BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 7 2009, pages 962–963
doi:10.1093/bioinformatics/btp097

Genome analysis

DIYA: a bacterial annotation pipeline for any genomics lab

Andrew C. Stewart[1], Brian Osborne[2] and Timothy D. Read[1,*,†]

[1]Genomics Department, Biological Defense Research Directorate, Naval Medical Research Center, Rockville, MD and [2]The BioTeam Inc., Middleton, MA, USA

Received on August 26, 2008; revised on February 12, 2009; accepted on February 16, 2009

Advance Access publication March 2, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: DIYA (Do-It-Yourself Annotator) is a modular and configurable open source pipeline software, written in Perl, used for the rapid annotation of bacterial genome sequences. The software is currently used to take DNA contigs as input, either in the form of complete genomes or the result of shotgun sequencing, and produce an annotated sequence in Genbank file format as output.

Availability: Distribution and source code are available at (https://sourceforge.net/projects/diyg/).

Contact: tread@emory.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Genome annotation is the process of embellishing raw DNA sequences with predictions of features such as genes and transcription factor binding sites. These assignments are necessary to identify important gene functions and to enable comparative analysis. There are now several web-based servers that allow anonymous public submission of bacterial genomes for high-quality automated annotation [e.g. Almeida *et al*., 2004; Aziz *et al*., 2008; Davila *et al*., 2005; Van Domselaar *et al*., 2005; IGS Annotation Engine (http://ae.igs.umaryland.edu/cgi/ae_pipeline_outline.cgi)]. However, there have been fewer open source tools developed that allow users to run microbial annotation pipelines at their own site.

In recent years there has been a revolution in genome sequencing, allowing for rapid shotgun draft sequence production of microbial genomes overnight (Mardis, 2008). These technologies have created many new small 'genome centers' (Zwick, 2005). DIYA (Do-It-Yourself Annotator) arose out of the desire for our group to be able to examine annotated microbial genomes on our own servers as soon as possible after generating the raw sequence data on Roche/454 Sequencing GS-FLX instruments. The essential properties of the required program were that it,

(1) Accepts as input either randomly ordered DNA contigs or complete genomes in fasta format or can download contigs from the NCBI based on Genome project ID.

(2) Uses open source annotation programs and biological databases.

(3) Is relatively straightforward to configure.

(4) Can be installed on a wide a range of hardware.

(5) Is modular; allowing for extension and customization of the pipeline.

(6) Outputs common file formats.

2 METHODS AND RESULTS

DIYA is written in object-oriented Perl and uses the Bioperl library (Stajich *et al*., 2002) for sequence conversion and annotation. Installation and configuration of DIYA requires basic knowledge of Perl and XML. Each DIYA component is tested on installation. All DIYA pipelines are composed of steps that are executed in a specific order, and each step is called either a 'parser' step or a 'script' step. The single configuration file must be edited to change the order of the steps, or add and delete steps. The pipeline is controlled by a master Perl module, diya.pm, which reads the configuration file and executes each step in the pipeline, passing the output from one step to the next step as input. The full DIYA pipeline is executed at the command-line using a simple Perl script which calls methods in diya.pm. Many genomes can be annotated simultaneously by running in batch mode using the Sun Grid Engine (SGE) scheduler.

For every DIYA parser step there will be a bioinformatics application that will analyze sequence and produce output. A corresponding Perl module parses that output and performs an action (for instance creating an annotated Genbank file). A script step is simpler than a parser step, and may do something like move a file, create a database or send an alert. An example pipeline is outlined below.

(1) Download a Genbank genome sequence using its Entrez ID. The nucleotide sequences for an NCBI project are downloaded as a fasta format file.

(2) Create a Genbank file. The unordered contigs in the downloaded fasta file are assembled into a 'pseudo-contig' in Genbank format.

(3) Gene finding using Glimmer3 (Salzberg, 1998). This is a parser step that runs Glimmer3 and parses its output using the g3-from-scratch.csh script. Glimmer3 is commonly used for identification for coding regions in microbial genomes (Delcher *et al*., 2002). The output for this step is a reannotated Genbank file containing coding regions predicted by Glimmer3.

