Visual Analytics for Large-Scale Bioinformatic Data Sets

Samuel Smith, James Hogan, Markus Rittenbruch, Daniel Johnson, Margot Brereton
School of Electrical Engineering and Computer Science, Queensland University of Technology
2 George Street, Brisbane, Queensland, Australia
{s165.smith, j.hogan, m.rittenbruch, dm.johnson, m.brereton}@qut.edu.au

ABSTRACT

Rapid advances in sequencing technologies (Next Generation Sequencing or NGS) have led to a vast increase in the quantity of bioinformatics data available, with this increasing scale presenting enormous challenges to researchers seeking to identify complex interactions. This paper is concerned with the domain of transcriptional regulation, and the use of visualisation to identify relationships between specific regulatory proteins (the transcription factors or TFs) and their associated target genes (TGs). We present preliminary work from an ongoing study which aims to determine the effectiveness of different visual representations and large scale displays in supporting discovery. Following an iterative process of implementation and evaluation, representations were tested by potential users in the bioinformatics domain to determine their efficacy, and to understand better the range of ad hoc practices among bioinformatics literate users. Results from two rounds of small scale user studies are considered with initial findings suggesting that bioinformaticians require richly detailed views of TF data, features to compare TF layouts between organisms quickly, and ways to keep track of interesting data points.

Author Keywords
Bioinformatics, visualisation, visual encodings, human computer interaction

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Recent improvements in genome sequencing technology have led to an exponential growth in the availability of genomic data. Data visualisation and visual analytics approaches are becoming increasingly prevalent in this context as they allow biologists, bioinformaticians and clinical researchers to gain insights into biological mechanisms that cannot be easily derived from more common text-based representations of data. Transcriptional regulation - the process through which regulatory proteins called transcription factors control the expression of a set of associated target genes - is now far better documented across species through initiatives such as the RegPrecise database (Novichkov et al., 2013). However there are few visualisation approaches that specifically address this area. Large scale displays are also of interest, due to the possibility of displaying a huge amount of information at the same time without scrolling (Andrews, Endert, Yost, & North, 2011, pp. 344-345).

Our research is concerned with enabling biologists and bioinformaticians to gain scientific insight from these data sets. The study is based on a purpose specific visualisation system called Regulon Explorer (Chua, Buckingham, Hogan, & Novichkov, 2015) that uses radial information displays to show the links between the TF and associated TGs, and to allow comparison across species. In order to capture the needs of bioinformaticians and other researchers, we followed a user-centred design approach to both evaluate and guide further development of the system.

BACKGROUND

There is currently little research into suitable visualisation and exploratory approaches that are most useful for examining regulatory networks. As a result the aim of this project was to use the Regulon Explorer application as a base to test which visual representations and interactive features would be most helpful to allow bioinformaticians and similar researchers to gain new insights into the underlying data. We particularly explored whether adding functionality to separate and group regulons would assist users in discovering patterns in the data. A regulon is a collection of genes or operons under regulation by the same protein.

Other bioinformatics systems have successfully used radial display visualisations. Examples of these visualisations are CGView (Stothard & Wishart, 2005) and DNAPlotter (Carver, Thomson, Bleasby, Berriman, & Parkhill, 2009), both used for exploring the makeup of genomes, and the BLAST Ring Image Generator (BRIG), which uses CGView to show entire genomes side-by-side for comparison purposes (Alikhan, Petty, Ben Zakour, & Beatson, 2011) – these visualisations, however, do not share the specific context of Regulon Explorer. More traditional radial displays such as pie charts were also explored to determine the strengths and weaknesses of this display method, such as the issues with comparing values that are similar in magnitude (Draper, Livnat, & Riesenfeld, 2009). These radial displays can allow the viewing of multiple one-to-many relationships at the same time, which is a requirement for viewing this type of data.
Preliminary Work
In regards to the design of the Regulon Explorer interface, there were a range of visual elements considered, including colour and layout. Colour has a variety of different possible uses in a visualisation and is a highly effective encoding (Mackinlay, 1986, pp. 124-125), but different ranges of colours are suitable for specific types of data (Silva, Sousa Santos, & Madeira, 2011). In Regulon Explorer, colour was chosen to represent a nominal data type, and so a palette of contrasting colours was selected. A challenge with a nominal colour palette that has been considered throughout the design is the need to limit it to a small number of categories, so that the amount of different colours does not overwhelm the user (Silva, et al., 2011, pp. 323-324).

