BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 20 2013, pages 2643–2644
doi:10.1093/bioinformatics/btt455

Genome analysis

Advance Access publication August 7, 2013

PARSEC: PAtteRn SEarch and Contextualization

Alexis Allot, Yannick-Noël Anno, Laetitia Poidevin, Raymond Ripp, Olivier Poch and Odile Lecompte*

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) CNRS/INSERM/UDS, 67404, Illkirch, France

Associate Editor: John Hancock

ABSTRACT

Summary: We present PARSEC (PAtteRn Search and Contextualization), a new open source platform for guided discovery, allowing localization and biological characterization of short genomic sites in entire eukaryotic genomes. PARSEC can search for a sequence or a degenerated pattern. The retrieved set of genomic sites can be characterized in terms of (i) conservation in model organisms, (ii) genomic context (proximity to genes) and (iii) function of neighboring genes. These modules allow the user to explore, visualize, filter and extract biological knowledge from a set of short genomic regions such as transcription factor binding sites.

Availability: Web site implemented in Java, JavaScript and C++, with all major browsers supported. Freely available at lbgi.fr/parsec. Source code is freely available at sourceforge.net/projects/genomicparsec.

Contact: odile.lecompte@unistra.fr

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

Received on May 6, 2013; revised on July 16, 2013; accepted on August 2, 2013

1 INTRODUCTION

Genomic sites are short genomic regions with a defined biological function (e.g. regulation of gene expression, splicing or epigenetic signals). Most genomic sites are represented by degenerated motifs with a scattered distribution and can either be specific to a given species or conserved between species. The characterization of these sites is a major challenge in the exploitation of next-generation sequencing data and in understanding genome expression. The *in silico* detection of these motifs in complete genomes is associated with a huge amount of noise. Consequently, the development of a computational platform for accurate definition of genomic sites requires the integration of various large-scale biological data resources to filter out false positives.

PARSEC (PAtteRn Search and Contextualization) is a modular web service designed for the rapid localization and characterization of genomic sites. The main program exploits an efficient data structure, namely, compressed suffix trees (CST) (Sadakane, 2007), to rapidly localize sequences or degenerated patterns in complete eukaryotic genomes (eight species are currently available: human, mouse, rat, chicken, zebrafish, drosophila, nematode and yeast). This pattern search module is linked to three in-house modules for conservation analysis, genomic context analysis and functional filtering, as well as to the GoMiner functional enrichment tool (Zeeberg *et al.*, 2005). It can be used to filter biologically meaningful genomic sequences, to complement transcriptomic data analysis and to further characterize known genomic sites.

In contrast to web services dedicated to genomic pattern searches like TagScan (Iseli *et al.*, 2007), or to functional interpretation of genomic regions like GREAT (McLean *et al.*, 2010), PARSEC proposes both search and characterization aspects in a single intuitive interface, guiding the user through the discovery path.

2 ARCHITECTURE AND IMPLEMENTATION

2.1 Web infrastructure

The PARSEC web service is a modular infrastructure composed of five modules (Fig. 1). These modules can be combined in various 'discovery pipelines', exploiting the relevant tools and results at each step of the analysis.

The pattern search module is based on a CST implementation (Valimaki *et al.*, 2007), which we extended to facilitate degenerated pattern searches (interpretation of basic regular expressions and recursive navigation of tree edges) and parallelization of chromosomal searches (management of an array of CST representing chromosomes). It is adapted for easy use as a native library for a Java program. The user can submit degenerated patterns (minimum of 5 bp) using the IUPAC code and can additionally allow up to two mismatches. The maximum number of hits for further characterization is set to 100 000 on the web server owing to memory limitations, but can be easily increased in a local installation. For the same reason, conservation analysis cannot be performed for patterns defined with one or two mismatches.

The conservation module, based on BlastZ pairwise alignments provided by University of California, Santa Cruz (UCSC) (Dreszer *et al.*, 2012), provides fast evolutionary screening and characterization of the identified genomic sites. The analysis can be customized, because of the three hierarchized levels of conservation stringency (from perfect alignment of query and target sites to conservation of the region around the query site) and the possibility to select the minimum number of organisms in which a site must be conserved. Depending on the biological question, the user can select phylogenetically close organisms for *phylogenetic shadowing* or more distant ones for *phylogenetic footprinting*.

*To whom correspondence should be addressed.

