Mini-Review

Correspondence
Maria S. Poptsova
maria.poptsova@gmail.com

Using comparative genome analysis to identify problems in annotated microbial genomes

Maria S. Poptsova and J. Peter Gogarten

Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3125, USA

Genome annotation is a tedious task that is mostly done by automated methods; however, the accuracy of these approaches has been questioned since the beginning of the sequencing era. Genome annotation is a multilevel process, and errors can emerge at different stages: during sequencing, as a result of gene-calling procedures, and in the process of assigning gene functions. Missed or wrongly annotated genes differentially impact different types of analyses. Here we discuss and demonstrate how the methods of comparative genome analysis can refine annotations by locating missing orthologues. We also discuss possible reasons for errors and show that the second-generation annotation systems, which combine multiple gene-calling programs with similarity-based methods, perform much better than the first annotation tools. Since old errors may propagate to the newly sequenced genomes, we emphasize that the problem of continuously updating popular public databases is an urgent and unresolved one. Due to the progress in genome-sequencing technologies, automated annotation techniques will remain the main approach in the future. Researchers need to be aware of the existing errors in the annotation of even well-studied genomes, such as *Escherichia coli*, and consider additional quality control for their results.

Introduction

Sequencing the first complete genome of *Haemophilus influenzae* in 1995 opened a new page in genome sciences. It took eight more years (till 2003) to increase the number of complete sequenced genomes to 100. This number was doubled by the year 2005, and by 2010 more than 1000 completely sequenced bacterial and archaeal genomes were available at GenBank, with approximately four times this number of genomes in the process of being sequenced. In September of 2009 the GOLD database (Liolios *et al.*, 2008) listed 5902 genome projects; 4242 of these were bacterial genome projects, of which 1154 were listed as complete and 966 in draft. With such a tempo it is obvious that the burden of genome annotation will be assigned mostly to automated methods. However, the accuracy of automated approaches has been questioned since the beginning of the sequencing era (Friedberg, 2006).

Genome annotation is a multi-level process that includes prediction not just of coding genes, but also of pseudo-genes, promoter regions, direct and inverted repeats, untranslated regions and other genome units. For a comprehensive review of genome and proteome annotation

see Reed *et al.* (2006) and Reeves *et al.* (2009). In this paper we briefly review the problems associated with identification of coding sequences (CDS) in bacterial and archaeal genomes and demonstrate how comparative genomics can help in the location of missed genes.

Bacterial and archaeal genomes, as well as those of some eukaryotic micro-organisms, have the considerable advantage of usually lacking introns, which makes the process of gene boundary identification much easier. Nevertheless bacterial or archaeal gene-calling procedures are not error free. In the absence of introns, it might have seemed that ORFs can be designated as any substring of DNA that begins with a start codon and ends with a stop codon. If we apply this rule to any bacterial or archaeal genome we will obtain many overlapping and short ORFs. A difficult task for gene prediction software is to decide which one from two overlapping ORFs represents a true gene. This is especially true for those ORFs that are encoded on the complementary DNA strands. Gene-prediction methods (see Do & Choi, 2006; Stothard & Wishart, 2006 for reviews) are based on hidden Markov models (HMMs), such as those implemented in Glimmer (Salzberg *et al.*, 1998) or GeneMark (Besemer *et al.*, 2001). To apply them to a new sequence, one must train prediction models on a set of known genes. HMM-based gene prediction methods proved to be very efficient; however, they have inherent restrictions. A training set must be chosen from the genome that has to be annotated. This implies that genes

Abbreviations: CDS, coding sequence(s); HMM, hidden Markov model.

A supplementary table with details of the 30 *Escherichia* strains searched for missing orthologues is available with the online version of this paper.

with atypical nucleotide composition could be missed. Another source of potential errors in gene prediction is the choice of a start codon, which can be one of the six potential candidates in prokaryotes (ATG, GTG, TTG, and in some cases ATT, CTG and ATC) (see Yada *et al.*, 2001 and Zhu *et al.*, 2004 on translation initiation site prediction in prokaryotes). Stop codons can also have a dual function, as stop codons TGA or TAG may encode selenocysteine and pyrrolysine. Another problem is to decide on the cutoff for filtering short ORFs that might encode small polypeptides. Some genes are missed because they have programmed frameshifts (Farabaugh, 1996). Second-generation annotation systems (see below) try to combine multiple gene-calling programs, but this sometimes leads to gene overcalling (i.e. gene called present when a gene is in fact absent) due to the conflict between multiple gene-prediction algorithms (Armengaud, 2009).

