BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 17 2008, pages 1959–1960
doi:10.1093/bioinformatics/btn361

Gene expression

Sircah: a tool for the detection and visualization of alternative transcripts

Eoghan D. Harrington[1] and Peer Bork[1,2,*]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg and [2]Max Delbrück Centre for Molecular Medicine, D-13092 Berlin, Germany

Received on February 22, 2008; revised on May 30, 2008; accepted on July 14, 2008

Advance Access publication July 17, 2008

Associate Editor: Trey Ideker

ABSTRACT

Summary: Sircah is a flexible tool for the detection, analysis and visualization of alternative transcripts. It takes as input gene models or spliced alignments and creates a database of alternative transcription events: alternative transcription initiation and polyadenylation, alternative 3′ and 5′ splice-site usage, skipped exons and retained introns. The results can be visualized in a variety of ways, allowing the creation of publication quality images.

Availability: The Sircah is available for download under a creative commons license along with additional documentation and a tutorial from http://www.bork.embl.de/Sircah.

Contact: bork@embl.de

1 INTRODUCTION

The development of high-throughput DNA sequencing technologies has been instrumental in uncovering the diversity of transcripts encoded in genomes. Already the sequencing of whole (full-length cDNA) and partial (EST) transcripts has revealed the extraordinary diversity of mammalian transcriptomes. Now, with the arrival of next-generation sequencing technologies, the study of these alternative transcripts is about to undergo another revolution. Initial studies suggest that the data generated by these technologies will dwarf the already considerable amounts of EST and cDNA data, not only increasing the depth (transcript abundance) and breadth (spatial/temporal/conditional) of transcriptome sampling, but also allowing the quantification of transcript levels (Emrich *et al.*, 2007). These data have the potential to provide a wealth of information on the regulation, function and conservation of alternative transcripts, however in order to realize this novel tools will have to be developed. Here we present Sircah, a tool designed to cope with the increasing scale of transcript data, to be flexible enough to permit novel types of alternative transcript analysis and provide compact visualizations of the results.

The major difference between Sircah and existing methods to detect alternative transcripts is that it is a downloadable tool. There are already large number of methods that use pipelines to generate web-accessible databases of alternative transcript data (Kim *et al.*, 2006; Stamm *et al.*, 2006); however, due to their complexity these methods are difficult to distribute and are therefore unavailable for individual research groups. Alternatively, a set of alignments can be submitted to the ASGS web server (Bollina *et al.*, 2006) for analysis; however, as this is dependent on the availability of third-party compute resources, it may not scale well to the level of whole transcriptome analysis. Sircah, on the other hand, is run from the commandline, making it suitable for batch use and therefore may be used for the analysis of single genes up to whole transcriptomes. The data models used in these analyses may be stored in a relational database, allowing flexible reanalysis of the data either by creating different visualizations or by comparing the alternative transcripts present in different subsets of the data. Additional flexibility is provided by the fact that Sircah is implemented as a Python package, allowing users to devise novel ways of analysing the data.

2 PROGRAM OVERVIEW

2.1 Input

Sircah takes as input transcript models in the GFF3 format allowing the user the flexibility to choose the sources of evidence for the use in detecting alternative transcription. Such transcript models may come from the gene prediction pipelines of genome databases or from spliced alignments of ESTs or proteins against the genome. Within the GFF3 file, the user may also specify the completeness of the transcript model used and may provide a set of tags, which can later be used to analyse subsets of data (Fig. 1).

2.2 Detection of alternative transcripts

Sircah uses a splice graph data model as first proposed by Heber *et al.* (2002) to represent the transcripts models in a non-redundant form. The nodes of the directed graph are exons, the edges introns and transcripts are represented as subpaths of the graph. Additionally overlapping exons are clustered into superexons. A series of rules are then applied to the data and based on the topology of the splice graph and the membership of the superexons, the following alternative events can be classified: alternative initiation exons, alternative termination exons, exons with alternative 3′ and/or 5′ splice sites, retained introns and skipped exons. A detailed description of the rules used is available on the Sircah website.

2.3 Visualization of alternative transcriptions

The splice graph and the transcript models used to construct it can be visualized in a variety of ways to show: (i) the alternative events detectable and the transcript models used, (ii) the different events

[*]To whom correspondence should be addressed.

