BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 7 2006, pages 885–886
doi:10.1093/bioinformatics/btl028

Sequence analysis

Look-Align: an interactive web-based multiple sequence alignment viewer with polymorphism analysis support

Payan Canaran[1], Lincoln Stein[1] and Doreen Ware[1,2,*]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA and
[2]USDA-ARS NAA Plant, Soil and Nutrition Laboratory Research Unit, Tower Road, Ithaca, NY 14853-2901, USA

Received on October 11, 2005; revised on January 24, 2006; accepted on January 26, 2006
Advance Access publication February 10, 2006
Associate Editor: Golan Yona

ABSTRACT

Summary: We have developed Look-Align, an interactive web-based viewer to display pre-computed multiple sequence alignments. Although initially developed to support the visualization needs of the maize diversity website Panzea (http://www.panzea.org), the viewer is a generic stand-alone tool that can be easily integrated into other websites.

Availability: Look-Align is written in Perl using open-source components and is available under an open-source license. Live installation and download information can be found at the Panzea website (http://www.panzea.org/software/alignment_viewer.html).

Contact: ware@cshl.edu

Supplementary information: The Supplementary information includes sample lists of multiple sequence alignment software and sample screenshots of the viewer.

INTRODUCTION

Multiple sequence alignments are used for a variety of applications in sequence analysis. They can be generated by a number of software programs (see reviews by Batzoglou, 2005; Nicholas *et al*., 2002; Notredame, 2002), many of which also offer visualization of the generated alignments (see Supplementary information). The great majority of applications with alignment visualization features were developed as interactive desktop applications and are not suited for deployment in a web environment.

To counter this difficulty, we have developed Look-Align, an interactive web-based viewer to display pre-computed multiple sequence alignments that have been generated by the Maize Diversity Project on its website, Panzea (http://www.panzea.org/). We have built in a variety of polymorphism analysis features, including the ability to display quality scores and variation statistics and to perform on-the-fly filtering of sequences. Although initially developed to support the visualization needs of the project website, the viewer is a generic stand-alone tool that can be easily integrated into other websites.

*To whom correspondence should be addressed.

WEB USER FEATURES

The main page of the viewer allows the user to query alignments using either unique alignment identifiers or identifiers associated with sequences in an alignment (e.g. gene or locus name). Depending on the data source, there may be one or more of these search methods available. The main page also displays the current values of the global parameters. These parameters apply to all operations performed within the viewer and can be customized by the user.

Searching for an alignment in the main page retrieves an overview of the alignment. If the search method supports retrieval of more than one alignment, the user is presented with a list of alignments to choose from. The overview contains a graphical representation of the complete alignment, and points of variation between sequences in the alignment are marked. These marks can be clicked to open base pair resolution alignments around that point. This alignment view displays a consensus sequence and base quality scores when available, with positions of variation highlighted. The user can change the region in display using the navigation options on top of the page.

Low-quality sequences and segments, or sequences that do not align well with the rest of the sequences in a given alignment, can cause variation point artifacts. To minimize these artifacts, the user can activate an optional on-the-fly filtering procedure that eliminates problematic sequences based on user-specified thresholds. The filtering procedure consists of three steps. (1) The software replaces all bases in a sequence that have a quality score lower than a specified threshold with an 'N', indicating an unknown base pair. Replaced bases are disregarded when calculating variation and percent identity. (2) The software aligns the processed sequences to a reference sequence and calculates the percent identities for each pair. Any sequence that has an identity percentage lower than the specified threshold is completely removed from the alignment set. To determine the reference sequence, the software calculates the percent identity for each possible pair of sequences. The software selects as the reference sequence the one that associates to the greatest number of sequences with a percent identity higher than the specified threshold. (3) The removal of sequences can create common gaps that exist

in all of the sequences in the alignment. The final step of the filtering process is to remove common gaps from all of the sequences. When filtering is in effect, filtering parameters, calculated reference sequence and eliminated sequences are displayed as a report on the overview page.

Both the overview and the alignment pages contain a utility that allows users to retrieve complete or partial alignments in text format.

SOFTWARE

Look-Align is written in Perl and uses readily available open-source components. It runs on a Linux/Unix environment running the Apache Web Server. The Panzea website is built on top of a MySQL database that uses GDPDM, the Genomic Diversity and Phenotype Data Model (http://www.maizegenetics.net/gdpdm). The viewer can run on top of GDPDM-based MySQL databases and flat files. It has a modular structure, and interface modules for the above two data source types are already available. More interface modules can be written by adapting those provided. For example, adapting the GDPDM interface module to retrieve data from another schema can be accomplished by modifying its data retrieval components. A similar adaptation can be performed on the flat file module to support other flat file formats.

The viewer's functional features and look-and-feel can be customized to fit each individual website's needs through a simple text-based configuration file. Some of the customizable features are the allowed value ranges and the default values for each parameter, details of the cookie used to store parameter information, amount of debugging information to be displayed, the style sheet used for formatting the displays, and page header/footer and additional website-specific information to be embedded in the displays. Although many components of the viewer are customizable, default parameter values provided in the installation package can be used with little modification for an initial fully-functional display.

Generating real-time image displays for large alignments would be impractical. Consequently, a utility is provided to pre-cache large alignment displays that would otherwise be time-consuming to generate on the fly. The pre-caching mechanism stores retrieved data and generated images in the cache directory defined by the configuration file. This directory can be re-built when a new data release is made. For the Panzea release at the time of this publication, on a server with 3.06 GHz CPUs, we chose to pre-cache alignments that have >40 sequences or >20 kb (e.g. 25 sequences of 800 bases each). In this dataset, using pre-caching, we are able to display alignments with >40 sequences with an average sequence length of ∼11 kb. Although, larger alignments can be displayed using the pre-caching feature, the viewer is not designed to handle very large segments such as genomic alignments. The caching feature can be turned on and off from the configuration file.

PLANNED IMPROVEMENTS

Improvements planned for future releases include an automatic installer, ability to upload data through the web interface, support for linking out to external sources, an interface for alignment editing, support for indels and additional interface modules.

ACKNOWLEDGEMENTS

The authors would like to thank John Doebley, Edward Buckler and Brandon Gaut for providing comments, interface suggestions and testing of the application, Todd Harris for critical review of the manuscript, Ken Youens-Clark for critical review of the manuscript and application code and Lalitha Krishnan for critical review of the installation instructions of the application. This work was funded by NSF project 0321467 and USDA ARS. Funding to pay the Open Access publication charges was provided by NSF #0321467.

Conflict of Interest: none declared.

REFERENCES

Batzoglou,S. (2005) The many faces of sequence alignment. *Brief Bioinform.*, **6**, 6–22.

Nicholas,H.B.,Jr *et al.* (2002) Strategies for multiple sequence alignment. *Biotechniques*, **32**, 572–574, 576, 578 passim.

Notredame,C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, **3**, 131–144.