

Algorithm to Improve Gene Consistency Across Bacterial Genomes

Judith D. Cohn¹, John Dunbar², Michael E. Wall³

1 Introduction

All of genomics depends on accurate identification of genes. Unfortunately, errors in gene predictions are not immediately obvious and consequently, the extent of errors is underappreciated. Pallejà, Harrington, & Bork [1] recently found nearly a thousand spurious overlaps in 338 bacterial genomes, indicating a serious deficiency in gene predictions. Recently we noticed extensive inconsistencies among orthologous genes across the *Burkholderia* genus. In particular, we found many genes where predictions appeared to be reasonable for all but one genome. This finding suggested that gene identification might be improved by using consistency among orthologues as an additional guide. Here, we describe an algorithm to improve consistency by integrating start site predictions across multiple genomes. We characterize its performance in predicting genes that are orthologous among five genomes in the *Burkholderia* genus.

2 Methods

To develop our algorithm, we used five completed genomes that span the genus *Burkholderia*: *B. thailandensis* e264, *B. pseudomallei* 1710b, *B. ambifaria* mc40-6, *B. vietnamiensis* g4, and *B. xenovorans* lb400. Genes were identified in each genome using the recently developed Prodigal bacterial gene prediction software [2]. For each gene, we obtained a list of all potential start positions and their respective quality scores.

Sets of orthologous genes across all five genomes were identified using a standardized sequence identity metric. Orthologs were defined as pan-reciprocal best BLASTP matches among all 5 genomes. Instances of paralogs in a genome with equal identity scores were omitted in order to simplify algorithm development.

For each selected set of orthologs, a multiple DNA sequence alignment was generated using MUSCLE [3]. Positions within the alignment where all genomes shared a possible start position were identified and ranked according to a combined quality score; the highest ranked site was then chosen as the start site for all of the genes. In many cases, all of the original Prodigal start sites were aligned; for these cases, the original gene predictions were unchanged.

3 Results and Discussion

We performed gene predictions using Prodigal for the *B. thailandensis*, *B. pseudomallei*, *B. ambifaria*, *B. vietnamiensis*, and *B. xenovorans* genomes. We found 2801 genes common to all 5 genomes. Of these genes, 2706 were synonymous with genes predicted by some version of Glimmer in the original Genbank genome records. For comparison, analysis of *B. thailandensis* alone yielded 5702 predicted genes, of which 5428 had a Genbank equivalent. For 46% of the 2706 ortholog sets, the original start sites in the Genbank records were aligned (i.e. consistent) across all five genomes. The consistency increased to 64% (1733 of 2706) using the Prodigal start site predictions. Using our algorithm, we were able to select a consistent start position for 98.4% of the ortholog sets (2664 of 2706). *Our algorithm therefore dramatically increased the consistency of gene predictions across these five genomes.*

In total, application of our algorithm yielded alternative start sites for 3031 genes in 931 ortholog sets. For 277 of these sets, only a single gene prediction was changed (Table 1). These are conservative

¹ Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA. E-mail: jcohn@lanl.gov

² Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA. E-mail: dunbar@lanl.gov

³ Computer, Computational, and Statistical Sciences Division, Bioscience Division, Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA. E-mail: mewall@lanl.gov

changes such as that illustrated in Fig. 1, which apparently corrects an error in a *B. xenovorans* gene start. In the 654 remaining cases, the start site was changed in two or more genes. In particular, in a large number of cases, 445 in total, the start site was changed for all five genes. Further analysis and experimental validation will be required to determine if the alternative start sites are the true translational starts *in vivo*.

Number of Genes Modified Per Ortholog Set	Number of Ortholog Sets
1	277
2	133
3	41
4	35
5	445

Table 1: Distribution of 931 modified ortholog sets with respect to number of modified genes per set.

```

B. xenovorans      atccgctaaccgATGgcattgcaacctagtccaactgcgagtggcgATGgtggcc
B. thailandensis  attcgctgaaccgATGgcgttccaacccacaccgctgcgctcgcgctcgtcgcg
B. pseudomallei   attcgctgaccgATGgcgttccaactcacaccgctgcgctcgcgctcgtcgcg
B. vietnamiensis  atccgctaaccgATGccgctgccgaccaatcagctgcggtcgccatggtggcc
B. ambifaria      atccgctaagcgATGtcgctcccgaccaatcagctccggctcgcaatggtggcc

```

Figure 1: Modification of a single start site due to application of the gene consistency algorithm. The original starts are indicated by bold capital letters, and the modified start for *B. xenovorans* is indicated by underlined, italic capital letters.

References

- [1] Pallejà A, ED Harrington, P Bork. 2008. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* 9:335
- [2] Prokaryotic Dynamic Programming Gene-finding Algorithm (Prodigal), <http://compbio.ornl.gov/prodigal>.
- [3] Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792-1797.