far from the perfect accuracy that some might expect. In the past, the accuracy of (bacterial) start-site prediction was closer to 80%, and many of the genomes in GenBank were predicted with earlier versions of gene finders. Note that all these accuracy figures are much lower for eukaryotic annotation. Some annotation pipelines include algorithms to adjust start sites, which can be done by looking closely at the boundaries of alignments to homologous proteins.

False positives represent another type of erroneous annotation: when the prediction of a gene-finding program does not match any previously known protein, the annotators (or the annotation pipeline software) must decide whether or not to include that prediction in the gene list. Over the years, annotation groups have used a variety of rules to make this decision, and they have inevitably included thousands of false predictions in the publicly available genome annotation. These predictions are mostly harmless unless they result in effort being expended trying to verify them. In some cases, too, they might 'hide' functional RNA genes or true genes in a different reading frame from that of the false prediction.

Perhaps the biggest problem with genome annotation is erroneous and inconsistent naming of genes. Much of this is due to the simple fact that our knowledge of genes has improved but the annotation has remained static. Thus a gene labeled 'hypothetical protein' a few years ago might now have a known function. A second problem is what's known as transitive catastrophe: the phenomenon whereby a name is transferred from one gene to another on the basis of sequence similarity (usually from a BLAST search) but where the original name is incorrect. As more genomes are annotated, and more BLAST searches are run, the name gets transferred to other proteins, and the original source of the name quickly becomes lost. It is well known in the genomics community that thousands of such transitive errors have propagated through sequence databases, and efforts are under way to try to clean up some of the mess. In the meantime, though, many genes remain incorrectly annotated.

Let us consider just one example, selected more or less at random from the bacterium *H. influenzae* Rd [4]. The gene *fdxH* encodes formate dehydrogenase, β subunit, GenBank accession number NP438180. When the genome was sequenced in 1995, this gene (encoding a 312 amino acid protein) was similar to very few other genes; even the orthologous *Escherichia coli* gene was not yet sequenced. It is very difficult today to reconstruct what the best BLAST hit was back then, but today there are 197 highly significant BLAST hits to 123 distinct species. Thus, it is pretty clear that this gene today should be well-annotated because of the multitude of highly similar proteins. Yet if we look at the list of matching proteins, we find a variety of names given, including not only the name found on NP438180 itself, but also: formate dehydrogenase-O β subunit; formate dehydrogenase, nitrate-inducible, iron-sulfur subunit; HybA protein; formate dehydrogenase-N, Fe-S β subunit, nitrate-inducible; hypothetical protein PaerPA_01004979; hypothetical protein Bpse11_03005113; 4Fe-4S ferredoxin, iron-sulfur binding; and Twin-arginine translocation pathway signal. Some of these names seem to be synonymous, but others clearly are not. To decide properly among them, we need to look at the source of each annotation and at the species to which it is attached.

Possible solutions

So if we can't always trust GenBank, what can we do? Clearly we cannot just ignore it. The scientific community must have a resource that contains the genes from all the species that have been sequenced. For the past 25 years, GenBank, EMBL and DDBJ have been enormously successful at providing these data. The pace of sequencing has changed the rules of the game, however: sequencing centers are pouring out genomes, annotating them rapidly and moving on. An archive of these annotations may be useful, but a static archive is insufficient.

One part of the solution is obvious: annotation must be regularly re-computed using the latest databases and software. For a small number of model organisms, this is already happening, but these species represent a tiny proportion of all known genes. Simply re-running an automated pipeline on all genomes is not sufficient, though, because that would over-write many of the carefully curated, manually annotated genes that have been produced in the past. Unfortunately, there is no standard label attached to such genes, so there is no way for an automated pipeline to know that they should be trusted. Therefore, we also need to launch an effort to start identifying those genes that are well annotated and, beyond that, to start recording the evidence used to annotate each gene.

Another solution is to create a new, expanded database that can display all the alternative annotations for any locus in a genome. If this were available, then scientists could be provided with links from any gene to alternative or overlapping gene predictions as well as alternative gene names. Along with each annotation could be a link to the evidence supporting it; for example, the date of a BLAST search or a citation to experiments contained in a journal article.

A wiki solution?

Various members of the genomics community have considered these and other solutions, but so far none have emerged as the standard. Several new databases have been

developed with alternative genome annotation, or with re-annotation (for example, the TIGR Comprehensive Microbial Resource [12]), but none of them has attracted nearly as much web traffic as GenBank or the other databases at NCBI [13]. The difficulties in changing this system are many: first, for example, there are some genomes for which GenBank is still the best source, and second, if another, better source of annotation exists, how is someone to discover it?

A relatively new model of sharing expertise through the Internet might offer a solution. This model is the 'wiki': a shared resource that anyone can edit. This open-editing framework for websites and data was first introduced in 1995, and it was initially viewed with skepticism by many in the Internet community, who argued that wiki-based websites would be filled with unreliable, inaccurate information. But the success of the online encyclopedia Wikipedia [14] has demonstrated that, despite the skeptics, a wiki site can be accurate, up-to-date and incredibly useful. Genome annotation has many of the same features of an encyclopedia: the information required to produce it is broad-based and the expertise is scattered around the scientific community in a very wide range of laboratories, most of whom are not connected to genome projects. I therefore propose that a 'genome wiki' might provide just the solution we need for genome annotation. A wiki would allow the community of experts to work out the best name for each gene, to indicate uncertainty where appropriate and to discuss alternative annotations. Although wikis will not (and should not) supplant well-curated model-organism databases, for the majority of species they might represent our best chance for creating accurate, up-to-date genome annotation.

Whether or not a genome wiki emerges, we will probably need an archival repository of annotation for many years to come. The international database consortium represented by GenBank, EMBL and DDBJ has served that purpose remarkably well for a long time and will continue to do so. Despite this success, the genomics community needs an accurate, continually updated source of genome annotation for every species, and we can hope that a solution to this problem will emerge in the near future.

References

1. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
2. The International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
4. White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, *et al.*: **Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1.** *Science* 1999, **286:**1571-1577.
5. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269:**496-512.
6. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER.** *Nucleic Acids Res* 1999, **27:**4636-4641.
7. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding.** *Nucleic Acids Res* 1998, **26:**1107-1115.
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
9. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
10. **EMBL Nucleotide Sequence Database** [http://www.ebi.ac.uk/embl/]
11. **DNA Data Bank of Japan** [http://www.ddbj.nig.ac.jp/]
12. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O: **The Comprehensive Microbial Resource.** *Nucleic Acids Res* 2001, **29:** 123-125.
13. **National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
14. **Wikipedia** [www.wikipedia.org]