Genes with wrong functional annotation and missed genes differentially impact different types of analyses. Comparative genome analysis usually consists of identification of orthologues and paralogues, and instances of gene gain and gene loss. Orthologues are used to infer evolutionary relationships between species, and a gene with a wrong annotation may be immediately detected, because in a phylogenetic tree it would be surrounded by genes with different annotations; however, this assumes that the annotation error is unique to individual genes and was not propagated between the genomes that contain the neighbouring genes. Short erroneously identified ORFs also would not greatly impact an orthologue collection as they most likely would not have orthologues in other genomes. They will mistakenly increase the set of ORFans, the origin of which is still a debated question (Siew & Fischer, 2003). However conclusions made about gene gain and loss, which are based on gene presence and absence analysis, could be wrong if a gene was missed as a result of gene-calling procedures. Missed genes have a particularly strong impact on the delineation of the core genome for a large and diverse group of organisms, and on identification of strain- or species-specific genes. The percentage of missed genes in one genome likely does not exceed 5–10%; however, even a single missed gene can lead to a wrong biological inference, especially if the missed gene is a key enzyme of a metabolic pathway.

Error estimation in annotated genomes

Sequencing errors in the early 1990s were estimated at the order of 0.1 % (Bork & Bairoch, 1996). These errors are responsible for frameshifts and stop-codon introduction (see below), and can influence recognition of a gene and its functional assignment. Next-generation sequencing considerably reduced cost and time, but did not greatly reduce sequencing error rate. For example, *de novo* assembly of the *Pseudomonas syringae* genome using Illumina/Solexa short sequence reads was done with an error rate of 0.33 % (Farrer *et al.*, 2009). The frequency of errors in genome

annotations is of another order. Here we have to distinguish between two different types of errors: errors in gene calling and errors in functional annotations. Errors in gene calling come from the gene-prediction methods, which are all based on Markov models. The difference in results depends first on the model itself and also on the training set. An evaluation of gene finders based on HMMs was done by Knapp & Chen (2007); the authors reported that no significant improvement in the quality of *de novo* gene prediction methods occurred during the previous 5 years. The programs were tested on human sequences, where CDS prediction is complicated by exonic and intronic structures. For additional assessment of the quality of gene-prediction methods see Majoros *et al.* (2003). Bakke *et al.* (2009) evaluated three second-generation gene-annotation systems on the genome of the archaeon *Halorhabdus utahensis* at a different level of annotation: from the performance of *de novo* gene-prediction models to the functional assignments of genes and pathways. Comparison of gene-calling methods showed that 90 % of all three annotations share exact stop sites with the other annotations, but only 48 % of identified genes share both start and stop sites. Palleja *et al.* (2008) performed an interesting investigation of overlapping CDS in prokaryotic genomes. They compared overlapping genes with their corresponding orthologues and found that more than 900 reported overlaps larger than 60 bp were not real overlaps, but annotation errors.

In addition to errors in CDS prediction, functional assignments are not error free either. One of the first estimations of the error rate of gene functional assignment was performed by comparing three published annotated genomes of *Mycoplasma genitalium* (Brenner, 1999). Of 702 annotated CDS, 55 cases (8 %) showed discrepancies in annotation. The errors intrinsic to the genome-annotation process were discussed by Devos & Valencia (2001), and the authors emphasized that most functional annotations in complete genomes are based on a relatively weak sequence identity. The estimation of potential errors of the first three published genomes of *Haemophilus influenzae*, *Mycoplasma genitalium* and *Methanococcus jannaschii* showed that, depending on the type of function, the expected rate of errors varies from less than 5 % to more than 40 % (Devos & Valencia, 2001). Bork (2000) estimated that 70 % accuracy in the annotation of functional and structural features has to be considered as a success (see also Table 1 in Bork, 2000). The reason for such a high error rate, as Bork pointed out, is that a functional annotation is based on knowledge databases that are still of insufficient quality. Nagy *et al.* (2008) discuss the identification of incorrectly predicted genes and proteins in public databases such as Ensembl and UniProtKB/TrEMBL, or by NCBI's GNOMON annotation pipeline. More recent estimates of error rates in curated sequence annotations stayed at the same level of 28–30 % (Jones *et al.*, 2007). Jones *et al.* (2007) raised the issue, already discussed by Devos & Valencia (2001), that functional

