Future Generation Computer Systems 26 (2010) 441–454

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

VisGenome with CartoonPlus: Supporting large scale genomic analyses via physical space deformation

Joanna Jakubowska [a,*], Ela Hunt [b], Matthew Chalmers [a]

[a] Department of Computing Science, University of Glasgow, UK

[b] Department of Computer and Information Sciences, University of Strathclyde, UK

ARTICLE INFO

Article history:
Received 15 October 2008
Received in revised form
26 June 2009
Accepted 3 August 2009
Available online 8 August 2009

ABSTRACT

We focus on visualisation techniques used in genome browsers and examine the available browsers with respect to their suitability for comparative genome analysis, and the legibility of display. Based on this investigation, we then report on a new technique, CartoonPlus, which improves the visual representation of data. We describe our use of smooth zooming and panning, and a new scaling algorithm and *focus on* options. CartoonPlus scaling allows the users to see data not in a physical size, but deformed to improve data legibility, depending on the data type chosen by the users. In VisGenome we have chosen genes as the basis for scaling. All genes have the same size and all other data is scaled in relationship to genes. Additionally, objects which are smaller than genes, such as micro array probes or markers, are scaled differently to reflect their partitioning into two categories: objects in a gene region and objects positioned between genes. This results in a significant legibility improvement and should enhance the understanding of genome maps. The technique may be useful in other information rich contexts, such as comparison of histograms or schema mapping.

1. Introduction

Medical researchers find it difficult to locate the correct biological information in the large amounts of available biological data and put it in the right context. Visualisation techniques are of great help to them, as they support data understanding and analysis. A great many genome browsers have been developed in the recent years, and we examined them to find the right browser or the best visual representation to support comparative genome analysis. This examination showed that none of the browsers satisfied our requirements and a new browser was needed. In response to this need, we developed a prototype of a new genome browser, VisGenome, which combines the available techniques.

The requirements for VisGenome [1–3] were identified in cooperation with medical researchers from a hospital. The users require visualisation support in two work scenarios. In single genome analysis they want to see large genome areas, such as quantitative trait loci (QTLs)[1] and chromosome bands, and small objects, such as microarray probes coloured and positioned to show which experiment they represent and how they relate to known genes and QTLs. In comparative analysis, the users want to view QTLs, genes and microarray probes in an area of synteny (see Section 3) in a way that makes it clear where the orthologous genes are positioned and what the gene relationships are. In both contexts it was important to lay out the information legibly and to enable intuitive navigation and perspective change, i.e. to show both context and detail at various levels of resolution. In looking for a suitable software solution we found that the majority of genome browsers show *only a selection of data for one chromosome* and do not adequately support comparative genome analysis, where genes from two species are being investigated. This is obvious, because the amount of available information is so large that it is impossible to show all data in one view. Expressionview [4], for example, shows QTLs and micro array probes and no other data. Some of the tools, such as Ensembl [5], show many types of data but use a number of different data views, which make the users disoriented and lost in the tool and data space. Moreover, Ensembl shows as much information as possible in one view, instead of offering a view or a panel with additional information. A large number of genome browsers show only a chromosome and do not allow one to see a comparison of two chromosomes from different species. Exceptions include SyntenyVista [6,7] and Cinteny [8] which show a comparative view of two genomes but are limited with regard to other data, such as micro array probes. On the other hand, SynView [9] visualises multi-species comparative genome data at a much higher level of abstraction.

* Corresponding author.
E-mail addresses: asia@dcs.gla.ac.uk (J. Jakubowska), ela.hunt@cis.strath.ac.uk (E. Hunt), matthew@dcs.gla.ac.uk (M. Chalmers).

[1] A quantitative trait locus (QTL) is a part of a chromosome which is correlated with a physical characteristic, such as height or disease. Micro array probes are used to test gene activity (expression).

0167-739X/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2009.08.007

Fig. 1. The comparative representation for the rat chromosome 18 and the mouse chromosome 18. The data is *scaled* by the scaling algorithm CartoonPlus which makes all genes the same size, and QTL size depends on genes. Genes ERF1_RAT and Etf1 are linked by a homology line and marked in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We aimed to find a solution which clearly presents all the available information, including all relevant information the biologists wish to see. The solution for data analysis should overcome both representational and cognitive problems.

