DYNAMIC QUERYING FOR PATTERN IDENTIFICATION IN MICROARRAY AND GENOMIC DATA

Harry Hochheiser[1], Eric H. Baehrecke[2] Stephen M. Mount[3], Ben Shneiderman[1]

[1]Unversity of Maryland, Department of Computer Science; [2]University of Maryland Biotechnology Research, Center for Biosystems Research; [3]University of Maryland, Department of Cell Biology and Molecular Genetics

ABSTRACT

Data sets involving linear ordered sequences are a recurring theme in bioinformatics. Dynamic query tools that support exploration of these data sets can be useful for identifying patterns of interest. This paper describes the use of one such tool – TimeSearcher - to interactively explore linear sequence data sets taken from two bioinformatics problems. Microarray time course data sets involve expression levels for large numbers of genes over multiple time points. TimeSearcher can be used to interactively search these data sets for genes with expression profiles of interest. The occurrence frequencies of short sequences of DNA in aligned exons can be used to identify sequences that play a role in the pre-mRNA splicing. TimeSearcher can be used to search these data sets for candidate splicing signals.

1. INTRODUCTION

Data sets involving linear ordered sequences of measurements are a recurring theme in bioinformatics work. Microarray time courses include gene expression levels for thousands of genes over multiple time points, providing biologists with a history of relative changes under different experimental conditions. Similarly, frequency counts for oligonucleotides in aligned sequences can help to identify signals for transcription and RNA processing.

Although these data sets involve fundamentally different questions and methods, they are both amenable to analysis via interactive querying for patterns involving differences in measured levels. For microarray results, this might involve finding genes that have low expression levels at one time point and higher levels at the next. For sequence data, the analogous query might involve identifying short sequences 5 nucleotides long (pentamers) that occur more frequently than expected in particular regions.

TimeSearcher (http://www.cs.umd.edu/hcil/timesearcher) [8] is a dynamic query tool originally designed for identification of time series data. This paper describes the use of TimeSearcher for identification of splicing signals in aligned sequence data, and patterns in microarray time course data.

2. TIMEBOXES AND TIMESEARCHER

Timeboxes are rectangular query regions drawn directly on a two-dimensional graph. The extent of the Timebox on the time (x) axis specifies the time period of interest, while the extent on the value (y) axis specifies a constraint on the range of values of interest in the given time period. More specifically, if an item in a data set is to satisfy a timebox that goes between (x_1,y_1) and (x_2,y_2), for every point in the time range $x_1 \le x \le x_2$, the value of that item must be in the range $y_1 \le y \le y_2$ (assuming $y_2 \ge y_1$ and $x_2 \ge x_1$). Multiple timeboxes can be drawn to specify conjunctive queries. Items in a data set must match all of the constraints implied by the active timeboxes in order to be included in the result set.

To create a timebox, the user clicks on the desired starting point of the timebox and drags the pointer to the desired location of the opposite corner. As this is identical to the mechanism used for creating rectangles in widely used drawing programs, this operation should be familiar to most users. Once the timebox is created, it may be dragged to a new location or resized via appropriate resize handles on the corners, using similarly familiar interactions. In all cases, the query is reprocessed with each mouse event. As the mouse is moved, the current position of the timebox is stored, and the result display is updated.

Construction of timeboxes is aided by the display of the graphs for all of the items in the data set directly on the query area. This "graph overview" display provides additional insight into the density, distributions, and patterns of change found among items in the data set (Figure 1).

3. MICROARRAY DATA

Numerous published reports of microarray data have used the examination of changes in gene expression levels over time to examine the effects of various stimuli on genetic expression. As these data sets typically involve measurements for thousands of genes taken over numerous (5-30) time points, interpretation is often a challenge. Currently, analyses of these data sets are generally conducted via some sort of mathematical grouping of genes with similar expression profiles. Clustering techniques that have been used include hierarchical clustering [5,14], self-organizing maps [17,19], and singular value decomposition [9]. The use of TimeSearcher to interactively query and explore these data sets may be a useful complement to such techniques.

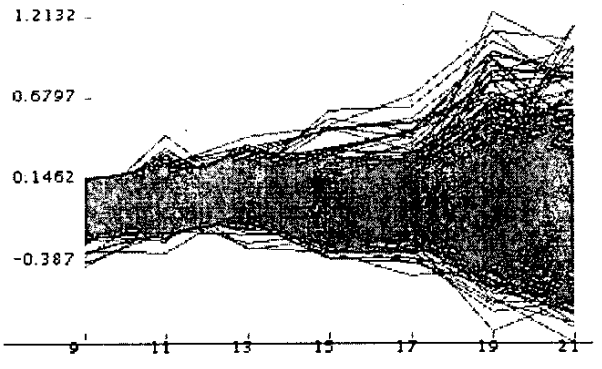

Figure 1: A "graph overview" display of 1051 profiles

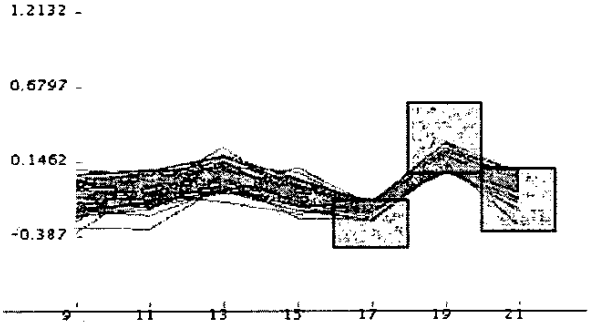

Figure 2: Genes with expression levels with local maxima at 19 hours

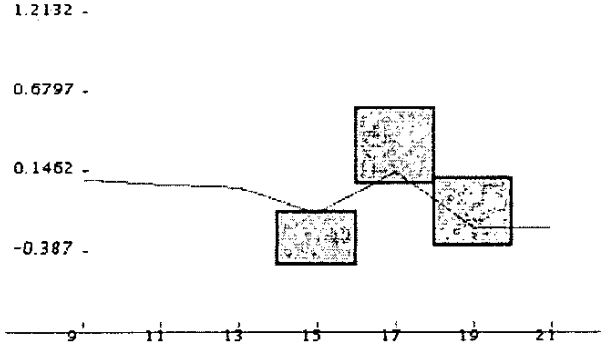

Figure 3: A single gene with a local maximum at 17 hours.

