IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10406-8

Colloquia: CSFI 2008

Comparative Genome Viewer

I. Molineris(*)(**) and G. Sales(**)
Dipartimento di Fisica Teorica, Università di Torino - Torino, Italy

(ricevuto il 10 Settembre 2009; pubblicato online il 25 Settembre 2009)

Summary. — The amount of information about genomes, both in the form of complete sequences and annotations, has been exponentially increasing in the last few years. As a result there is the need for tools providing a graphical representation of such information that should be comprehensive and intuitive. Visual representation is especially important in the comparative genomics field since it should provide a combined view of data belonging to different genomes. We believe that existing tools are limited in this respect as they focus on a single genome at a time (conservation histograms) or compress alignment representation to a single dimension. We have therefore developed a web-based tool called Comparative Genome Viewer (CGV): it integrates a bidimensional representation of alignments between two regions, both at small and big scales, with the richness of annotations present in other genome browsers. We give access to our system through a web-based interface that provides the user with an interactive representation that can be updated in real time using the mouse to move from region to region and to zoom in on intereseting details.

PACS 42.62.Be – Biological and medical applications.
PACS 07.05.Rm – Data presentation and visualization: algorithms and implementation.
PACS 87.18.Wd – Genomics.

1. – Introduction

The availability of the complete sequence of the human genome [1] opened a new era for geneticists. With thousands of microbial [2] and over 150 eukaryotic genomes sequenced within the last few years, we are now provided with a wealth of DNA sequences. Comparative genomics is an attempt to take advantage of such situation; the results achieved in this field are affecting most areas of biology [3].

As the amount of information about genomes increases exponentially, there is the need for advanced bioinformatic tools to analyze them [4]. Sequence alignment is one

(*) E-mail: molineri@to.infn.it
(**) These authors contributed equally to this work.

of the most widespread techniques used to take advantage of the diversity of the available information. It is a way of arranging two or more sequences in order to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships [4].

In this context visualization plays an essential and defining role: providing a graphical representation of complex information being both comprehensive and intuitive it is a critical point in scientific research [5]. For example, the display of the experimental data required for the annotation of genomes has been the subject of extensive reaserch [6]. Genome browsers, such as the one provided by the University of California Santa Cruz (UCSC), collect and organize such data and make it available in various formats allowing the user to extract and visualize useful information.

In this paper we propose a new method to visualize genomic alignments and to integrate the information that they give with already known annotations about them available through the UCSC genome browser.

A variety of computational approaches have been applied to the problem of sequence aligment, including slow but exhaustive methods like dynamic programming [7], and efficient, but not as thorough heuristic or probabilistic algorithms designed for large-scale database searches [8].

The specific search strategy notwistanding, almost all available softwares represent alignments textually, using rows arranged so that matching positions (characters) appear in the same column. Other graphical representations are limited to histograms conveying the percentage of sequence conservation (or some other similarity measure) among species in each given genomic position [9].

Our goal is to provide a richer graphical representation of alignments in order to allow the researcher to visually identify certain sequence features such as insertions, deletions, segment-shufflings and repeats. Having such visualization at our disposale, both at the scale of small sequences (10 to 100 basepairs) up to hundreds of thousands of nucleotides, it becomes possible to quickly hypothesize wath kind of processes (other than point mutations) operated on sequences during evolution [10].

2. – Comparative Genome Viewer

Our tool, the Comparative Genome Viewer (CGV), is inspired by another visualization scheme, one of the first that have been developed in the field: the dot-plot [11]. A dot-plot implicitly produces a family of alignments for a pair of sequences. It is qualitative and simple, though it rapidly becomes impractical when the scale increases. To plot one, two sequences are written along the two sides of a matrix and a dot is placed at any point where the characters in the corresponding row and column match.

This tecnique has, however, at least three big drawback: 1) there is no statistical indication of the relevance of alignments, 2) the graphical representation is very noisy (diagonal lines corresponding to relevant alignments are mixed among many small lines and dots deriving from random alignments), 3) filling the matrix becomes very time demanding as the size increases (in fact, it scales as N^2).