While large-scale displays have several advantages, such as not requiring users to keep objects in memory (Ball & North, 2007, p. 381) and allowing a greater number of users to collaborate (Andrews, et al., 2011, pp. 348-349), there were a number of important issues to consider. The shape of the visual elements is also an issue that is present when considering the scalability of Regulon Explorer. In a large scale display, objects that require the user to view them at an extreme angle or distance may have their apparent shape distorted (Andrews, et al., 2011, p. 342), which will be taken into account as our visualisation is scaled up. As size does not implicitly affect the type of shapes that can be used in the visualisation, any shapes that are selected need to work in both small and large scale and avoid clutter. The decision to experiment with a pie-chart like wedge layout in later work was driven by this issue. As the visualisation already included multiple views, and large scale displays allow more views to be displayed, there are additional advantages: for example, reducing the cognitive load for comparing two views by showing them on screen at the same time, and being able to show relationships between views (Baldonado, Woodruff, & Kuchinsky, 2000, pp. 113-116).

Another more subtle issue in visualisation that has been considered is the effects of “non-target” elements. Exploring the literature in preattentive processing in human sight initially showed that the similarities between non-target objects - those elements a user is not currently interested in - in a display can affect how readily the target objects “pop out” to the user (Duncan & Humphreys, 1989, p. 442). This is especially true when non-targets vary greatly in the characteristic that is unique to the target - for example, when looking for a target with a specific colour (Duncan & Humphreys, 1989, p. 456). As the visualisation is scaled up, the issue of non-targets will become an important issue to examine.

Regulon Explorer
Regulon Explorer (or TRNDiff) (Chua, et al., 2015) is the visualisation system that was designed and evaluated as part of this project. The aim of the visualisation was to analyse networks of genes known as “regulons”, which show how a specific gene or protein regulates the functions of a collection of genes. These regulated genes are “homologous”, meaning they are common across multiple species. Using the Regulon Explorer users were able to examine the composition of each regulon, use the visualisation to compare between different regulons and determine whether their similarities or differences provide valuable insights into possible evolutionary relationships or characteristics, such as their contribution to organism virulence.

The Regulon Explorer was based on a previous implementation of TRNDiff developed as a new way to explore transcriptional regulation, as part of Chua’s investigation into methods for increasing the accuracy of transcription factor binding site prediction (Chua, 2012). Regulon Explorer uses a radial-type display - colloquially known as a “wagon-wheel” for its appearance - to display the regulatory layout of different genomes (see Figure 1). The centre node (A) represents the regulator gene, while the outer nodes represent target genes (B), with genes of the same type - homologs - positioned in the same spot across different networks. While there can be degrees or strengths of regulation, at present, spokes only display the presence or absence of regulation. If a spoke is absent, it indicates that regulation is not present (C).

![Figure 1: Regulon Explorer’s “wagon-wheel” display. Components are A) the regulator gene/protein, B) a regulated target gene, and C) an unregulated target gene. Gaps indicate a gene is completely absent.](image)

There is no limit to how many genes can be displayed at once, but crowding does occur with some larger datasets. Regulated genes can also be selected for identification purposes, and users can also filter to show genes with certain gene functions. Colour is used to indicate the gene function when the filter is inactive, with a fixed palette of seven colours to represent this nominal data. The networks are labelled by the genome name, with further information present in tooltips that appear when the user hovers over a gene node. At the current display size, 12 of these networks can fit on a laptop-sized display at once.