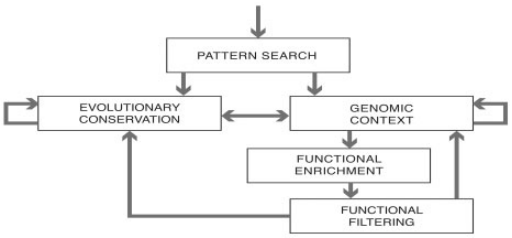

Fig. 1. Non-linear organization of modules in PARSEC, allowing multiple exploitation scenarios. The user chooses the analysis path according to results obtained at each previous step

The genomic context module, also based on data from UCSC tables (ensGene, EnsemblSource, ensemblToGeneName tRNAs), allows the detection of gene candidates spatially linked to the site of interest. Again, a high level of customization is provided. Analyses can be performed relative to the two gene boundaries or to the transcription start site. Several types of genes are available: protein coding, miRNA, snRNA, snoRNA, tRNA, etc. The user can retrieve all genes in proximity of the site or can choose only the closest gene to the site. The full set of alternative transcripts of proximal genes with their relative distances to the site is also provided and can be visualized through a link to UCSC.

For the enrichment module, we developed a layer allowing the use of GoMiner as a simple Java library. It allows the functional characterization of a previously identified set of proximal genes and query sites. This can be useful for identification of processes regulated by a transcription factor for example.

The functional filtering step allows the selection of sites potentially related to specific molecular functions, biological processes or cellular components.

At each step, the results can be saved as browser extensible data (BED) or comma-separated values (CSV) format files and include various links to external resources [UCSC, Ensembl (Flicek *et al.*, 2012)] for additional information. The algorithmic complexities and running times of the different modules are provided in Supplementary Dataset S1.

2.2 Installing PARSEC on a local server

The PARSEC web infrastructure relies on an information manager that allows automated retrieval of primary data (genomes, genes and BlastZ alignments) from UCSC and their integration in the PARSEC database. A new species can easily be added to PARSEC by specifying its NCBI taxonomy identifier, its UCSC genome version and some other parameters. This genome information is added to the PARSEC startup genome loading configuration file, so that the new genome is directly available after a simple web service restart. Genetic information is added to the PARSEC database, using a set of intelligent parsers, which check, for example, whether all the required gene types were found for the given organism and, if necessary, add the missing gene types. Each database entry is labeled with its source, and genetic or alignment entries can be easily updated by running retrieval commands on existing organisms. The source code, the information manager command line program and the .WAR archive for deployment on a Tomcat server are all freely available.

3 CASE STUDY

PARSEC can be used for various exploitation scenarios. For instance, the user can retrieve the set of genes potentially regulated by a transcription factor. To illustrate this, we searched for the DR5 (direct repeats separated by 5 bp) Retinoic Acid Response Element (RARE) in the masked human genome (version hg19), using the consensus pattern RGKTSANNNNNRGKTSA (Lalevee *et al.*, 2011). We localized 14 251 sites in 5.3 s. Using the conservation module, we then selected the sites conserved ('Aligned site' and 'Conserved site' parameters) in at least one amphibian or fish species. We found 178 sites in 20.8 s. We then filtered the 104 sites located near a transcription start site (5000 bp upstream, 5000 bp downstream) in the human genome using the genomic context module (0.13 s). We then analyzed the nearby protein coding genes with the functional enrichment module. Several GO categories were identified (false discovery rate $< 10^{-5}$), including embryo development and retinoic acid receptor signaling pathway. These categories related to retinoic acid regulation (Kumar and Duester, 2010) demonstrate PARSEC's ability to perform biologically meaningful genomic site analysis. Another case study (Supplementary Dataset S2) illustrates the usage of mismatches in PARSEC to improve the definition of the neuron-restrictive silencer factor binding site.

ACKNOWLEDGEMENTS

The authors are grateful to Vincent Laudet and Cécile Rochette-Egly for helpful discussions during the development of PARSEC. They thank Julie Thompson for critical reading of the manuscript.

Funding: This work was supported by the Agence Nationale de la Recherche [*Puzzle-Fit*: 09-PIRI-0018-02 and BIPBIP: ANR10-BINF03-05].

Conflict of Interest: none declared.

REFERENCES

Dreszer,T.R. *et al.* (2012) The UCSC genome browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.

Flicek,P. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.

Iseli,C. *et al.* (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.

Kumar,S. and Duester,G. (2010) Retinoic acid signaling in perioptic mesenchyme represses Wnt signaling via induction of Pitx2 and Dkk2. *Dev. Biol.*, **340**, 67–74.

Lalevee,S. *et al.* (2011) Genome-wide *in silico* identification of new conserved and functional retinoic acid receptor response elements (direct repeats separated by 5 bp). *J. Biol. Chem.*, **286**, 33322–33334.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Sadakane,K. (2007) Compressed suffix trees with full functionality. *Theory Comput. Syst.*, **41**, 589–607.

Valimaki,N. *et al.* (2007) Compressed suffix tree—a basis for genome-scale sequence analysis. *Bioinformatics*, **23**, 629–630.

Zeeberg,B.R. *et al.* (2005) High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, **6**, 168.