prediction is often made based on low sequence identity, which is not sufficient to accurately pinpoint function. In recent reviews on the genome-annotation process (Lee *et al.*, 2007; Reeves *et al.*, 2009) the authors acknowledge the fact that the functional annotation is most difficult, and that errors in well-known and popular databases propagate to the newly annotated genomes. For example, several ABC transporter operons in *Thermotoga* genomes were initially annotated as oligopeptide transporters because of their similarity to archaeal homologues. However, experimental studies and inspection of neighbouring genes revealed saccharides and oligosaccharides as the transported substrate (Nanavati *et al.*, 2006). In this example, the original annotation correctly identified the operons as encoding ABC transporters, but was wrong in identifying the transported substrate. The complexity of an often hierarchical annotation makes it extremely difficult to estimate the real extent of error rates in functional annotation. The process of error correction will be gradual (see also the discussion of possible solutions below) and the methods of comparative genomics may be of great use.

Comparative genome analysis as a refining tool

Despite a high error rate in gene functional assignment, the aid of automated annotation cannot be discarded, as it is always more effective to refine an existing annotation than to do a new one without prior automated annotation from scratch. Furthermore, comparative genome analysis often is not based on gene annotations, but uses the called gene sequences directly to identify orthologues, an approach that can be very useful in identifying missing orthologues. The identification of orthologous genes is usually based on similarity scores and can include information from phylogenetic reconstruction and from neighbouring genes (Arigon *et al.*, 2008; Poptsova, 2008). Some analyses in comparative genomics do not depend on functional annotation; nevertheless, in analysing absence and presence data for orthologous genes researchers need to exercise caution. While the rate of false negatives in gene calling is small, the combined effect of missed identifications in analysing several genomes is often cumulative. If, for example, an individual gene has a 0.5 % chance of being missed, in calculating the core genome for a group of 25 individual genomes, over 10 % of the core genes would be erroneously excluded from the gene set giving the strict core. In the case of delineating the core genome, one solution is to relax the definition of a core gene (Lapierre & Gogarten, 2009), but for other applications, like phylogenetic profiling, solutions may be more complicated. The erroneous absence of a gene could be due to either a gene-calling or a sequencing error. If only one base is changed (due to mutation or sequencing error), causing a frameshift or a stop-codon, an orthologue with an important function such as ATP synthase subunit, ribosomal protein or tRNA synthetase can be missed in gene calling. However, as we illustrate below, today many genes missed

in gene calling do not contain changed nucleotides that led to frameshifts or stop codons.

In that respect, similarity-based methods are very useful to improve gene annotation (Medigue & Moszer, 2007; Windsor & Mitchell-Olds, 2006). The comparative analysis of the yeast *Saccharomyces cerevisiae* and three closely related yeast species revealed gene-calling errors in 15 % of the genes (Kellis *et al.*, 2003). The method of ORF comparison between four species revealed 500 genes as meaningful, having frameshift indels and in-frame stop codons. Also, about 40 missed small genes were recognized in the intergenic areas. A similar approach was applied to the fungal pathogen *Cryptococcus neoformans* by Tenney *et al.* (2004). Apart from identifying about 200 new genes, the authors validated prediction by reverse transcription PCR for 80 % of the newly discovered genes. The comparative genome approaches proved to be efficient even in more complex genomes such as mouse, rat and human (Tenney *et al.*, 2004), with more than 900 newly discovered genes.

