On the experimental annotation of tomato BACs sequences: reliable alignment for useful genomic analyses

Traini A (1,2), Chiusano ML (2)

(1) PhD fellow in Computational Biology, Interdepartmental Research Center for Computational and Biotechnological Sciences, Second University of Naples, Naples, Italy
(2) Department of Structural and Functional Biology, University "Federico II", Naples, Italy

Motivation

Curated gene annotation is a challenging computational problem in genomics. Reliable results are commonly achieved with spliced alignment of full-length cDNAs or expressed sequence tags (ESTs) with sufficient overlap to cover the entire mRNA. Moreover, predictive approaches are based on curated Gene Models obtained by experimental effort too. Many standalone programs are available for mapping and aligning expressed tags to a genome sequence. With the aim to contribute to the bioinformatics efforts of the International Tomato Genome Sequencing project we tested the software proposed by the International Committee for cDNA/EST to genome mapping. We investigated which algorithm is more reliable and in which context, comparing and evaluating some of the most frequently used specialized software to provide a trusty Reference dataset of Tomato Gene Models.

Methods

We tested sim4 [1], based on the BLASTZ algorithm [2], Galahad, included in the grail-exp package [3], blEST [1] and SIBsim4 [1], ALL based on sim4 algorithm [1], and GeneSeqer [4][5] as software commonly used for solving the task of mapping cDNAs/ESTs to genomic sequences. Software results depend on the specific parameter usage but also on specific similarity thresholds. We compared the resulting data from different software considering all common results and discussing each software specific feature. To provide a reliable informative benchwork for Tomato genome annotation based on experimental results, we set up a Gbrowse [6] based platform reporting the results from the different methods. The datasets used in the present analysis are the Tomato expressed sequences available from dbEST [7] and from the genome sequencing effort at the SOL Genomics Network (SGN)[8].

Results

This work summarizes the analysis and the results of specific software solving the cDNA/EST to genome mapping problem. Different algorithms and even different parameters and similarity thresholds influence the quality of the resulting alignments. We present here our evaluation of the software considered and we propose multiple algorithm usage under different constraints to provide exhaustive and reliable information when experimentally annotating genome sequence data. indeed To further support annotators, we provide different results: i) best alignments with a low error margin of the genomic region alignment and ii) and less stringent results extending the number of retained alignments. This may be useful because highly scored complete EST alignment as well as medium level similarities and partial alignments per EST may be of interest, either when considering reliable gene models predicted by experimental data or when looking for related gene loci in evolutionary analysis.

Availability: http://www.cab.unina.it/

Contact email: chiusano@unina.it

References

1. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. Sep;8(9):967-74 (1998).
2. S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller, Human-Mouse Alignments with BLASTZ, Genome Res. Vol 13, Issue 1, 103-107, January

(2003).
3. D. Hyatt, J. Snoddy, D. Schmoyer, G. Chen, K. Fischer, M. Parang, I. Vokler, S. Petrov, P. Locascio, V. Olman, Miriam Land, M. Shah, and E. Uberbacher, Improved Analysis and Annotation Tools for Whole-Genome Computational Annotation and Analysis: GRAIL-EXP Genome Analysis Toolkit and Related Analysis Tools, Genome Sequencing & Biology Meeting, May 2000
4. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinformatics. May 1;20(7):1157-69. Epub 2004 Feb 5 (2004).
5. Shannon D. Schlueter, Qunfeng Dong and Volker Brendel, GeneSeqer@PlantGDB: gene structure prediction in plant genomes, Nucleic Acids Research, Vol. 31, No. 13, 3597-3600, (2003).
6. Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E. Stajich, J.E., Harris, T.W., Arva, A. and Lewis S. (2002) The generic genome browser: A building block for a model organism system database. Genome Res. 12, 1599-1610.
7. Boguski M.S., Lowe T.M., Tolstoshev C.M.,dbEST-database for "expressed sequence tags", Nat Genet., Aug; 4(4): 332-333 (1993). home-page: http://ncbi.nih.gov/dbEST/
8. home-page: www.sgn.cornell.edu

Supplementary informations

This work is supported by the Agronanotech Project (MIPAF, ITALY)