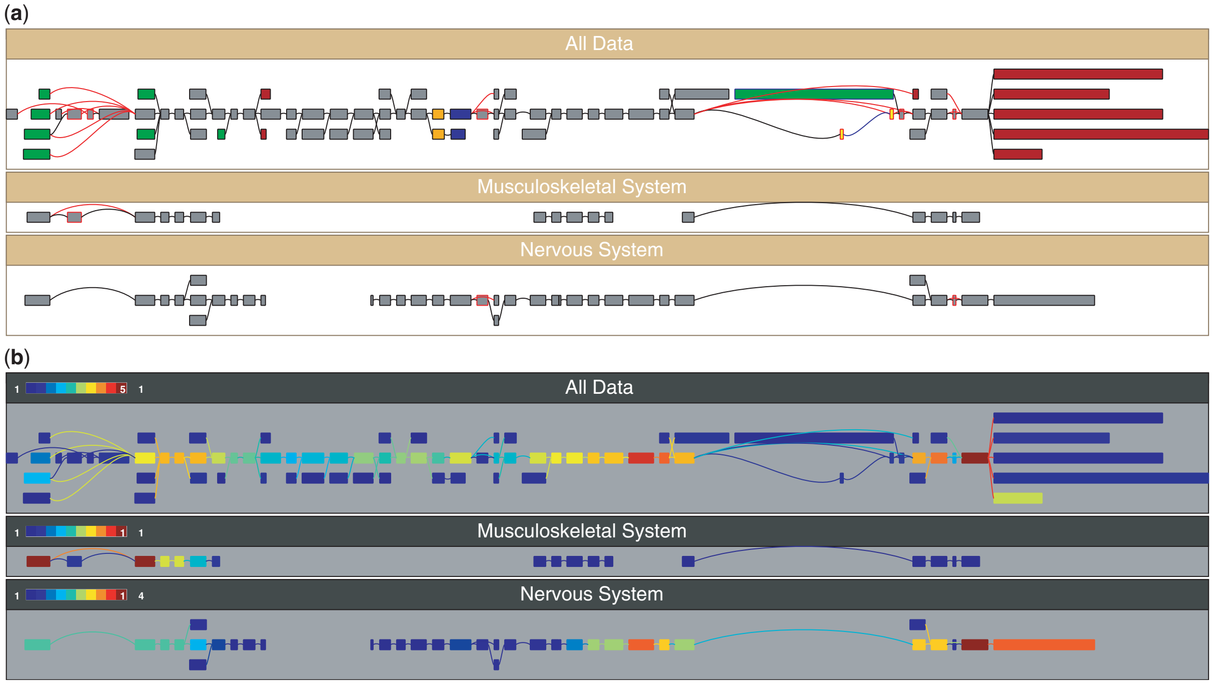

Fig. 1. Sircah visualizations of the myosin 6 gene. (**a**) The splice graph constructed from EST and cDNA alignments. The introns and exons are coloured according to the type of alternative transcription event that they are involved in, for instance skipped exons are outlined in red. The full colour scheme is available on the Sircah website. The bottom panels show the splice graphs created from the subset of ESTs that are tagged as brain or muscle in the EVOC ontology (Kelso *et al.*, 2003). (**b**) The same plot as in (a) but this time coloured according to EST coverage rather than alternative event. These figures are created as part of tutorial on the Sircah website.

detectable in subsets of the data and (iii) the coverage of introns and exons by transcript models (Fig. 1). The visualization is created in the SVG format, allowing the creation of publication quality images. In addition, the IDs of all the elements in the SVG file are directly mappable to the IDs used in the data objects, allowing the user to alter graphic after it is generated and even create interactive graphics using javascript.

2.4 Analysis of subsets of evidence

One of the most powerful features of Sircah is the ability to create splice graphs based on a subset of the total data. This allows the user to compare alternative transcription events under different conditions. For example, by tagging the EST alignments with the EVOC expression ontology (Kelso *et al.*, 2003) one can examine the tissue distribution of alternative transcription events (Fig. 1).

2.5 Data Serialization

In order to be able to carry out such analyses, it is important to be able to save and reload the data models described above. To facilitate this Sircah can serialize its data to either an XML file using the Elementtree python module or to a relational database (sqlite3, MySQL and PostgreSQL are currently supported) using the SQLAlchemy python module.

ACKNOWLEDGEMENT

We would like to acknowledge the helpful suggestions given by the Bork group, E. Kriventseva and N. Paul.

Funding: This work was supported by the European Commission through the ASD (LSHG-CT-2003-503329) and Eurasnet (LSHG-CT-2005-518238) grants.

Conflict of Interest: none declared.

REFERENCES

Bollina,D. *et al.* (2006) Asgs: an alternative splicing graph web service. *Nucleic Acids Res.*, **34** (Web Server issue), W444–W447.

Emrich,S.J. *et al.* (2007) Gene discovery and annotation using lcm-454 transcriptome sequencing. *Genome Res.*, **17**, 69–73.

Heber,S. *et al.* (2002) Splicing graphs and est assembly problem. *Bioinformatics*, **18** (Suppl. 1), S181–S188.

Kelso,J. *et al.* (2003) evoc: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.

Kim,N. *et al.* (2006) The asap ii database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, **35** (Database Issue), D93–D98.

Stamm,S. *et al.* (2006) Asd: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34** (Database issue), D46–D55.