Here, we describe single and comparative genome representations. A single representation is a view which shows data for one chromosome. A comparative representation illustrates relationships between two or more chromosomes.

Our contribution is a short review of the available browsers and a scaling algorithm which we call CartoonPlus. CartoonPlus allows the users to see data more clearly by choosing one kind of data as basis and scaling other data types in relationship to the basis. The solution does not show data in its natural size but allows one to see relationships between different kinds of data more clearly, especially in a comparative representation.

The paper is organised as follows. Section 2 provides the background about the existing browsers and visualisation techniques and their usefulness for medical researchers. Section 3 introduces the visualisation techniques we used in VisGenome and provides details of our new algorithm. We discuss our work in Section 4 and the last section concludes.

2. Related work

This section surveys the existing genomics visualisation software and then sets it in the context of known visualisation techniques. This justifies the need for our browser and clarifies what type of data scaling is necessary.

Here we survey the following genomics visualisation software: ACeDB [10], DerBrowser [11], Apollo [12], Artemis [13], BugView [14], SyntenyVista [6,7], Ensembl [5], Sockeye [15], K-BROWSER [16], GBrowse [17], NCBI Map Viewer [18], eQTL Explorer [19] and Expressionview [4]. We compare the systems in order to understand the problems and possible solutions to data visualisation. At the end of this section we present a table summarising our findings with regard to the users' needs.

ACeDB [10], see Fig. 3, was one of the first tools for genome visualisation. It offers a graphical representation containing many objects where various object types are coloured following a prescribed schema. The users can view textual details by double clicking on an object. ACeDB offers simple zooming activated via zoom buttons. It offers three types of sequence view: a genetic map (Fig. 3), a physical map, and a sequence window which shows the DNA or AA letters. All views offer pop-up menus. ACeDB has a comparative genome display, but this display does not offer scaling options or view adjustment, and is not suitable in our context. The web version uses server-side generated images and the options for view adjustment are limited.

DerBrowser [11], see Figs. 4 and 5, was designed at the time of the human genome sequencing project. It is a Java applet supporting interactive visualisation of one chromosome, or of a chromosome part. It can use a local database to produce web pages showing all the information describing a given map object. It displays genes, chromosome bands (chromosome parts coloured light or dark in the karyotype pictures), markers, or any other object on a map. It provides an illusion of smooth zooming (a slider), and supports the hiding of objects, based on object type. It also offers search functions. In tests with human and rat genome data, as required by our user group, in both single and comparative genome representations we observed that the scaling was not flexible enough and the comparative representation was illegible (not shown).

Artemis [13], see Fig. 6, is a genome viewer and annotation tool that visualises sequence features and the results of analyses within the context of the sequence, and its translation from DNA to protein. Artemis was designed for smaller genomes. Properties of the sequence can be plotted, and each plot allows dynamic modification of the window size used for the calculation. The sequence and plots can be zoomed together into the single base level or out for the complete genome. Artemis provides two sequence windows to view the same sequence at different zoom levels simultaneously. The tool can be run as an applet within a web browser. As no comparative genome views are offered, Artemis is not appropriate in our research context.

Apollo [12], see Fig. 7, is a sequence annotation viewer and editor. It allows the biologist to improve on the genomic feature descriptions derived from automated analyses and computational pipelines. It connects to various databases and supports the comparison of existing annotations with other biological data. Within the various views offered by the package, annotations can be created, deleted, merged, split, classified and commented upon. The tool allows the view to be scaled using zoom buttons and provides a degree of semantic zooming. Some features are not displayed at low zoom levels and appear more precisely only when the user zooms in on them. The users can move to a specific position by specifying a coordinate, gene name, or short sequence string, or by using the horizontal scroll bar. Apollo can display features on two genomes at the same time. In a synteny view it shows a horizontal representation with one chromosome at the top and the other one at the bottom. Coloured bands connect genes on one chromosome to their homologues on the other chromosome. It is hard to relate the bands to gene names. Those bands take up a