An automated approach to the in-silico identification of chimeric mRNAs

Alberti S (1), Trerotola M (1), Emerson A (2), Rossi E (2)

(1) Unit of Cancer Pathology, Center for Excellence in Research on Aging, University "G. D' Annunzio", Via Colle dell' Ara, 66013 Chieti Scalo (Chieti), Italy.
(2) High Performance Systems Division, CINECA, via Magnanelli 6/3, 40033 Casalecchio di Reno (BO), Italy.

Motivation

Chimeric mRNAs from two different genes largely arise by mRNA trans-splicing.
mRNA transsplicing post-transcriptionally joins heterologous mRNAs at canonical exon-exon borders, essentially following the rules of canonical cis-splicing. As the construction of cDNA libraries frequently causes cDNA fusion artefacts, largely because of incorrect ligation of independent cDNAs or of abnormal reverse-transcription, a key issue is how to distinguish between bona fide chimeras and in vitro artefacts. We have developed a bioinformatics retrieval strategy, the In Silico Trans-splicing Retrieval System (ISTReS), in order to distinguish between the two. The ISTRes pipeline consists of the following steps: 1. Map the cDNA databank onto the human genome by Blast analysis, masking human repetitive DNA. 2. Filter the Blast output according to score, match length and percentage identity. 3. Group the query sequence segments (mRNA exons) in longer concatamers, each mapping only onto one chromosome. 4. Check for possible chimeric sequences by comparing concatamers. 5. Remove possible cDNA fusion artefacts (e.g. 'sense/antisense' sequences). 6. Structural analysis of remaining sequences to locate mRNA cleavage or poly-A addition signals to provide further evidence of chimeric joins. The procedure has been successfully validated against a set of known chimeric sequences and has also detected two novel chimeric mRNAs [1]. The authors of this work estimate that about 1% of the hybrid sequences in current mRNA databanks are canonically trans-spliced. The aim of this study was to extend the ISTReS procedure to larger datasets.

Methods

Steps 2-6 of the trans-splicing detection system were implemented with custom Perl scripts, many of them re-written for efficiency and to reflect changes in strategy since the previous study. Although computationally inexpensive the algorithms are often quite complex and most of the programs have undergone major revisions. Indeed, we have found that progress in developing the trans-splicing retrieval system for larger datasets does not depend on the computationally intensive Blast analysis but instead on the validation of the analysis programs. In order to validate the ISTReS procedure the scientific experts in the team need to be able to execute each individual component of the pipeline as well as the whole pipeline itself. The situation is complicated by the requirements for supercomputing resources and large data storage, thus necessitating direct logon access to the computers in question. The common technique of providing a web-interface to hide the underlying computer implementation is in impractical for such a complex system which is still evolving and being tested. To accelerate the refining of the ISTReS procedure and to provide a more convenient environment for the end-user, it was decided to create a workflow description of the pipeline and to implement the various components as web services. The workflow was constructed with the Taverna workflow editor, while the web services were created with the Soaplab environment. The latter is particularly convenient because it generates web services by "wrapping" already existing programs, thereby avoiding re-programming of the applications. Note that due to difficulties in implementing asynchronous web services with available tools, for the moment the Blast analyses have not been exported as web services.

Results

We show below an image of an example Taverna workflow which implements some of the key steps of the ISTReS pipeline. This and similar workflows are currently being used to refine some of the analysis steps in ISTReS. Candidate chimeras identified by ISTReS analysis with selected cDNA databanks will be reported in a future work.

Contact email: a.emerson@cineca.it