Journal of Phylogenetics & Evolutionary Biology

Brocchieri, J Phylogen Evolution Biol 2015, 3:3
DOI: 10.4172/2329-9002.1000e114

Editorial Open Access

Where Are The Genes Missing From Prokaryotic Genomes?

Luciano Brocchieri*

Department of Molecular Genetics and Microbiology and Genetics Institute, University of Florida, Gainesville, USA

Despite the remarkable achievements of modern computational methods in predicting prokaryotic genes, evolutionary conservation and other analyses suggest that a significant number of genes are missing from published genome annotations [1-3]. Particularly in genomes of high GC content, the presence of coding regions missing from annotations can often be identified simply and reliably by graphically matching annotated coding regions to appropriately visualized compositional properties of the genome sequence, with the method of "frame analysis" [4]. With this method the GC content of the sequence is measured within a moving window (say, about 200 nt wide) in three subsequences, each composed of every third nucleotide, and starting, respectively, at the first, second, or third position of the genome sequence. The GC contents of the three subsequences are plotted along the genome generating three "curves" called "S-profiles" [5]. Whenever a coding region is traversed, each of the three sub-sequences will be superimposed to first, second, or third positions of the codons, depending on the coding strand and phase of the gene. Thus, each S-profile will represent the GC content of one of the three codon positions of the gene. Since in coding regions of high GC content the three codon positions have very different GC content, very high in third and relatively low in second codon positions (Figure 1), this will reflect in characteristic contrasts between S-profiles, visually corresponding to characteristic "bubbles", informative of the presence, phase, and coding strand of the underlying gene (Figure 2). When annotated genes are matched to the S-profiles, missed genes will become obvious, as unmatched bubbles. To facilitate using frame analysis on a genomic scale and extend its usage to sequences of any composition, we recently developed a method for quantifying and generalizing the information provided by frame analysis [3], and implemented these methods in the N-Profile Analysis Computational Tool (NPACT), available at http://genome.ufl.edu/npact/, which identifies sequence regions with significant 3-base periodicity ("Hits") corresponding to S-profile bubbles, looks for associated open reading frames (ORFs), compares them to genome annotations, and displays missing ORFs and corresponding hits together with S-profiles and pre-annotated genes.

Applying our methods, we identified in 1,000 genomes of any *GC* content, a plethora of significant 3-base periodicities corresponding to non-annotated ORFs, of which more than 46,000 were evolutionarily conserved in sequence and in length across bacterial genera or phyla, thus "discovering" in these genomes many "new" genes, and providing a useful tool for the amelioration of prokaryotic genome annotations. ORFs associated with significant three-base-periodicities are many more than those conserved, and we estimate that many are likely not to encode genes (they may for example indicate the presence of pseudo genes, or overlap in different frames regions of anomalous composition of true genes). However, an integral part of the method implemented by NPACT is the visual representation and comparison of 3-periodic ORFs with corresponding regions of significant periodicity, frame analysis profiles, and the position of pre-annotated genes, by which the "quality" of the newly-identified ORFs can be evaluated. Because "newly-identified" ORFs can be added to a pre-existing collection of genes, we expect the method to be at least as sensitive as the method used to obtain the set of pre-annotated genes. Since our method includes human intervention, its specificity is difficult to estimate across all genomes we tested, and depend on how the visual information if offers is evaluated by the user. In our detailed analysis of two strains of *A. dehalogenans* [3], a "quality" was assigned to each newly identified ORF, representing our confidence that it encoded a gene, uniquely based on visual analysis of compositional contrasts and relation of the ORF with neighboring genes. We then compared ORFs of different qualities with information from conservation and found that the vast majority of ORFs of good or best quality were also conserved, thus most likely improving on the sensitivity of the annotation, with a low rate of false positives (high specificity). However, when we compared our collection of "new" genes with collections of genes automatically predicted in the same genomes by the popular gene prediction methods Prodigal [6], Glimmer 3.0 [7], GeneMark HMM [8], and GeneMark 2.5 [9], we found in these collections 80% of our conserved newly-discovered genes, and 48,000 of our newly-discovered non-conserved 3-periodic ORFs. In fact, about 75% of all conserved genes were predicted by all popular prediction methods, indicating that corroboration of prediction by multiple methods increases the probability that a gene is correctly predicted.

Among all distinct genes predicted by Prodigal, Glimmer 3.0, GeneMark HMM, or GeneMark 2.5 in the 1000 genomes, there are about 20% more genes than those annotated with the same genomes. One third of the genes excluded from the annotations are uniquely predicted by Glimmer 3.0, but more than 25% are predicted by all methods (including the earlier method GeneMark 2.5), suggesting that the exclusion of many predicted genes from published annotations is not a consequence of the unavailability of the most sensitive methods at the time of annotation. Annotators may have had good reasons for excluding many predicted genes. Specificity of prediction methods (which predicted genes are false?) is difficult to evaluate, and in fact it may be that quality controls implemented in annotation pipelines improve specificity more than they decrease sensitivity. We found that more than 80% of the excluded predicted genes indeed are not conserved across genera. It is however surprising that almost 20% of the excluded predicted genes are highly conserved in sequence and in length, and that among them many are well-characterized genes. Among our newly-predicted ORFs, those that are conserved and predicted are also among the longest (with an average length of 693 nt), excluding the possibility that they had been rejected from annotations because they were shorter than a pre-established minimum-length threshold.