For prokaryotic species, comparative analysis is easier since many genomes of closely related species and for multiple strains are available. Here we attempt to demonstrate how refinement of existing annotations of closely related organisms can be effectively done through a search for missing orthologues from incomplete orthologous families. We applied this approach to a test case of 30 *Escherichia* strains available at GenBank by January 2010: 29 *Escherichia coli* strains and one *Escherichia fergusonii* strain (see complete list in Supplementary Table S1, available with the online version of this paper). We found that most orthologues missing in one of the 30 genomes are missing only in the annotation files, but that their sequence is found in the complete genome by BLAST searches on the nucleotide level. We distinguish three possible causes of gene absences in genome annotation: (1) the ORF was missed due to limitations of the ORF-finding programs used in annotation; (2) a mutation or sequencing error introduced a stop codon in the middle of the sequence; and (3) insertion/deletion created a frameshift. Orthologous gene families were selected with the BranchClust method (Poptsova & Gogarten, 2007), which distributes all genes from a given set of genomes into complete and incomplete families of orthologues. For this illustration we selected only families with a gene absent in one of the 30 *Escherichia* genomes. A total of 762 families were reported with one gene missing. Surprisingly, this list includes conserved orthologous genes such as subunits of ATP synthases, aminoacyl-tRNA synthetases, ribosomal proteins, subunits of DNA polymerase I, II and III, and some other well-known proteins (see full list in Supplementary Table S1) that are presumably essential for survival. These genes were indeed absent in the respective annotation files.

To test if the missed genes were really lost in the full genome sequences, annotated as pseudogenes, or missed only in the genome annotation, we performed the following analyses. One of the 29 orthologues was used

as a query for a TBLASTN (Altschul *et al.*, 1990) search of the full genome in which the gene was reported as missing. This target genome was used as a single nucleotide sequence. Of the 762 genes reported missing in one of the genomes, only 30 % did not produce significant hits (these genes are likely absent in the target genome); 2 % of significant hits were less than 90 % identical over the entire length of the query, but the remaining 68 % of the genes had significant hits with more than 90 % identity over the entire protein length of the query (Fig. 1d).

Significant hits with more than 90 % identity over the entire protein length of the query were analysed in more detail, and three distinct cases of possible reasons for missing ORFs were recognized (Fig. 1a–c). The first case comprises ORFs with 99–100 % identity with their corresponding orthologue that were missed at the stage of ORF prediction such as, for example, four ATP synthase subunits in *E. coli* APEC O1. As illustrated in Fig. 1(a), an alternative annotation for hypothetical proteins (hp) exists on the opposite strand in the place where ATP synthase subunits should be located. In *E. coli* UTI89 both ATP synthase subunits and the hypothetical proteins on the opposite strand are identified, but in the *E. coli* APEC O1 annotation only the hypothetical proteins are listed, and annotations of the ATP synthase subunits are absent. We determined that 71 % of the missing ORFs have an alternative annotation either on the same or on the opposite strand (Fig. 1e). These errors reflect the difficulty of some ORF-calling programs to decide on overlapping potential ORFs, which for many programs is more difficult than finding protein-coding regions (Aggarwal *et al.*, 2003; Higgs & Attwood, 2005). Surprisingly, this largest category of missed ORFs does not include any frameshifts or stop codons.

The second type of error is the introduction of a stop codon in the middle of the sequence that leads to shortened ORFs (Fig. 1b). One example is a nucleotide substitution that introduces a stop codon in the tyrosyl-tRNA synthetase missing in the annotation of *E. coli* 536. The third type of error is an insertion or a deletion that creates a frameshift, and the resulting ORF has a completely different amino acid sequence downstream of the mutation. An example of an insertion leading to a frameshift in the middle of the sequence is depicted in Fig. 1(c) for the case of the phenylalanyl-tRNA synthetase β -subunit that is missing in the annotation of *E. coli* CFT073. If these genes were not essential genes, one would consider classifying them as pseudogenes, because the method of pseudogene identification is based on searching for stop codons and frameshifts in a gene sequence. The genes from all these examples are conserved orthologues, and many encode essential functions that seem unlikely to be lost or become pseudogenes. Thus we conclude that in many instances we observe consequences of sequencing errors. All significant hits with a hit length less than the entire length of the query are truncated orthologues and true candidates for pseudogenes (Liu *et al.*, 2004). Of the genes detected by our method as

truncated ORFs, 34 % are reported in .gbk files as possible pseudogenes (see Supplementary Table S1).

