BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 1 2015, pages 112–113
doi:10.1093/bioinformatics/btu547

Sequence analysis

Advance Access publication September 4, 2014

mod_bio: Apache modules for Next-Generation sequencing data

Pierre Lindenbaum[1,2,3,4,*] and Richard Redon[1,2,3,4]

[1]Institut National de la Santé et de la Recherche Médicale (INSERM) Unité Mixte de Recherche (UMR) 1087, L'Institut du Thorax, [2]Centre National de la Recherche Scientifique (CNRS) UMR 6291, [3]Centre Hospitalier Universitaire (CHU) de Nantes, L'Institut du Thorax, Service de Cardiologie, 44000 Nantes and [4]Université de Nantes, 44000 Nantes, France

Associate Editor: Inanc Birol

ABSTRACT

Summary: We describe *mod_bio*, a set of modules for the Apache HTTP server that allows the users to access and query fastq, tabix, fasta and bam files through a Web browser. Those data are made available in plain text, HTML, XML, JSON and JSON-P. A javascript-based genome browser using the JSON-P communication technique is provided as an example of cross-domain Web service.

Availability and implementation: https://github.com/lindenb/mod_bio.

Contact: pierre.lindenbaum@univ-nantes.fr

Received on April 28, 2014; revised on July 18, 2014; accepted on August 7, 2014

1 INTRODUCTION

The open-source Apache HTTP Server is the most widely used Web server (http://en.wikipedia.org/wiki/Apache_HTTP_Server). It has a generalized programming interface extending the functionality of the basic server. Those pieces of C code, named 'modules', register custom hooks in the core server. A module has an access to the server's data structures and is able to customize the response back to the client. The Apache server is used by the main centers in bioinformatics like the NCBI, UCSC, Ensembl, 1000Genomes and most of the time, they provide an online area where the users can download some structured raw files. It is often a FTP server, which can go with a HTTP server: at the time of writing, the 1000Genomes data are available through both protocols. Common tasks such as getting the first lines of a file or accessing the data in a given genomic region are easily provided using standard linux tools or using the coordinate-sorted index of samtools (Li *et al.*, 2009) and tabix (Li, 2011), but this index needs to be downloaded, and it remains difficult for the biologists to quickly get an overview of those data. An interactive program such as Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011) allows viewing the content of a remote BAM file but the index for the file must be downloaded too, and interaction is limited to the software itself. The Distributed Annotation System (DAS) (Jenkinson *et al.*, 2008) protocol was also developed to answer those coordinate-based queries, but the response is limited to the XML format, and it requires to define a registry and some entry points, and it doesn't allow to fetch the original data. Ultimately, it becomes hard to develop some Web interactive applications, which need to access remote data without downloading the whole dataset, especially because of cross-site scripting security issues. A first way to circumvent this problem is to return a JSON-P document, a communication technique used in Javascript programs to request data from a server in a different domain. Another solution is to enable the server to use the 'Cross-origin resource sharing' (CORS) protocol, a strategy implemented by the 'Dalliance genome browser' (Down *et al.*, 2011). Dalliance uses CORS in combination of the HTTP 'Range:' header to fetch chunks of NGS data, but it depends of multiple javascript libraries, and the whole indexes for the BAM or Tabix files must be loaded in the browser's memory.

2 RESULTS

To answer those issues, we have developed *mod_bio*, a set of apache modules providing a user-friendly overview of the bioinformatics files through a Web browser. Modules are activated using the files extensions and thus, there is no need to index documents in a database on the server side. In practice, when a remote directory is listed in a Web browser, the files managed by *mod_bio* are displayed with some extra hyperlinks that offer the possibility to display the contents in a browser using the following alternative formats: modules are able to print the data in HTML or plain text but also provide an option to retrieve the data using the XML or JSON formats.

Coordinate-sorted indexed files like BAM or tabix-indexed files can be queried by specifiying a genomic range in the URL (Fig. 1). A module named 'mod_fastq' handles Fastq files (Cock *et al.*, 2010) and displays the first short-reads of a file, giving an overview of the length of the reads of the platform used for sequencing. A second module named 'mod_faidx' retrieves fragments of Fasta reference sequences indexed with '*samtools faidx*' (Li *et al.*, 2009). Another module 'mod_tabix' retrieves genomic data such as VCF or GFF files that have been compressed with '*bgzip*' and indexed with '*tabix*' (Li, 2011). Finally, the module named 'mod_bam' handles the SAM alignments in a BAM file (Li *et al.*, 2009). If the server does not support CORS a supplementary parameter '*callback*' in the URL turns each JSON output into a JSON-P document. The module *mod_bio* interacts smoothly with the other apache modules like 'mod_deflate', which compresses the response sent to the client over the network.

