Project Proposal - CPSC 533C

Andrew Carbonetto

Due: Nov 4th, 2004

1 Description of domain, task and dataset targeting

This project is targeted to a project that I have previously worked on, in the bioinformatics field. Currently, there has been little success towards finding an optimal solution. Thus, a lot of the current work is only approximate.

Let me first define the dataset: Input:

- A sequence of ordered nucleotides, $N = \{A, C, G, T\}^*$.
- A set of motifs with attributes: start_location, end_location, similarity_score, motif_name, family_name, species_name)

The set of motifs are a set of potential Transcription Factor Binding Sites (TFBSs) that are predicted based on the inputed sequence of nucleotides. How they are predicted is a complicated question that is currently being researched. Although the number and combination of predicted TFBSs differ from algorithm to algorithm, finding *significant* groupings (*complexes*) of TFBS remains an even more difficult question. Biologist have a few ideas of the properties of a significant TFBS.

One reason that these sets tend to be so difficult to understand is because of the density of predicted TFBSs. Some programs will predict several hundred points upon a sequence of small size (1000 nucleotides). Recently, interactions between TFBSs found to be 40'000 nucleotides apart were found, leading biologists to believe that many large sequences should be analyzed for potential TFBSs.

The other reason that TFBSs are difficult to predict, is due to the fact that TFBSs are not consistent. Instead, potential sites are predicted based on a frequency matrix (the matrix empirically tallies the count of each binding site). Thus, predictors tend to rate the similarity of the binding site to a consensus sequence or frequency matrix. This similarity score is very important, can be used as a filter.

A visualization of large portions of the sequence, with the corresponding TFBS locations, would be helpful in this non-trivial pattern recognition exercise.

Some sites with website TFBSs prediction access:

- **matInspector**: http://www.genomatix.de/
- **TSSG/TSSW** http://www.softberry.com/berry.phtml
  no graphical output
- **consite** http://mordor.cgb.ki.se/cgi-bin/CONSITE/consite/
  W. Wasserman at CMMT/UBC contributed to this site
- **rVista** http://genome.lbl.gov/vista/index.shtml

2 Personal Expertise

I have performed several TFBS searches, many with little to no success. The analysis tends to be very time consuming, because of the large predicted set of TFBSs. One would prefer to not filter the set, because of the possible loss of significant points. A quick method to view filtered or unfiltered data could be helpful.

Although I've experience with the field, I've not taken a university-level graphics course, and am unfamiliar with the current graphics software (see Figure 1 for an example of a search).

3 Your proposed infovis solution

I propose a customized Parallel Coordinates solution to the problem. This solution will allow multiple sequences to be place in parallel so that patterns of TFBSs (called a promoter complex) that are common between the sequences can be easily visualized.

There are several filtering techniques that are shown to be useful, including filtering by conservation between sequences (roughly: number of nucleotides common to both sequences), filtering for most similar TFBS motifs (each motif has a frequency matrix for scoring) and filtering by species.

Interactive filters would be useful to implement.

Since clutter is an issue, it would be nice to implement a mouse-over pop up of the information of each TFBS.

Since the sequences are long, navigation across these sequences is required. A limited zooming function might be necessary depending on the number of viewable motifs on the screen at one time.

Links between TFBS common on each parallel sequence should be represented as a line across the coordinates. The color of the line can be determined by either the TFBS name, or the TFBS species. The brightness of the line can represent the similarity of the motif (this can be represented in log function of brightness, and respective to the assigned filter).

By clicking on a link, both sequences can be navigated so that the line is at right angles to both sequences. Only links with TFBSs on the screen would be viewed, to reduce clutter.

As optional prospects to be included: Aggregation of patterns of motifs. Aggregation of conservation represented on the graph. Zooming up to the nucleotide level, and out to the whole viewable sequence level. The same sequence viewable as parallel coordinates, so that pattern can be found at multiple locations on a given sequence.