Table 1 provides the list of gene-calling software used for annotation of the 30 *Escherichia* genomes. Fig. 1(e) summarizes how existing annotation is distributed over the number of missing ORFs that were found to produce significant hits with more than 90 % nucleotide identity over the entire protein length of the query. Of the missing genes, 61 % are genes whose sequences in the genomes with missing annotation do not have either stop codon or frameshift mutation. An alternative annotation exists for 71 % of these genes. About a third (34 %) of the missing genes were annotated as pseudogenes based on stop codon or frameshift introduction. An additional 4 % of missing genes with either stop codon or frameshift are not annotated as pseudogenes, but are potential candidates for pseudogenes. However the examples in Fig. 1(b, c) demonstrate that one should be cautious even with pseudogene assignment.

Alternative annotation resources and second-generation genome-annotation tools

For *E. coli*, a species of considerable interest, alternative annotation resources such as EcoCyc (Keseler *et al.*, 2009), GenoBase (Riley *et al.*, 2006) and EcoGene (Rudd, 2000) are available for two K12 strains: MG1655 and W3110. Mistakes in ORF boundaries and pseudogene identification are demonstrated and discussed by Riley *et al.* (2006) for these two strains. Recently a major effort was undertaken in reannotation of 20 *E. coli* species (Touchon *et al.*, 2009). For many applications these alternative annotations are superior to the NCBI genome sequences, which have not been updated recently. However, many researchers use the NCBI site as a main resource of conveniently and centrally available bioinformatics data. Needless to say, most of the bacterial species do not have alternative annotation resources.

Earlier genome-annotation tools did not have a huge repository of sequences to use in similarity-based methods. The latest genome-annotation tools combine both *de novo* gene prediction (usually employing existing HMM-based programs) and similarity-based methods. From Table 1 one can see that most of the *E. coli* strains annotated in the years 2008–2009 used second-generation annotation systems: MaGe (Vallenet *et al.*, 2006), RAST (Aziz *et al.*, 2008), or in-house tools of the JGI or the J. Craig Venter Institute.

For *de novo* gene prediction, the MaGe annotation system uses the AMIgene (Bocs *et al.*, 2003) program, based on HMM models. It then applies the methods of comparative genomics in finding orthologues in closely related genomes, and the orthologues found serve as a basis for the reconstruction of synteny blocks. The system is linked to knowledge databases, and functional assignment can be done in either automatic or manual fashion. The RAST server is another new fully automated tool for annotating

Fig. 1. Analysis of ORFs missing in one out of 30 completely annotated *Escherichia* genomes. (a) An example of several consecutive ORFs that were not recognized in one genome, even though the complete, uninterrupted ORFs are present. The operon is correctly annotated in the other genomes; however, in *E. coli* UT189 additional hypothetical proteins are identified as being encoded on the complementary strand, and in APEC O1 only these hypothetical proteins are identified as ORFs at the expense of the ATP synthase subunit ORFs. (b, c) Examples of substitutions in the nucleotide sequence that led to stop codons (b) or frameshifts (c). Given the importance of the encoded proteins, it seems likely that in some of these instances the mutations are not real but reflect sequencing errors. (d) Distribution of missing genes with respect to BLASTN searches using the correctly annotated ORF from one of the other genomes as query. (e) Distribution of error types for those missing ORFs that had a match with more than 90 % sequence identity in the genome for which the ORF was missing in the annotation.

Table 1. Programs and methods used for ORF calling for 30 *E. coli* strains available at NCBI in January 2010