Our analyses indicated that sensitivity of prediction methods may not be the most significant limiting factor in achieving accurate annotations, and that many genes may be excluded from genome annotations because of how difficult and time consuming it is to distinguish true genes from false predictions on a genomic scale. Most prokaryotic genomes contain thousands of predicted genes, and their

***Corresponding author:** Luciano Brocchieri, Cancer and Genetics Research Complex, 2033 Mowry Rd, Gainesville, FL 32606, USA, Tel: 352.273.8131; E-mail: lucianob@ufl.edu

Received July 08, 2015; **Accepted** July 10, 2015; **Published** July 16, 2015

Citation: Brocchieri L (2015) Where Are The Genes Missing From Prokaryotic Genomes? J Phylogen Evolution Biol 3: e114. doi:10.4172/2329-9002.1000e114

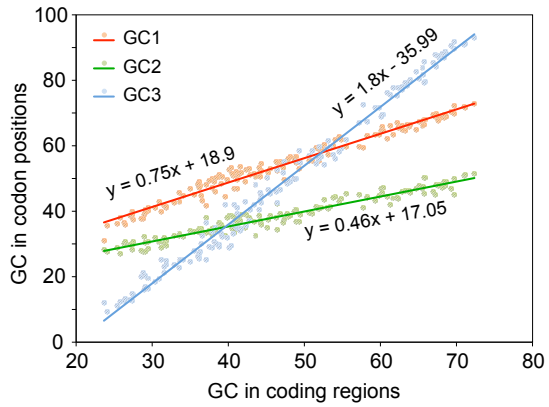

Figure 1: The average GC contents at the three codon positions, measured over all coding regions annotated within a genome, is plotted against the overall GC content of the same coding regions, for 200 genomes.

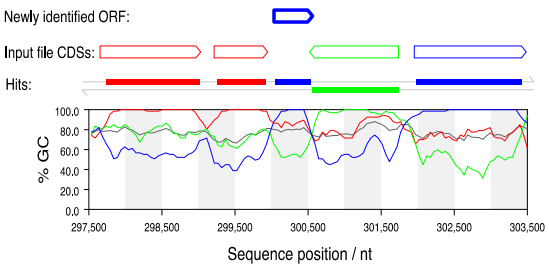

Figure 2: Phase-specific S-profiles of a segment of the *Anaeromyxobacter dehalogenans* 2CP-C genome, matched to annotated coding sequences ("Input file CDSs"), represented as arrows pointing towards the 3' end of the gene and colored according to the S-profile of the subsequence coinciding with their third codon positions (see text). Sequence segments of statistically significant 3-base periodicity ("Hits") corresponding to regions of contrast between S-profiles are also represented and colored according to the expected phase and coding strand of a coding region of that composition. One such segment corresponds to an ORF not included in the published annotation ("Newly identified ORF").

conservation analysis through similarity searches can require hours of computational time, and of tedious expert analysis. This process can be expensive in terms of human resources, prone to shifting qualitative assessments and dependent on arbitrary decisions on thresholds of "significance" (e.g., minimum levels of similarity for assessing conservation). It appears quite possible that the development of tools that help annotators in comparing multiple sets of gene predictions and in exploring sequence features across entire genomes, such as Artemis [10] or NPACT [3], may have a greater impact on the amelioration of genome annotations than any remaining improvement in sensitivity of computational prokaryotic-gene-prediction methods.

Acknowledgements

Work supported by NIH Grant 5R01GM87485-2.

References

1. Warren AS, Archuleta J, Feng WC, Setubal JC (2010) Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics 11: 131.
2. Wood DE, Lin H, Levy-Moonshine A, Swaminathan R, Chang YC, et al. (2012) Thousands of missed genes found in bacterial genomes and their analysis with COMBREX. Biology Direct 7: 37.
3. Oden S, Brocchieri L (2015) Quantitative frame analysis and the annotation of GC-rich (and other) prokaryotic genomes. An application to *Anaeromyxobacter dehalogenans*. Bioinformatics btv 339.
4. Bibb, MJ, Findlay PR, Johnson MW (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. Gene 30: 157-166.
5. Brocchieri L, Kledal TN, Karlin S, Mocarski ES (2005) Predicting coding potential from genome sequence: application to betaherpes viruses infecting rats and mice. J Virol 79: 7570-7596.
6. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.
7. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673-679.
8. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: New solutions for gene finding. Nucleic Acids Res 26: 1107-1115.
9. Borodovsky M, McIninch J (1993) GeneMark: Parallel gene recognition for both DNA strands. Computers Chem 17: 123-133.
10. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16: 944-945.