Hence, Apache server using *mod_bio* are turned into a real Web service for bioinformatics. *mod_bio* was developed in the C programming language. It is heavily based on 'htslib', the core C library of samtools, bcftools and tabix (https://github.com/samtools/htslib).

*To whom correspondence should be addressed.

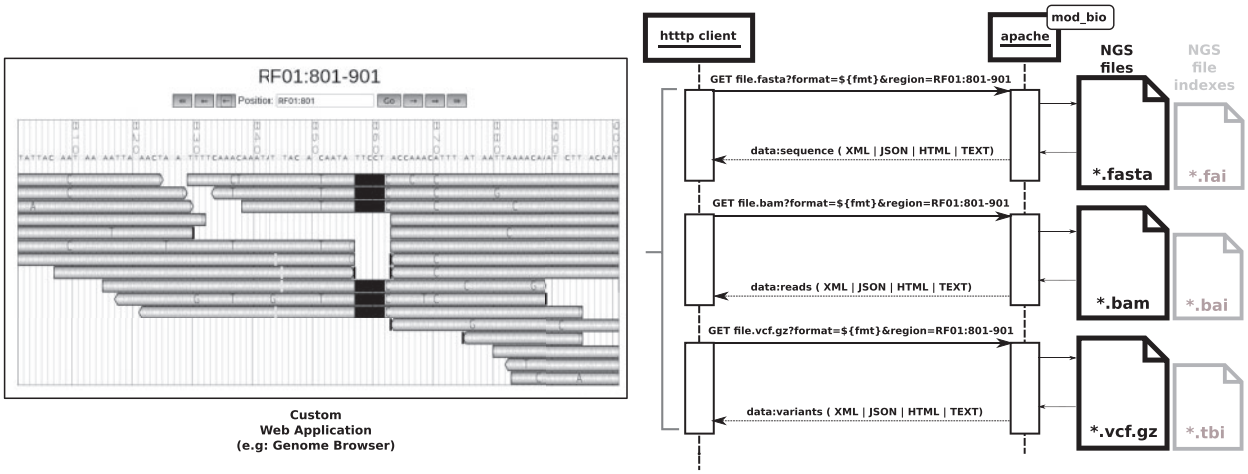

Fig. 1. The client sends GET requests to the HTTP server running *mod_bio*. Parameters like '*region*' and '*format*' modify the normal response of the server which returns structured data (XML, JSON, . . .) back to the client where it can be gathered by a custom tool such as a 'Genome Browser'

In the end, for the system administrators, *mod_bio* provides a quick way to expose their data through a Web service or a Web browser. Biologists will find here an interface to explore the data. Future developments will include the support of other indexed formats like BigWig, BigBed and Cram (Bonfield, 2014; Kent *et al*., 2010) and consolidating the output formats to fit formal specifications like those defined by the 'Global Alliance for Genomics and Health' (https://github.com/ga4gh/schemas).

The source code of *mod_bio* is available on https://github.com/lindenb/mod_bio. We also have installed a demo server of *mod_bio* on http://cardioserve.nantes.inserm.fr/lindenb/mod_bio/00EXAMPLE.html. This demonstration includes a JSON-P dynamic Web browser displaying a BAM and its reference sequence using the Canvas HTML5 element.

We believe that *mod_bio* can be used to release structured scientific data publicly. It provides a quick interface to get an overview of 'Next-Generation Sequencing' output files.

ACKNOWLEDGEMENTS

The authors want to thank the bioinformatics core facility of Nantes (Biogenouest) for technical support, Dr Julien Barc, Dr Marta Sánchez and Dr Stéphanie Bonnaud for the dust removal.

Funding: This work was supported by the Inserm (ATIP-Avenir program) and the French Regional Council of Pays-de-la-Loire ('VaCaRMe Project').

Conflict of interest: none declared.

REFERENCES

Bonfield,J.K. (2014) The scramble conversion Tool. *Bioinformatics*, **30**, 2818–2819.
Cock,P.J.A. *et al.* (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res*., **38**, 1767–1771.
Down,T.A. *et al.* (2011) Dalliance: interactive genome viewing on the web. *Bioinformatics*, **27**, 889–890.
Jenkinson,A.M. *et al.* (2008) Integrating biological data–the Distributed Annotation System. *BMC Bioinformatics*, **9** (**Suppl. 8**), S3.
Kent,W.J. *et al.* (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.