<i>E. coli</i> strain	Year of submission	ORF calling programs/methods	No. of missed ORFs
1 O127:H6 str. E2348/69	2009	GeneHacker	14
2 536	2006	ERGO and PEDANT tools	1
3 55989	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	0
4 APEC (APEC O1:K1:H7)	2008	GeneQuest from DNASTAR	113
5 BL21_DE3	2009	JGI-ORNL and JGI-PGF	0
6 BW2952	2009	Derived from the annotation of <i>E. coli</i> strain MG1655, then Glimmer3, Artemis	2
7 B_REL606	2009	AMIGene software, ORPHEUS, Artemis	1
8 CFT073	2002	Magpie (based on Glimmer)	3
9 ATCC 8739	2008	JGI-ORNL and JGI-PGF	1
10 E24377A	2008	Glimmer	0
11 ED1a	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	0
12 HS	2008	Glimmer	6
13 IAI1	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	0
14 IAI39	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	3
15 K-12 substr. DH10B	2008	ASAP package	38
16 K-12 substr. MG1655	1997, 2006 (update)	No description on CDS finding available; annotation update from ecogene.org 2006	0
17 K-12 substr. W3110	2006 (update)	No description on CDS finding available; annotation update from ecogene.org 2006	0
18 O103:H2 str. 12009	2009	GeneHacker	1
19 O111:H- str. 11128	2009	GeneHacker	0
20 O157:H7 str. Sakai	2001	Genome Gambler, Glimmer	1
21 O157:H7 EDL933	2001	GeneMark	151
22 O157:H7 str. EC4115	2008	In-house tools of J. Craig Venter Institute	1
23 O157:H7 str. TW14359	2009	Prokaryotic Genome Annotation Tool (University of Washington), RAST	0
24 O26:H11 str. 11368	2009	GeneHacker	0
25 S88	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	0
26 SE11	2008	Glimmer	0
27 SMS-3-5	2008	No description on CDS finding available	1
28 UMN026	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	0
29 UTI89	2006	Glimmer, ORPHEUS, CRITICA, and GENEMARKS	0
30 <i>E. fergusonii</i> ATCC 35469	2009	AMIGene software, MaGe annotation system (all the results are stored in ColiScope)	4

bacterial and archaeal genomes. It is implemented in the SEED framework (Overbeek *et al.*, 2005) and uses a subsystem approach, which is based on the assumption that different experts annotate single subsystems over the complete collection of genomes, rather than one annotation expert annotating all of the genes in a single genome. The initial gene prediction is made using Glimmer software. The automated functional annotation is based on specific sets of orthologous families created in SEED. The annotated genomes can be conveniently browsed in the SEED environment. In our test study of 30 *Escherichia* genomes, missing conserved orthologues were not detected in the genomes annotated within the last 2 years.

The general tendency of improvement in genome-annotation packages comes from our increasing knowledge of genomes and from progress in methods and bioinformatics software. The most important issue today is to find a solution for how to update or even reannotate previously annotated genomes, and how the detected errors can be continuously corrected. It is currently difficult to update GenBank, EMBL, Swiss-Prot, or any other database, for researchers who did not generate the original record. An update mechanism exists only for specialized sites and specialized groups. Even the results of reannotation of 20 *E. coli* species, discussed earlier in this section, so far have not made it into the GenBank repository. The issue of updat-

ing widely used databases was discussed by Salzberg (2007), with the proposal of a wiki solution for genome reannotation. But such an approach may not be applicable for large repositories such as the NCBI. The growing genome repositories require a new level of periodic quality assessment.

Prospects for a perfect annotation

In the future, proteomics will allow for more accurate gene calling, because experimental protein expression data can be used to verify automatic gene identification (Ansong *et al.*, 2008; Armengaud, 2009). Correlation of mass spectra from expressed proteins with a nucleotide database translated into the six reading frames was reported in 1995 (Yates *et al.*, 1995). At that time producing high-throughput experimental data for proteins was difficult. Nowadays, proteomics is so advanced that it can be used not only as a refining tool, but also at the primary stage of genome annotation (Armengaud, 2009). A combined automatic annotation and proteomic analysis was done for the genome annotation of *Mycoplasma mobilis* by Jaffe *et al.* (2004). Another example of combining automatic and experimental approaches is the genome annotation of *Deinococcus deserti* (de Groot *et al.*, 2009). The method identified 15 novel genes and 11 cases of reversal. Optimistic forecasts for improved accuracy have been specifically made for eukaryotes, whose genes are often divided into exons and introns (Brent, 2008). These forecasts are based on the success of *cis*- and *trans*-alignment of cDNAs for correction of intron–exon boundaries and for detecting genes *de novo*. Genes that are expressed only under special conditions, or whose expression is below the detection level, pose a problem for proteomic and cDNA validation.