This pipeline could be modified by adding other steps. For example, DIYA comes with code and configuration files for the identification of non-coding RNAs, tRNAs (using tRNAscan-SE; Lowe and Eddy, 1997), and for performing Blast or RPS Blast analysis (Altschul *et al*., 1990) of coding regions extracted from Genbank files. Currently, gene product names are based on the simplistic scheme of transferring annotation from the best Blast or RPS Blast match. This is appropriate for many end uses of the information,

*To whom correspondence should be addressed.

†Present address: Division of Infectious Diseases and Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA.

such as resequencing or first pass annotation but could be extended in the future by more sophisticated product-naming parsers.

The output from the analyses are Genbank files containing annotations. This file format is very familiar to biologists and can be used in a wide variety of commercial and open source software. We routinely use the Genbank outputs files as input to General Feature Format (GFF) converters, and the results are visualized in the open source genome browser, GBrowse (Stein *et al*., 2002).

A DIYA script called gbconvert creates ASN.1 (Abstract Syntax Notation One) format files for easy submission of the genome project to the National Center for Biotechnology Information (NCBI). Information on eight genomes submitted to NCBI via the DIYA pipeline can be found in Supplementary Table 1. The gbconvert script contains more than 100 text-formatting rules derived from interaction with NCBI staff.

3 CONCLUSIONS

DIYA as currently implemented is a lightweight microbial annotation pipeline producing data suitable for rapid visualization of bacterial genomes. We have used DIYA to annotate more than 50 bacterial genomes to date (including the Yersinia genomes listed in the Supplementary Table 1) as a basis for large-scale comparative analysis. Since the program can be installed locally, the user can have control over how often, and with what priority, jobs are run. Local control is also important if the user is concerned about posting preliminary data on hard drives outside their institution. The recently published PIPA (Protein Identification Pipeline) software (Yu *et al*., 2008) automates the querying of multiple databases and organizes the output and can accept the Genbank output file from the DIYA pipeline as input. An alternative use of DIYA is to reannotate genomes already submitted to NCBI. This can be done by simply supplying the Entrez genome project database ID to DIYA, which will download and annotate all associated DNA molecules.

Our plans for this project are to gradually make more modules available. Functions we are looking to add to the DIYA pipeline include software for detection of prophages, CRISPR elements (Sorek *et al*., 2008) and pseudogenes. In the future we plan to integrate DIYA into a virtual appliance for easy deployment across everything from laboratory workstations to cloud computing facilities.

ACKNOWLEDGEMENTS

We would like to extend our gratitude to Peter Chen, Mike Cariaso, Jason Zhang, Dan Sommer, Mihai Pop, Art Delcher and Bill Klimke for their assistance. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense nor the US Government. Some of the authors are employees of the US Government. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government'. Title 17 U.S.C. §101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.

Funding: Joint Science and Technology Office for Chemical and Biological Defense (TMTI0068_07_NM_T); Defense Threat Reduction Agency Initiative (to T.D.R.).

Conflict of Interest: none declared.

REFERENCES

Almeida,L.G. *et al*. (2004) A system for automated bacterial (genome) integrated annotation–sabia. *Bioinformatics*, **20**, 2832–2833.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Aziz,R.K. *et al*. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Davila,A.M. *et al*. (2005) GARSA: genomic analysis resources for sequence annotation. *Bioinformatics*, **21**, 4302–4303.

Delcher,A.L. *et al*. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, **30**, 2478–2483.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Mardis,E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.

Salzberg,S.L. *et al*. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.

Stajich,J.E. *et al*. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Stein *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599.

Sorek,R. *et al*. (2008) CRISPR–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.

Van Domselaar,G.H. *et al*. (2005) BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.*, **33**, W455–W459.

Yu,C. *et al*. (2008) The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics*, **9**, 52.

Zwick,M.E. (2005) A genome sequencing center in every lab. *Eur. J. Hum. Genet.*, **13**, 1167–1168.