Genome analysis

DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication

Yasuhiro Tanizawa, Takatomo Fujisawa and Yasukazu Nakamura*

Center for Information Biology, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima 411-8540, Japan

*To whom correspondence should be addressed.

Received on August 15, 2017; revised on October 27, 2017; editorial decision on October 30, 2017; accepted on November 6, 2017

Abstract

Summary: We developed a prokaryotic genome annotation pipeline, DFAST, that also supports genome submission to public sequence databases. DFAST was originally started as an on-line annotation server, and to date, over 7000 jobs have been processed since its first launch in 2016. Here, we present a newly implemented background annotation engine for DFAST, which is also available as a standalone command-line program. The new engine can annotate a typical-sized bacterial genome within 10 min, with rich information such as pseudogenes, translation exceptions and orthologous gene assignment between given reference genomes. In addition, the modular framework of DFAST allows users to customize the annotation workflow easily and will also facilitate extensions for new functions and incorporation of new tools in the future.

Availability and implementation: The software is implemented in Python 3 and runs in both Python 2.7 and 3.4—on Macintosh and Linux systems. It is freely available at https://github.com/nigyta/dfast_core/under the GPLv3 license with external binaries bundled in the software distribution. An on-line version is also available at https://dfast.nig.ac.jp/.

Contact: yn@nig.ac.jp

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Most scientific journals require newly obtained sequence data to be deposited in the International Nucleotide Sequence Database Collaboration (INSDC) as a condition of publication (Cochrane et al., 2016). However, submission of annotated genomes to public databases remains a burden for researchers. The NCBI provides an annotation service called Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016) incorporated in its submission system, but it is only available for GenBank submitters. The on-line server Microbial Genome Annotation Pipeline (MiGAP) (Sugawara et al., 2009) partly supports DDBJ submission; however, it requires extensive manual revision. To address these issues, we recently developed a web-based pipeline called DDBJ Fast Annotation and Submission Tool (DFAST), aiming to assist users to submit their genomes to DDBJ (Tanizawa et al., 2016). The original version of DFAST employs the lightweight command-line program Prokka (Seemann, 2014) as an annotation engine, combined with curated reference databases and a graphical user interface to create submission files to DDBJ.

Here, we report a new implementation of the background engine of DFAST, which is called DFAST-core to differentiate it from its web version. The new version features unique functions, such as pseudogene annotation and orthologous assignments between reference genomes. DFAST-core is also available as a standalone program, providing a flexible local annotation platform. Hereinafter, we simply refer to it as DFAST in this report.

2 Materials and methods

DFAST accepts a FASTA-formatted file as a minimum required input, and users can customize parameters, tools and reference
Data source/Annotation tool INSDC RefSeq Data source/Annotation tool INSDC RefSeq a RefSeqb DFAST Prokka MiGAP  
Table 1. Comparison of annotation results of E.coli O26: H11 str. 11368

<table>
<thead>
<tr>
<th></th>
<th>INSDCb</th>
<th>RefSeqb</th>
<th>DFAST</th>
<th>Prokka</th>
<th>MiGAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total CDS</td>
<td>5795</td>
<td>6243</td>
<td>5740</td>
<td>5759</td>
<td>5721</td>
</tr>
<tr>
<td>Pseudogenec</td>
<td>276</td>
<td>337 (250/87)</td>
<td>344 (158/186)</td>
<td>[30]</td>
<td>—</td>
</tr>
<tr>
<td>Selenoprotein</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>With COG number</td>
<td>—</td>
<td>—</td>
<td>3965</td>
<td>—</td>
<td>4392</td>
</tr>
<tr>
<td>Unknown function</td>
<td>1203</td>
<td>1514</td>
<td>1347</td>
<td>2068</td>
<td>418</td>
</tr>
<tr>
<td>tRNA</td>
<td>101</td>
<td>101</td>
<td>105</td>
<td>105</td>
<td>100</td>
</tr>
<tr>
<td>rRNA</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>CRISPR array</td>
<td>—</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>—</td>
</tr>
<tr>
<td>Running time</td>
<td>—</td>
<td>—</td>
<td>3 m 27 s</td>
<td>3 m 20 s</td>
<td>4 h 43 m</td>
</tr>
</tbody>
</table>

Note: Numbers represent annotated features and running time. DFAST and Prokka were run on a 4-core Macintosh laptop with default settings.

aOriginal annotation by submitters (GCA_000091005.1).
bAnnotated by PGAP (GCF_000091005.1).
cNumbers in parentheses denote internal stop codon/frameshift and partial genes, respectively.
dCandidates for pseudogenes are mentioned in the log file, not in the result.
Our simple strategy to find pseudogenes depends on the accuracy of reference databases. However, when references from close relatives are available, DFAST outperforms other tools. Among 158 CDSs in which internal stop codons or frameshifts were identified, 123 were found to be consistent with the INSDC data (78%). Although the comparison is not straightforward as annotation formats are different, 97 out of 250 identified by PGAP were consistent (39%). Notably, DFAST succeeded in annotating all 3 selenoproteins present in the query genome.

Another major advantage of our pipeline is its speed. The running time of DFAST is comparable with that of Prokka, yet the default reference database of DFAST (417 922 sequences in total) is 20 times larger than that of Prokka (18 276 sequences). This is mostly attributable to the efficient algorithm of GHOSTX. If BLASTP is used instead, running time will increase up to 40 min under the same condition. In accordance with the database size, the number of genes with assigned function was larger than Prokka, although smaller than MiGAP, which conducts sequence search against a more comprehensive database such as UniProtKB/TrEMBL.

In general, DFAST performs well with the default settings on well-characterized organisms, such as Actinobacteria, Firmicutes and Proteobacteria. The annotation of the genomes from less-studied species, for which references of close relatives are not present in the default database, may contain relatively large number of uncharacterized genes. In such cases, providing additional references will improve the results as demonstrated in Supplementary Notes.

Acknowledgements

We thank Yoshitoshi Ogura and Keiji Nakamura for providing us the sample genome data, Martin Frith for suggestions on pseudogene detection, Yutaka Akiyama and Shuji Suzuki for technical advice for GHOSTX, and Masanori Arita for reviewing the manuscript. MetaGeneAnnotator is redistributed by courtesy of Hideki Noguchi.

Funding

This work was supported by JSPS KAKENHI [grant number 16H06279].

Conflict of Interest: none declared.

References