Frontiers in Genetics

Methods
9 June 2014

FAST: Fast Analysis of Sequences Toolbox

Travis J. Lawrence [1], **Dana L. Carper** [1], **Katherine C.H. Amrine** [1,2], **Kyle Kauffman** [3], **Claudia Canales** [3] **and David H. Ardell** [1,*]

[1] *Program in Quantitative and Systems Biology, University of California, Merced, CA, USA*
[2] *Present address: Department of Viticulture and Enology, University of California, Davis, CA, USA*
[3] *School of Natural Sciences, University of California, Merced, CA, USA*

Correspondence*:
David Ardell
University of California, Merced, Quantitative and Systems Biology, 5200 North Lake Road, Merced , CA , 95343, USA, dhard@ucmerced.edu

ABSTRACT

Refer to `http://www.frontiersin.org/Genetics/authorguidelines` or **Table??** for abstract requirement and length according to article type.

Keywords: Text Text Text Text Text Text Text Text

1 INTRODUCTION

The field of molecular biology has changed significantly with the advent of next generation sequencing technology. It is now commonplace to analyze gigabases worth of data per experiment. Traditionally programs were developed for visualization and for basic sequence manipulation by a GUI interface (**Smith et al.**, 1994; **Rampp et al.**, 2006). Most available bioinformatic toolkits are designed for specific types of data or analysis requiring several toolkits to be installed. Moreover, each toolkit often requires a different file format making data analysis difficult.

The FAST utilities are modeled after the standard Unix toolkit(**Peek**, 2001), follow the Unix philosophy of “do one thing and do it well” (**Stutz**, 2000), and are written in PERL using bioperl packages (**Stajich et al.**, 2002). This makes FAST utilities easy to adopt if you are familiar with the Unix toolkit and allows fast sequence analysis even on large datasets. FAST utilities have a uniform interface requiring FASTA formatted files and are capable of reading data from STDIN. This allows quick prototyping of sequence analysis problems by piping data between several utilities. Additionally, fasconvert can convert to/from fasta from/to several formats increasing the flexibility and usability of FAST. Extensive documentation has been developed for each utility along with useful error messages following the recommendations of **Seemann** (2013) to increase usability. Lastly, FAST is open source, which makes it available to anyone free of cost. This is in line with the call to make science more assessable, open, and reproducible by other scientists and the public (**Groves and Godlee**, 2012).

FAST is split into three categories selection, transformation, and annotation and analysis. The selection category contains utilities designed to select sequences and sites from alignments based on several different criteria. For example fasgrep selects sequences by matching a regular expression to the ID, description, or sequence. The transformation utilities are used to modify the ID, description, sequence, or order of sequences using several criteria. For example, fastaxsort sorts sequences within a multifasta file based on NCBI taxonomy (**Benson et al.**, 2009; **Sayers et al.**, 2009). The annotation and analysis category contains utilities to calculate sequence composition, codon usage, sequence length, and basic population genetic statistics. Additionally these utilities can also append the results of the analysis to the sequence description, which then can be used as selection criteria by the utilities in the selection category.

Some utilities within FAST have overlapping function with those found within other toolkits. For example sequence composition, sequence translation, and codon usage are available in the EMBOSS package (**Rice et al.**, 2000). Another example is the Bioinformatics Toolbox (**White et al.**, 2014) that has utilities to select only unique sequences and extract sequences from Genbank files based on gene name, which have overlapping function with fasuniq and gbfeat2fas respectively. However, the utilities in EMBOSS (**Rice et al.**, 2000) and Bioinformatics Toolbox (**White et al.**, 2014) lack a uniform interface, are not modeled after the Unix toolkit, and do not have the ability to use regular expressions to select and manipulate sequences. However, FAST also contains several utilities that have unique functionality. For example gbalncut takes a multiple sequence alignment annotated with a genomic feature and a genbank file and allows you to select certain regions of the alignment such as all the exons or the coding sequence. Another example is fastaxsort that allows sorting of a multifasta file based on NCBI taxonomy (**Benson et al.**, 2009; **Sayers et al.**, 2009).

2 DESCRIPTION

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text. might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. might want to know about text text text text

2.1 SELECTION UTILITIES

2.2 TRANSFORMATION UTILITIES

2.3 ANNOTATION AND ANALYSIS UTILITIES

2.4 USABILITY AND SCALABILITY

3 DISCUSSION

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Additional Requirements:

3.1 DATA SHARING

Frontiers supports the policy of data sharing, and authors are advised to make freely available any materials and information described in their article, and any data relevant to the article (while not compromising confidentiality in the context of human-subject research) that may be reasonably requested by others for the purpose of academic and non-commercial research. In regards to deposition of data and data sharing through databases, Frontiers urges authors to comply with the current best practices within their discipline.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

The statement about the authors and contributors can be up to several sentences long, describing the tasks of individual authors referred to by their initials and should be included at the end of the manuscript before the References section.

ACKNOWLEDGEMENT

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Funding: Text Text Text Text Text Text Text Text.

SUPPLEMENTAL DATA

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

REFERENCES

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009), GenBank., *Nucleic acids research*, 37, Database issue, D26–31, doi:10.1093/nar/gkn723

Groves, T. and Godlee, F. (2012), Open science and reproducible research., *BMJ (Clinical research ed.)*, 344, jun26_1, e4383, doi:10.1136/bmj.e4383

Peek, J. (2001), Why Use a Command Line Instead of Windows?, `http://www.linuxdevcenter.com/pub/a/linux/2001/11/15/learnunixos.html`

Rampp, M., Soddemann, T., and Lederer, H. (2006), The MIGenAS integrated bioinformatics toolkit for web-based sequence analysis., *Nucleic acids research*, 34, Web Server issue, W15–9, doi:10.1093/nar/gkl254

Rice, P., Longden, I., and Bleasby, A. (2000), EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, 16, 6, 276–277, doi:10.1016/S0168-9525(00)02024-2

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2009), Database resources of the National Center for Biotechnology Information., *Nucleic acids research*, 37, Database issue, D5–15, doi:10.1093/nar/gkn741

Seemann, T. (2013), Ten recommendations for creating usable bioinformatics command line software., *GigaScience*, 2, 1, 15, doi:10.1186/2047-217X-2-15

Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W., and Gillevet, P. M. (1994), The genetic data environment an expandable GUI for multiple sequence analysis., *Computer applications in the biosciences : CABIOS*, 10, 6, 671–5

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002), The Bioperl toolkit: Perl modules for the life sciences., *Genome research*, 12, 10, 1611–8, doi:10.1101/gr.361602

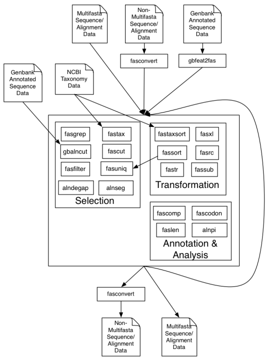

Figure 1. Enter the caption for your figure here. Repeat as necessary for each of your figures

Stutz, M. (2000), Linux and the Tools Philosophy, `http://www.linuxdevcenter.com/pub/a/linux/2000/07/25/LivingLinux.html`

White, B. P., Pilgrim, E. M., Boykin, L. M., Stein, E. D., and Mazor, R. D. (2014), Comparison of four species-delimitation methods applied to a DNA barcode data set of insect larvae for use in routine bioassessment, *Freshwater Science*, 33, 1, 338–348, doi:10.1086/674982

FIGURES

Figure 1. Enter the caption for your figure here. Repeat as necessary for each of your figures.