Information is retrieved from the RegPrecise database - a collection of recent data on regulatory proteins and the genes they regulate across multiple species (Novichkov, et al., 2013) - and includes information such as gene/regulator name, function and locus tag. At present, IDs used in the application are taken from RegPrecise and do not correspond to IDs in other genomics databases such as BLAST and GenBank - linking our system to other sources is a future implementation goal.
The majority of datasets within the RegPrecise database do not contain more than 30 individual genes. The wagon wheel display is well suited to display networks of that size. However, a number of outlier regulons can have up to 150 genes, making this type of visualisation unsuitable. For the purpose of this project we focused on the most common case, rather than aiming to offer different visualisation options for different regulon sizes (see Figure 2).

Several features were added to assist researchers to identify patterns and make new discoveries about datasets. One of these was a feature to allow users to rearrange and “group” genes. Each regulon network could be dragged around the display area and dropped in a new location, and when the grouping function was selected, networks could be dropped into separate areas of the display. Users were also able to select a “side-by-side” view of two networks and compare them using the logical operations AND, OR and XOR.

METHODS

We conducted two preliminary evaluation sessions to examine the existing functionality and visual representations in Regulon Explorer. The simple, more obvious features were tested first, followed by the more complicated features. Zooming functionality, allowing side by side comparisons and highlighting selected genes across all networks, based on user feedback from the first session were added before the second evaluation session.

The goal was to determine the features that bioinformatics researchers would find most useful to help them to gain insights into the represented data.

The first evaluation session, conducted in November 2014, involved five participants (4 male, 1 female) and the second, occurring in May 2015, involved three (1 male, 2 female) participants. All of these participants were chosen as they had some background in biology, though not all of them were bioinformatics experts.

Each of the two sessions consisted of a two-phase evaluation process, a trial consisting of the use of the software by a single participant (20 minutes on average) and a follow-up contextual interview (10 minutes on average). During the first step, participants explored an artificial dataset - based on real data from the Yersinia pestis species - and were asked to use the radial visualisation to compare different Regulog networks and determine whether they could identify relationships and regularities within the data. The available data was prepared by the research team to ensure that there was a distinct pattern of infectious and safe regulons that participants could detect and classify using the grouping functions. We decided not to advise participants about the existence of the pattern, but instead encouraged them to find patterns they felt were “interesting”. This was to ascertain whether the current implementation enabled the discovery of new patterns and relationships.

During both sessions participants were encouraged to vocalise their thoughts and feelings about the visualisation during the exploration, using a think aloud approach. Afterwards, a contextual interview allowed participants to elaborate and reflect on their use of the system. The evaluations were intended to run around 45 minutes in the first set and 30 minutes in the second set, though actual evaluation time varied significantly between participants. The main goal of these evaluation sessions was to elicit feedback from potential users to understand how the system and the visual representations were used and to determine whether our system could be a viable tool to assist users in this domain with their research. We were further looking for specific feedback with regards to the usability and applicability of the system that would allow us to improve future versions of the software.

All evaluation sessions were audio-recorded with the permission of the participants. We used an open coding approach to analyse the resulting data and draw out common issues and themes.

RESULTS AND DISCUSSION

The overall feedback from both evaluation sessions was mixed, but generally positive. Several important issues with the current software as well as suggestions for improvement were identified. Lengths of evaluation times differed, with the shortest total times being slightly less than 20 minutes, and the longest being over 50 minutes.

Several main usability issues were identified by the participants. They had some difficulty in examining the large number of dense networks due to the small screen size and the small fixed size of the networks, and therefore suggested the ability to zoom in and out of the
networks to more easily examine them. It was also difficult for them to notice the differences in regulation pattern between two networks at a glance, meaning that a side-by-side logical comparison was another suggestion.

Participants were generally pleased with the amount of detail available, but said that there should have been a way to show the type of regulation, rather than to just show whether it was present or not. Including this feature would rely on whether this data could be incorporated into the visualisation from the RegPrecise database or another source.