4 A scenario of use

We are showing several ideas on how to view the potential TFBSs, and then apply filters and focus techniques to find patterns.

In figure 2, we see how the points relate to each other. Unfortunately, with only a few viewable TFBSs, the number of edges between all similar TFBS begins to clutter. So we need to apply some techniques to isolate the more probable TFBSs.

In figure 3, we can see some of the techniques that were listed in the "proposed infovis solution" section above.

In figure 4,we can see the results of a mouse over and click to a specific predicted TFBS.

One idea that cannot be shown very clearly, is the animation. We'd like to implement a smooth transition between two sections of a sequence, and a slow-start slow-stop technique might be the best solution (with fast zoom-in and fast zoom-out if necessary).

5 Proposed implementation approach

With my limited experience of graphics and graphical languages, I believe getting to learn the basics of the language to be my greatest hurdle. Here are the potential languages that I will use:

- **ILOG** Although excellent for parallel coordinates, this tool might provide too many options (since my proposal is a simplified version). If I can limit clutter to a minimum, this might be a good framework.
- **prefuse** This might require a bit of management and time, but this toolkit might be ideal for my proposal.

6 milestones

This is a potential calendar of events for the project (items in italics are important due dates for other classes):

Nov 12 Read up on some recent topic related papers to make sure there are no new techniques available in the area that I should know of.

Nov 12 A sample usage of each potential language

Nov 14 *Midterm in cs540*

Nov 19 Have a better idea with which language to use (possibly narrowed down topic to one language)

Dec 3 *project reports due cs304*

Dec 10 A base case for the project (as in figure 1), with scrolling included of any sort

Dec 13 *project presentations for cs540*

Dec 14 Mouse-over abilities that tie up with matrix view

Dec 17 Implementation of most filters, scrolling and mouse-over. Project source should be finished, or near completion (debugging needed?).

Dec 19 Project Due

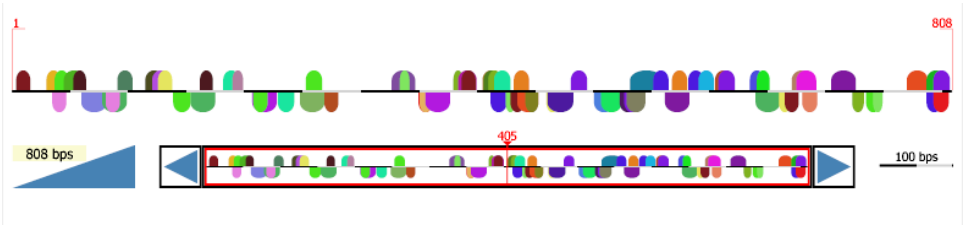


Figure 1: A graphic of the potential binding sites provided by matInspector of the Genomatix suite (http://www.genomatix.de/). Each semi-circle represents a potential binding site on the 808bp given sequence.

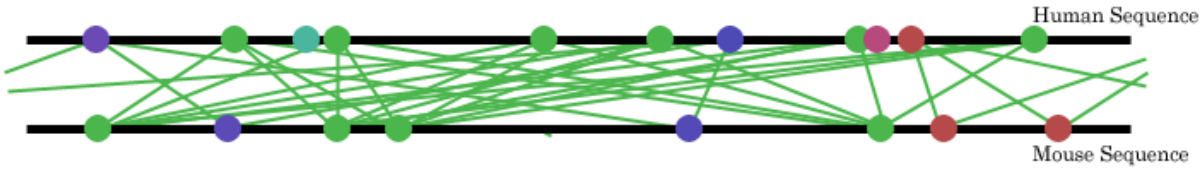


Figure 2: An scenario of usage. No filtering has been applied. This shows two sequences side by side, with links between commonly named TFBSs. The sliders on the top show the relative position of the sequence in focus.

