Title: Track-based Genome Browsers: Analysis, Challenges, and the Next Steps
Name: Jillian Slind
Contact: jslind@cs.ubc.ca

Domain:

A common task in bioinformatics and molecular biology analysis today is the analysis of a given sequence of interest with respect to that artifact's surroundings. In other words, the task is to find out where a target sequence is coded within a particular genome and determine the genomic context of that target sequence. There is a vast amount of useful information that can be found by performing these searches, such as upstream or downstream regulatory sequences that control how frequently your sequence becomes expressed, or nearby products that can be thought to be expressed simultaneously or in tandem with your target sequence.

There are a few ways to accomplish this contextual analysis of a particular sequence:

The first way is to extract the sequences of portions of a genome upstream and downstream of the target sequence, and query a database for artifacts that align with you particular sequences. However, not only is this a cumbersome process, but not as accessible to other users. Additionally, it is impossible to predict exactly how far up- or down-stream some interesting artifacts may turn up initially. If you only extract one kb (kilo-base, 1000 bases or 1000 letters) on either side of your sequence, you might miss some interesting artifacts that exist 2 kb up- or down-stream. If you have extracted too large of the sequence appearing before and after your sequence, the sheer number of unrelated sequences might occlude the ones that would be of interest to your investigation. Extracting many different sizes of flanking sequences is impractical due to the tedious nature of the task, not to mention the space and time required.

Thankfully, many organizations have worked to develop a more intuitive way to view your sequence in the context of a genome. Genome Browsers can show the context of your sequence within a genome graphically, allowing for direct analysis or exploration upstream and downstream. The genomes loaded into these browsers have annotations and markings within their sequences identifying other coding regions or other regions of potential interest within the genome. Thus, by graphically exploring upstream and downstream of a particular sequence, interesting regions can be identified quickly and efficiently.

A common form of genome browser for the purposes of sequence analysis is the track-based genome browser. These browsers center the genome being viewed on the sequence of interest, and show the flanking regions of the genome. One advantage of these track-based genome browsers is that the user can elect to view more of the genome on either side of the sequence by scrolling, or adjusting the level of zoom. Coloured lines or blocks encode the genome and sequence of interest,

and additional sequences are shown by a variety of methods dependent on the particular genome browser.

Not all track-based genome browsers, however, are created equally. From my experiences in dealing with them and teaching others how to use them, often the encoding methods can cause confusion if not done properly, and navigation can become cumbersome. Thus, there is often a learning curve involved before the user can take advantage of the tools made available through the browser itself. If the learning curve is too steep, a potential user could get frustrated and abandon the genome browser without getting the information they seek.

Task:

The goal of this project is to comparatively analyze two different genome browsers by trying to gather important information about a dataset with each, and taking to account not only the amount of detail gathered from the analyses, but the ease at which the information is extracted. The focus of this analysis will be on the visualization techniques incorporated and their effectiveness at making the analysis user-friendly. This will be completed in three major steps:

1) Determining the data set for analysis. The goal of this is to get a wide variety of data with different genomic contexts. For example, one gene could be highly regulated based on expression of a flanking gene. Another gene might not be regulated, but might be known to have similar expression behavior to a few other genes – it would be of interest to see whether or not these genes exist within the same region. If so, what regulates them? Another gene of interest might be one that is regularly modified in cancerous cells.

2) Based on the characteristics of the data, choosing two genome browsers for analyzing the data. One browser of current interest is the Galaxy Genome Browser, as the Galaxy platform of bioinformatics tools is quickly growing platform in not only capability but also popularity (it was popular at ISMB this year). Some other popular genome browsers are the ones featured by UCSC, and Ensembl.

3) Running the analyses in tandem, and figuring out what information can be extracted, the relative ease of extraction of information, and recording the visualization techniques used to make this extraction more accessible.

Data:

All data will be downloaded from NCBI's ftp site, once the appropriate genes and genomic regions have been selected.