Last but not least, ultra-high-throughput next-generation sequencing technologies have stimulated the launch of numerous projects of *de novo* genome sequencing and of resequencing existing genomes. As a result, we will have more accurate genome sequences, and many sequencing errors will be fixed in the existing genomes. Also, this new technology will help to define more accurately the full complement of RNA transcripts in cells under different conditions. The accumulated experimental knowledge will provide positive feedback for the *de novo* gene-prediction methods that will be trained on experimentally verified data, and this will help to improve the theoretical models. Similarity-based methods will continue to be extremely useful for the refinement of the annotation as more and more closely related genomes become available.

Concluding remarks

Genome annotation includes gene prediction and functional annotation of predicted genes. Errors can accumulate at different stages, from genome sequencing to the assignment of metabolic pathways. In our examination of genes detected as missing in one out of 29 *E. coli* strains

and one strain of the closely related species *E. fergusonii*, we found most of the errors in the earlier annotated genomes. Missed conserved orthologues were not detected in the genomes that were sequenced and annotated in the previous 2 years and with the use of the second-generation annotation systems that employ multiple *de novo* gene-prediction and similarity-based methods. Thus the most pertinent issue is not that annotations contain unavoidable errors, but rather the need to fix them quickly and efficiently in the large data repositories that are widely used by researchers.

In using the collections of all translated ORFs contained in a genome, researchers need to be aware that an ORF could be missed or assigned an incorrect function even in the genomes of well-studied organisms. Before reporting the results based on gene presence or absence, the genome's nucleotide sequence should be used to confirm a gene's absence.

Since the time and effort needed for experimental identification of gene function are not comparable with those required for automated prediction methods, preference for the latter will stay firm in the near future. We observe a progress in the quality of gene prediction in the second-generation annotation systems, at least at the level of missed genes detection. This is explained by the extensive use of similarity-based methods to scan for genes in closely related species. Hopefully, the progress in next-generation sequencing technologies will help to further improve the *de novo* gene-prediction models.

Acknowledgements

This work was supported through the NASA Exobiology (NNX08AQ10G) and Applied Information System Research Program (NNG04GP90G).

References

- Aggarwal, G., Worthey, E. A., McDonagh, P. D. & Myler, P. J. (2003). Importing statistical measures into Artemis enhances gene identification in the *Leishmania* genome project. *BMC Bioinformatics* **4**, 23.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S. & Smith, R. D. (2008). Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief Funct Genomic Proteomic* **7**, 50–62.
- Arigon, A. M., Perriere, G. & Gouy, M. (2008). Automatic identification of large collections of protein-coding or rRNA sequences. *Biochimie* **90**, 609–614.
- Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. *Curr Opin Microbiol* **12**, 292–300.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M. & other authors (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75.

