e2g - A Web-Based Tool for Efficiently Aligning Genomic Sequence to EST and cDNA data

Alexander Sczyrba, Jan Krüger, Robert Giegerich [1]

Keywords: EST alignment, gene structure prediction, suffix array

1 Introduction.

High throughput cDNA and EST sequencing projects have generated a vast amount of data representing the transcribed portion of the organisms in study. As soon as (parts of) the sequence of the associated genome becomes available, gene structures can be determined by mapping the cDNA data to the genomic sequence. This allows the detection of genes missed by gene prediction tools and the determination of splice variants of already known genes.

While several tools for mapping ESTs and cDNAs to genomic sequence already exist [1, 2, 3], they can hardly be used in an interactive web-based application because of the huge amount of data to be searched against. `e2g` is a web-based tool which efficiently aligns genomic sequence to indexed cDNA and EST databases. This allows users to rapidly detect the exon-intron structure of genes, including variants, in the genomic region of interest.

`e2g` is online available on the Bielefeld University Bioinformatics Server at:

`http://bibiserv.techfak.uni-bielefeld.de/e2g/`

2 Method.

The web interface accepts either (i) genomic sequence or (ii) genomic sequence and cDNA/EST data as input. In the first case the sequence will be matched against a database of cDNAs and ESTs. (Currently, databases for human and mouse are available.) In the second case, the user provides the cDNA data to be matched against the genomic sequence. In both cases, repeats and low-complexity regions will be masked before matching.

As good efficiency is critical for the approach, an enhanced suffix array is built on the server as an persistent index of the EST sequences using `mkvtree` [4, 5]. The index efficiently represents all substrings of the database sequences and allows to solve matching tasks in time independent of the size of the index, done using the string matching algorithm `vmatch` [4, 5]. Matches can be computed either exactly, or approximately by extending the exact seeds using the X-drop strategy [6]. Matching the 16.5kb genomic sequence of the example in figure 1 against all mouse EST data (approx. 2.5 GB), takes 50 seconds on a SUN UltraSparc III (800 Mhz) with 64 GB RAM. Allowing mismatches using the X-drop strategy (99% identity, seedlength 15) increases the running time to 73 seconds.

3 Web interface.

Figure 1 shows a screenshot of the `e2g` web interface. Matches of the ESTs reveal the exon-intron structure of the gene. On the top the annotation of the submitted genomic region is shown. This can be overlayed by dragging a transparent image over the lower part of the window, allowing the user to easily compare the annotated gene structure to the matches

[1] Technische Fakultät, Bielefeld University, D-33594 Bielefeld, Germany.
E-mail: {asczyrba,jkrueger,robert}@TechFak.Uni-Bielefeld.DE

found. Information about the matches of the circled exons is shown on top. The alignment of each match can be calculated on the fly, reusing the existing index (see popup window in the lower left).

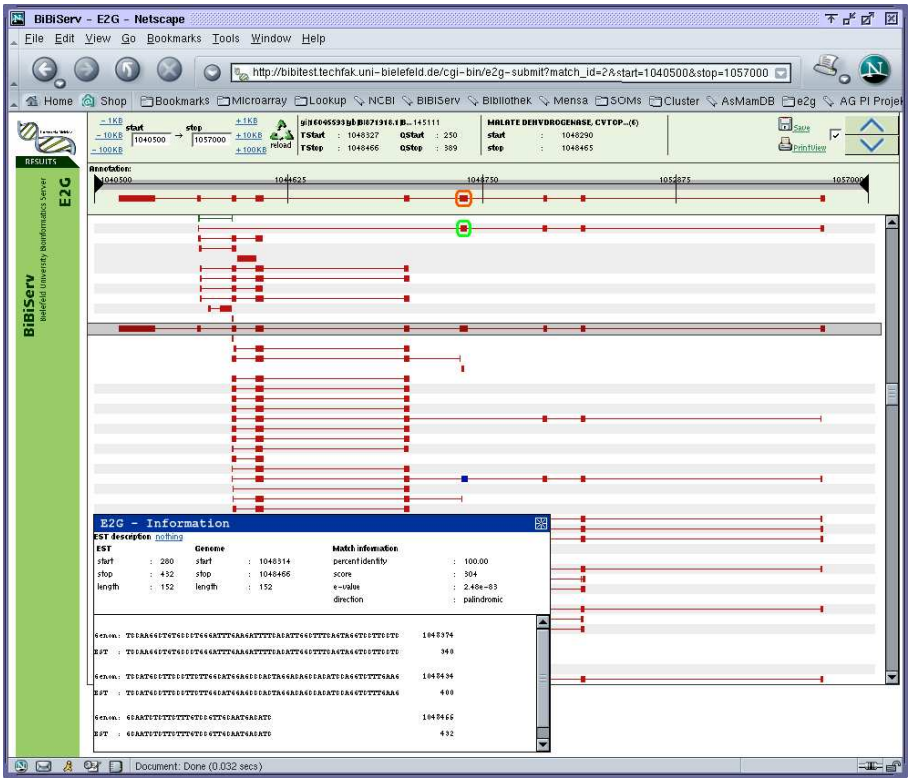

Figure 1: Screenshot of the `e2g` web interface.

References

[1] R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS*, 13(4):477–8, 1997.

[2] W.J. Kent. BLAT-The BLAST-Like Alignment Tool. *Genome Res.*, 12(4):656–664, 2002.

[3] C. Del Val, K.H. Glatting, and S. Suhai. cDNA2Genome: A tool for mapping and annotating cDNAs. *BMC Bioinformatics*, 4(1):39, 2003.

[4] M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch. The Enhanced Suffix Array and its Applications to Genome Analysis. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics*. Springer-Verlag, Lecture Notes in Computer Science, 2002.

[5] M.I. Abouelhoda, E. Ohlebusch, and S. Kurtz. Optimal Exact String Matching Based on Suffix Arrays. In *Proceedings of the Ninth International Symposium on String Processing and Information Retrieval*, pages 31–43. Springer-Verlag, Lecture Notes in Computer Science 2476, 2002.

[6] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7:203–14, 2000.