Personal Expertise:

I have a B. Sc. Honours in Bioinformatics, and a B. Sc. Honours in Biochemistry. Throughout my educational career, I have worked with a variety of genome browsers, as well as instructed several other students on the use of these browsers. Thus, I have a lot of first-hand experience in the types of information to look for with these browsers, as well as potential frustrations that users of these browsers may have.

Analysis Plan:

As stated in the goal, this analysis will be performed in a series of steps.

1) Determine the data for analysis (as stated above). Since the human genome is the one most extensively researched, all of the genes for analysis will be for the human genome.
2) Determine the browsers for comparison. Hopefully, I can get access to Galaxy's Genome Browser as one, and then ideally a popular/mainstream browser such as the one from Ensembl or UCSC. Ideally, the second genome browser will be online, because not every user will have fully updated data at all times. Having up to date information is key for the field of bioinformatics. One browser that I probably won't use is the Ensembl browser, just because I have had so much practice using that one.
3) Format the data for analysis for the genome browsers required, and develop a set of questions to answer using these browsers, as well as requirements to check for.
4) Run the data on these browsers, and determine the visualization techniques used for each aspect of the browser (How does the user navigate? How are different types for regions encoded? How much detail is shown?). Answer the questions that were determined previously and note every step taken to answer these questions, as well as any presented challenges.
5) Compare and contrast these browsers and the techniques used, and give reasons for preferring one over the other. Compare these browsers to other browsers in the literature.
6) From the live analysis of genome browsers as well as analyzing other browsers in the literature, think of things that could be the next steps for future genome browsers, or updates for current genome browsers.

Schedule:

Since this proposal is late, here is the proposed schedule. Note the heaviness in the first week or two is purposeful – the long weekend will allow me to make up for lost time.

Wednesday, 9 November 2011	Submit Project Proposal, continue related work search
Friday, 11 November, 2011	Decide on data to analyze, start generating list of tasks

Monday, 14 November, 2011	Decide on genome browsers to use and acquire access to them, determine characteristics and prepare to present initial findings
Wednesday, 16 November, 2011	Project midway presentation
Friday, 18 November, 2011	Fully determine task set for data analysis, have data fully acquired, start analysis
Monday, 28 November, 2011	Finish analysis using the browsers, finish related work search, start summarizing results into presentation slides. (Make sure to be prepared for unrelated in class presentation!)
Friday, 2 December, 2011	Finish preliminary presentation slides, start write-up
Friday, 9 December, 2011	Use write-up draft to prepare for presentation, editing slides as need be
Monday, 12 December, 2011	Present project, start finishing the write up
Wednesday, 14 December, 2011	Hand in the write up

Previous Work:

(1) is a general review of three major genome browsers, namely Enselmbl, NCBI's MapViewer and UCSC's genome browser. This review goes over the basic functionality of each one and compares them based on what information can be extracted, but not a great emphasis on usability. (2) Is a very general review of a multitude of genome visualization techniques, with a section on genome browsers, giving a generalized overview of what was out there at the time, (March 2010), but not so much in terms of a comparative analysis. (3) Discusses challenges involved with using genome browsers to infer biological knowledge ((this is important for the analysis/testing aspect)). Galaxy's genome browser was introduced a short while ago at BioVis 2011 (4). Ensembl's genome browser is found here: http://tinyurl.com/epigenomebrowser-hlab. NCBI's MapViewer is found here: http://tinyurl.com/mapviewer-hlab. UCSC's genome browser is found here: http://genome.ucsc.edu/.

References:

1. Furey, TS. Comparison of human (and other) genome browsers. *Human Genomics*. Jan 2006, 2(4): 266-70.
2. Nielsen, C. et al. Visualizing genomes: techniques and challenges. *Nature Methods (Supplement)*. Mar 2010, 7(3).
3. Cline, W & Kent, J. Understanding genome browsing. *Nature Biotechnology*. Feb 2009, 27(2): 153-5.
4. Goecks, J. et al. The Galaxy Track Browser: Transforming the Genome Browser from Visualization Tool to Analysis Tool. *IEEE Symposium on Biological Data Visualization*. October 23-24, 2011. pp: 39-46.