- Bakke, P., Carney, N., Deloache, W., Gearing, M., Ingvorsen, K., Lotz, M., McNair, J., Penumetcha, P., Simpson, S. & other authors (2009). Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS One* 4, e6291.
- Besemer, J., Lomsadze, A. & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29, 2607–2618.
- Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. & Medigue, C. (2003). AMIGene: annotation of microbial genes. *Nucleic Acids Res* 31, 3723–3726.
- Bork, P. (2000). Powers and pitfalls in sequence analysis: the 70 % hurdle. *Genome Res* 10, 398–400.
- Bork, P. & Bairoch, A. (1996). Go hunting in sequence databases but watch out for the traps. *Trends Genet* 12, 425–427.
- Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet* 15, 132–133.
- Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet* 9, 62–73.
- de Groot, A., Dulerio, R., Ortet, P., Blanchard, L., Guérin, P., Fernandez, B., Vacherie, B., Dossat, C., Jolivet, E. & other authors (2009). Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. *PLoS Genet* 5, e1000434.
- Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genet* 17, 429–431.
- Do, J. H. & Choi, D. K. (2006). Computational approaches to gene prediction. *J Microbiol* 44, 137–144.
- Farabaugh, P. J. (1996). Programmed translational frameshifting. *Annu Rev Genet* 30, 507–528.
- Farrer, R. A., Kemen, E., Jones, J. D. & Studholme, D. J. (2009). De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol Lett* 291, 103–111.
- Friedberg, I. (2006). Automated protein function prediction – the genomic challenge. *Brief Bioinform* 7, 225–242.
- Higgs, P. G. & Attwood, T. K. (2005). *Bioinformatics and Molecular Evolution*. Malden, MA: Blackwell.
- Jaffe, J. D., Stange-Thomann, N., Smith, C., DeCaprio, D., Fisher, S., Butler, J., Calvo, S., Elkins, T., FitzGerald, M. G. & other authors (2004). The complete genome and proteome of *Mycoplasma mobile*. *Genome Res* 14, 1447–1461.
- Jones, C. E., Brown, A. L. & Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* 8, 170.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241–254.
- Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S. & other authors (2009). EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res* 37, D464–D470.
- Knapp, K. & Chen, Y. P. (2007). An evaluation of contemporary hidden Markov model gene finders with a predicted exon taxonomy. *Nucleic Acids Res* 35, 317–324.
- Lapierre, P. & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet* 25, 107–110.
- Lee, D., Redfern, O. & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 8, 995–1005.
- Liolios, K., Mavromatis, K., Tavernarakis, N. & Kyrpides, N. C. (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.
- Liu, Y., Harrison, P. M., Kunin, V. & Gerstein, M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol* 5, R64.
- Majoros, W. H., Pertea, M., Antonescu, C. & Salzberg, S. L. (2003). GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic gene finders. *Nucleic Acids Res* 31, 3601–3604.
- Medigue, C. & Moszer, I. (2007). Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol* 158, 724–736.
- Nagy, A., Hegyi, H., Farkas, K., Tordai, H., Kozma, E., Banyai, L. & Patthy, L. (2008). Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics* 9, 353.
- Nanavati, D. M., Thirangoon, K. & Noll, K. M. (2006). Several archaeal homologs of putative oligopeptide-binding proteins encoded by *Thermotoga maritima* bind sugars. *Appl Environ Microbiol* 72, 1336–1345.
- Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T. & other authors (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33, 5691–5702.
- Palleja, A., Harrington, E. D. & Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* 9, 335.
- Poptsova, M. S. (2008). Computational techniques for orthologous gene prediction in prokaryotes. In *Computational Methods for Understanding Bacterial and Archaeal Genomes*, pp. 209–232. Edited by Y. Xu & J. P. Gogarten. London: Imperial College Press.
- Poptsova, M. S. & Gogarten, J. P. (2007). BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* 8, 120.
- Reed, J. L., Famili, I., Thiele, I. & Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nat Rev Genet* 7, 130–141.
- Reeves, G. A., Talavera, D. & Thornton, J. M. (2009). Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface* 6, 129–147.
- Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R., Glasner, J. D., Horiuchi, T., Keseler, I. M. & other authors (2006). *Escherichia coli* K-12: a cooperatively developed annotation snapshot – 2005. *Nucleic Acids Res* 34, 1–9.
- Rudd, K. E. (2000). EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res* 28, 60–64.
- Salzberg, S. L. (2007). Genome re-annotation: a wiki solution? *Genome Biol* 8, 102.
- Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 26, 544–548.
- Siew, N. & Fischer, D. (2003). Unravelling the ORFan puzzle. *Comp Funct Genomics* 4, 432–441.
- Stothard, P. & Wishart, D. S. (2006). Automated bacterial genome analysis and annotation. *Curr Opin Microbiol* 9, 505–510.
- Tenney, A. E., Brown, R. H., Vaske, C., Lodge, J. K., Doering, T. L. & Brent, M. R. (2004). Gene prediction and verification in a compact genome with numerous small introns. *Genome Res* 14, 2330–2335.

- Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C. & other authors (2009).** Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**, e1000344.
- Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C. & Médigue, C. (2006).** MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34**, 53–65.
- Windsor, A. J. & Mitchell-Olds, T. (2006).** Comparative genomics as a tool for gene discovery. *Curr Opin Biotechnol* **17**, 161–167.
- Yada, T., Totoki, Y., Takagi, T. & Nakai, K. (2001).** A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res* **8**, 97–106.
- Yates, J. R., III, Eng, J. K. & McCormack, A. L. (1995).** Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **67**, 3202–3210.
- Zhu, H. Q., Hu, G. Q., Ouyang, Z. Q., Wang, J. & She, Z. S. (2004).** Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics* **20**, 3308–3317.