To overcome these issues we start by performing whole genome, high-sensitivity alignments using the Basic Local Alignment Search Tool (BLAST) software [8]. BLAST finds homologous sequences by means of a heuristic method that locates short exact matches between them. This heuristics allow BLAST to skip a full (and potentially very slow) comparison at the cost of a small loss in sensitivity. Moreover BLAST statistically score

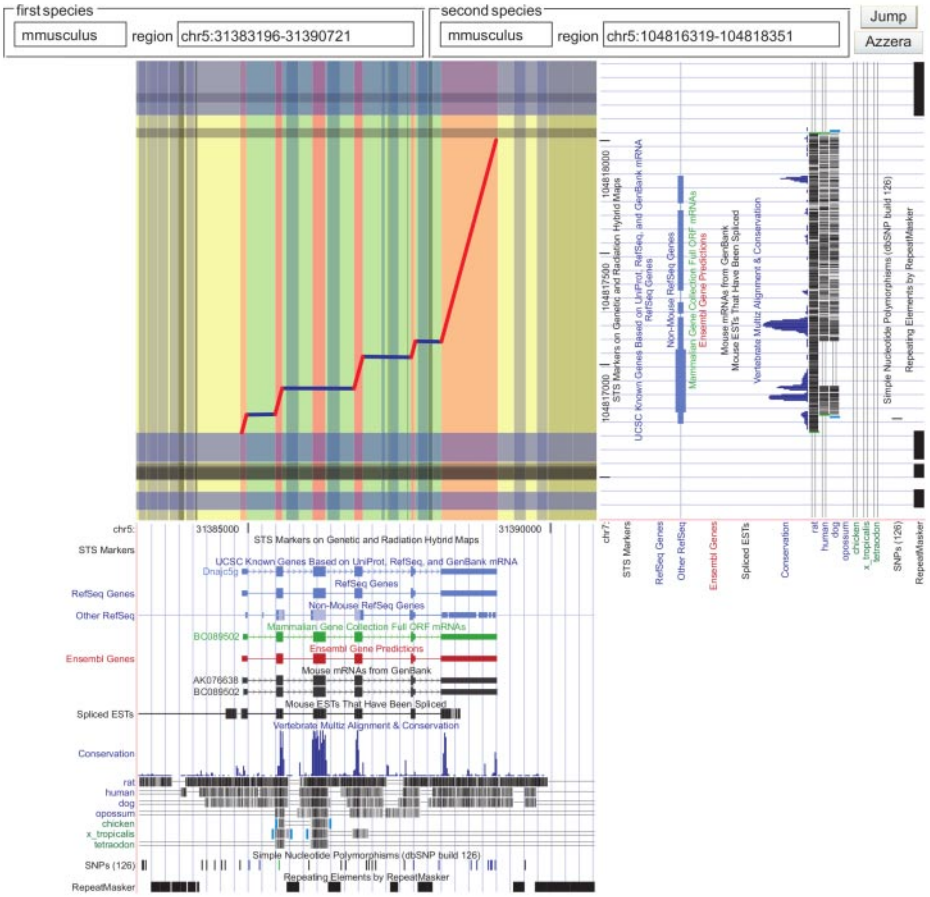

Fig. 1. – (Color on-line) Screenshot of the CGV interface. The central panel shows the dot-plot–like graph. The tracks given by the UCSC genome browser are reported alongside the central panel ad as stripes on the central panel. The aligments (red segment on the foreground of the central panel) coincide with the exons of a gene on the horizontal axis. The graph represents the aligment of a gene with a processed pseudogene [13] on the same genome: our visualization algorithms correctly recognize the missing introns of the pseudogene as gaps (blue segments) that coincide with the introns of the gene.

the likelihood of a particular alignment; it is thus possible to filter out results most likely arosen just by chance [12].

We represent each alignment as a segment in a Cartesian plane whose axes correspond to the two sequences being aligned: the resulting image is much cleaner than those produced by a standard dot-plotter; each reported alignment is statistically significant. It is moreover possible to compute such graph for long sequences, for which a standard dot-plot would require both an unfeasible amount of time and image size.