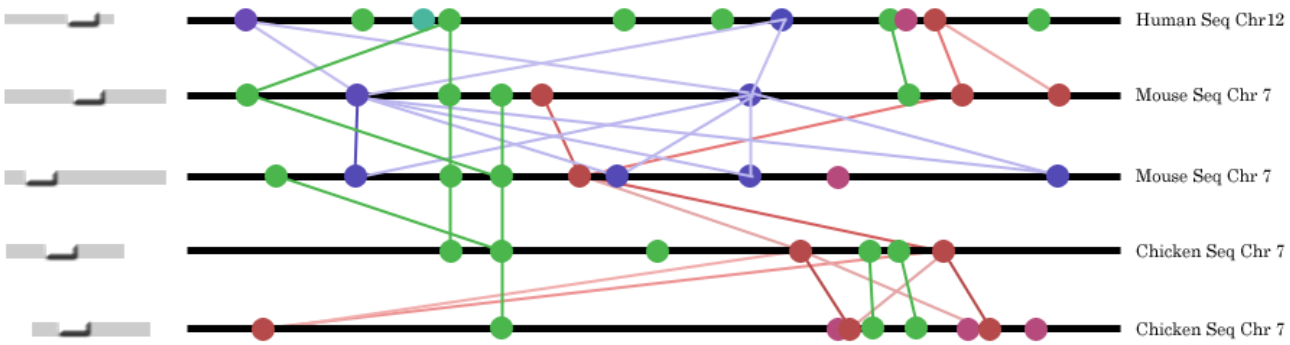
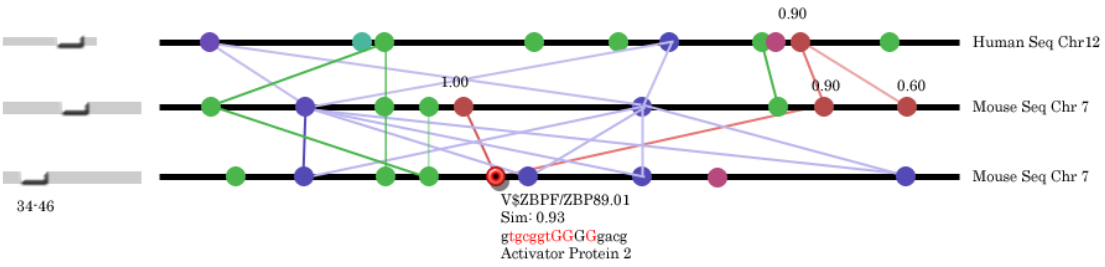


Figure 3: An example of using several sequences in parallel. Several points on both the Mouse and Chicken sequence are being viewed at the same time, for comparison. The sequences have been centered around the green TFBSs on the middle-left hand side. Several filters have been applied, so that fewer edges are seen (to reduce clutter). The similarity between two potential TFBSs is represented by the edge brightness.

A



B

Inspecting sequence m20543 (1 - 808):

	Family/matrix	Further Information	Opt.	Position from - to	Str.	Core sim.	Matrix sim.	Sequence
	V$AP2F/AP2.01	Activator protein 2	0.89	4 - 16	(+)	0.857	0.931	cgCCCTccggccg
	V$EGRF/EGR1.01	Egr-1/Krox-24/NGFI-A immediate-early gene product	0.79	30 - 44	(+)	0.797	0.809	gtgcggtGGGGgaag
	V$ZBPF/ZBP89.01	Zinc finger transcription factor ZBP-89	0.93	34 - 46	(-)	1.000	0.930	ccttCCCCcacc
	V$CREB/ATF6.02	Activating transcription factor 6, member of b-zip family, induced by ER stress	0.85	43 - 63	(+)	1.000	0.885	aggggtcGACGtggctcagct

Figure 4: (A) Selection of a point brings up appropriate information (text) about that point, and all related points. (B) It should also bring up a list of potential binding sites and highlight the appropriate selection.