Features that were added between the two evaluations included a new display that allowed two networks to be compared side by side, and to use a logical comparison such as AND, OR and XOR to highlight the differences between them. A function for zooming in and out of the networks was also added, as well as a feature for highlighting all homologous genes when one gene node was selected. A feature for scaling the distance between regulator node and gene node based on their actual physical distance in the genome was also trialled.

In general, the use of colour was an issue that emerged in both the first and second evaluation session - Different colours were used to encode different gene functions in the graph, with the gene function names present in the hover tooltip. Participants generally felt that the use of colours was appropriate, and stood out - "At a glance, the colour stands out most" was stated by one - however, they thought that a legend showing how colours were mapped to gene functions would allow them to identify genes more quickly. The grouping function was the one feature that participants had trouble discovering, as the study setup did not require them to extensively drag and drop networks around during the sessions. Multiple participants also suggested that there should be a way to annotate, label or otherwise keep track of interesting data points they had found.

One of the most important issues that we encountered was that few of the participants used the grouping function to try to identify a pattern in the data. It is possible that the density of the data (in both size and detail), in conjunction with the fact that the ability to drag and drop data within the system was not very visible, made the participants' reluctant or unable to explore this feature in depth – one participant described the process of looking for the presence of genes as “laborious”. However, multiple participants identified that they could use the visualisation to “see what [the networks] share and what they don't.”

Based on our observations and participant feedback, there is an indication that adding features that allow users to keep track of and/or annotate important data points will allow them to more easily identify patterns and then use the grouping function to separate the dataset into groups based on that. Finally, an automated grouping function which uses Hamming and Euclidean distances to perform K-means grouping will help users see which networks have similar characteristics, and allow them to fine-tune the grouping depending on their own observations.

The suggestion of a legend is also important to consider, as colour is a very effective visual encoding and must be used properly - for example, using similar colours in the normal display and the logical side-by-side comparison and not indicating their meaning caused confusion. Being able to better identify genes with similar function across networks should increase the probability of users making insights into the data they analyse with the tool. It is also possible to have several data characteristics that can be encoded with colour, and allow users to select which one they wish to use at any one time.

Finally, it is important to note the small sample size in these evaluations. The small number of participants was not intentional, but was due to the difficulty in recruitment in the specific domains we were targeting. Future evaluations with more participants would be able to determine whether the findings in these preliminary evaluations are accurate.

CONCLUSIONS AND FUTURE PLANS

While further extensive user testing is required, the preliminary feedback from the first two evaluation sessions has been valuable in order to capture some of the requirements that bioinformaticians and other related researchers might require in a bioinformatics visualisation relating to regulatory relationships. These include being able to easily understand the meaning of data represented by colours, an accessible way to view a network in detail, to keep track of interesting data points, and methods to directly compare networks to each other.

We are currently entering the second stage of our project. Because of the increasing size of bioinformatics datasets, this stage will explore how large-scale interactive displays can be used to allow bioinformaticians to explore large and complex datasets at scale. This requires reorganising the interface for better support for multi-touch interaction and high pixel-density wall displays, such as QUT’s Cube facility (Rittenbruch et al., 2013). We expected that large-scale displays will allow for the displays of more complex networks and relationships as well as foster collaboration between users, potentially leading to greater insight into datasets (Andrews, et al., 2011, pp. 348-349). The main focus will be to determine if the visual encodings that are viable in small scale are also viable in large scale, allowing a truly scalable bioinformatics visualisation. Because of the touch-based nature of the Cube and other large-scale displays, this will also require Regulon Explorer to be modified to allow touch controls.

Variations to the radial display have also been considered for testing, such as using a pie chart-like wedge representation to aid users in interpreting data at small zoom levels or from a long distance away. A rough implementation of automatic grouping using a mix of Euclidean and Hamming distance has also been added.

It is our intention that our preliminary exploration of the visual representation and exploration in the bioinformatics domain will be useful to researchers in related areas and a starting point for further research.
REFERENCES