Our approach is very fast because all alignments are precomputed. We cluster the alignments in a hierarchical structure to store them efficiently in a relational database; this form has the added benefit of making queries extremely fast (up to 40 times with respect to the naïve format). Given a pair of regions in chromosomal coordinates, our

system recovers all the overlapping alignments (segments) and presents them graphically with usually no appreciable delay; several seconds or minutes would be required, on the other hand, to run the BLAST software.

Another CGV feature is the integration of existing knowledge about the sequences.

We import in our database the most relevant information about gene structures [14] and DNA-repeats [15] from the UCSC genome browser and use them to plot the background of CGV output. Each annotated feature is represented as a colored horizontal or vertical stripe so that its projection with the corresponding axis determines the genomic segment originally tagged with such feature. As a result, our graphical representation allows, for example, to immediatelly understand whether an aligment starts on both sequences in correspondence with an intron-exon junction.

To further integrate external information, we recover gene annotations and other tracks from the UCSC genome browser and we show them at both sides of our plot (see fig. 1).

* * *

We are grateful to M. Caselle that supported this work.

REFERENCES

[1] Subramanian G., Adams M. D., Venter J. C. and Broder S., *JAMA*, **286** (2001) 2296.
[2] Markowitz Victor, Szeto Ernest, Palaniappan Krishna, Grechkin Yuri, Chu Ken, Dubchak Inna, Anderson Iain, Lykidis Athanasios, Mavromatis Konstantinos, Ivanova Natalia and Kyrpides Nikos, *Nucleic Acids Res.*, **36** (2007) D528-3, 10.1093/nar/gkm846.
[3] Woodage Trevor and Broder Samuel, *J. Gastroenterol.*, **38** Suppl. 15 (2003) 68.
[4] Ponting Chris P. and Goodstadt Leo, *J. Cell Sci.*, **116** (2003) 6, 10.1242/jcs.00197.
[5] Wilkinson Leland, *The Grammar of Graphics* (Springer) 2005.
[6] Pevsner Jonathan, *Methods Mol. Biol. (Clifton, N.J.)*, **537** (2009) 277, 10.1007/978-1-59745-251-9_14.
[7] Smith T. F. and Waterman M. S., *J. Mol. Biol.*, **147** (1981) 195.
[8] Altschul S., Gish W., Miller W., Myers E. and Lipman D., *J. Mol. Biol.*, **215** (1990) 403, 10.1006/jmbi.1990.9999.
[9] Mayor Chris, Brudno Michael, Schwartz Jody, Poliakov Alexander, Rubin Edward, Frazer Kelly, Pachter Lior and Dubchak Inna, *Bioinformatics*, **16** (2000) 1046, 10.1093/bioinformatics/16.11.1046.
[10] Brudno Michael, Malde Sanket, Poliakov Alexander, Do Chuong B., Couronne Olivier, Dubchak Inna and Batzoglou Serafim, *Bioinformatics (Oxford, England)*, **19** Suppl. 1 (2003) i54.
[11] W. Mount David, *Bioinformatics: Sequence and Genome Analysis* (CSH Press) 2004, Chapt. 3, p. 76.
[12] Karlin S. and Altschul S. F., *Proc. Natl. Acad. Sci. U.S.A.*, **87** (1990) 2264.
[13] Mighell A., Smith N., Robinson P. and Markham A., *FEBS Lett.*, **468** (2000) 109.
[14] Gerstein Mark, Bruce Can, Rozowsky Joel, Zheng Deyou, Du Jiang, Korbel Jan, Emanuelsson Olof, Zhang Zhengdong, Weissman Sherman and Snyder Michael, *Genome Res.*, **17** (2007) 669, 10.1101/gr.6339607.
[15] Richard Guy-Franck, Kerrest Alix and Dujon Bernard, *Microbiol. Mol. Biol. Rev.*, **72** (2008) 686, 10.1128/MMBR.00011-08.