In order to study the shift between anaerobic to aerobic metabolism, DeRisi, Iyer, and Brown examined gene expression changes in yeast (*Saccharomyces cerevisiae*) cells at several points in time after their placement in fresh medium [4]. Microarray measurements were made every 2 hours between 9 and 21 hours after initial placement, for a total of 7 time points. Figure 1 shows a "graph overview" of a subset of the data set, containing 1051 profiles overlaid on a single pair of axes. Starting from this overview, the user might draw a series of timeboxes to identify some genes with expression levels that had local peaks at 19 hours (Figure 2). This provides a subset of 57 genes that appear to have strongly similar expression profiles.

To identify potential regulatory genes, this query can be shifted one time period earlier, thus finding genes that had similar peaks at 17 hours. To do this, the user lasso-selects the three query boxes and drags them to the left. The resulting query identifies a single profile with a similar local peak at 17 hours (Figure 3). As this single gene precedes the expression of the genes identified in Figure 2, it might be considered for examination as a potential downstream regulator of those genes with local maxima at 19 hours.

4. NUCLEOTIDE SEQUENCE DATA

Molecular biologists interested in understanding the process of converting genes into proteins must examine and understand the structure of nucleotide sequences. Often, this work involves analyzing the frequencies of subsequence/oligonucleotide/words at different positions within aligned DNA sequences. A variety of computational and statistical tools have been proposed to help with the challenge of analyzing the large volume of sequence information that is available [6,7,12]. As with the microarray data sets, interactive exploration complements these tools.

Specifically, we have been using TimeSearcher to identify consensus branch site splicing signals in the plant *Arabidopsis thaliana*. These are secondary signals in the RNA transcripts of genes that help to determine which sequences (introns) are removed from RNA. The segments that remain are known as exons.

The data set being used for this purpose was generated from the genomic sequences surrounding 8550 internal exons that were internally truncated and aligned with respect to their boundaries [13]. It consists of the normalized frequencies of each of the 1024 possible pentamers at each of 192 possible positions.

Figure 4 shows a "data envelope" overview of the whole data set. Formed by plotting a contour defined by the minimum and maximum values of any item in the data set at a given time point, this display can provide useful feedback when the graph

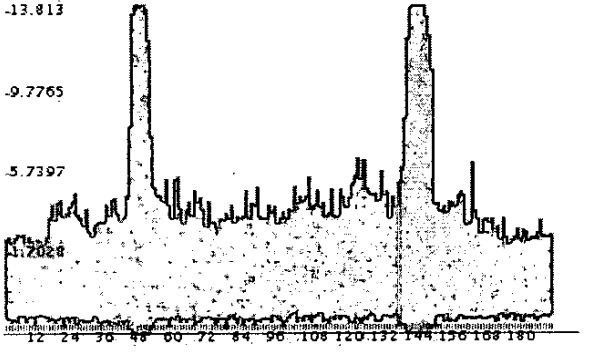

Figure 4: "Data Envelope" overview of pentamer frequency distributions in *Arabidopsis thaliana*

envelope (Figure 1) becomes too cluttered. Two peaks, indicating the boundaries between the exon in the middle and the introns on the ends, are immediately apparent. These peaks represent well-known conservation of sequences at splice sites.

To identify candidate splicing signals, we create a query using two timeboxes. One component of the query will identify those pentamers that are frequently found 23-27 nucleotides upstream of intron-exon boundaries. The second identifies pentamers that are infrequently found elsewhere within introns (Figure 5). Taken together, these criteria identify candidate branch site consensus sequences (e.g. CTAAT, CTGAT) that correspond closely to known examples [15]. In addition to identifying known consensus motifs, TimeSearcher was useful for identifying their location.

5. RELATED WORK

A variety of methods have been proposed for analyzing microarray data sets. Mathematical clustering of gene expression profiles, together with clustering mosaic plots, has been used to examine time series and other microarray results [4,5,9,11,17,19]. As these results generally provide output in static forms, interactive exploration is generally not supported. Similarly, much of the work to date in analysis of oligomer count data has focused heavily on computational and statistical approaches [6,7,12]. These computational approaches can be very helpful for understanding these data sets, but they suffer from problems such as lack of interactivity and possible sensitivity to parameters. TimeSearcher and other tools that let users work more directly from the data may help users "see" their data better. Ideally, interactive approaches such as TimeSearcher would be integrated with these computational approaches.

Recently, tools that apply principles of information visualization to microarray data have been developed. The hierarchical clustering explorer (HCE) provides support for dynamic querying of dendrograms that result from hierarchical clustering algorithms [14]. VxInsight has been used to visualize clustered gene expression profiles in a 3d-projected "mountain" [11].

Interactive tools for querying time series data might be used to find patterns in microarray and sequence data. QuerySketch is an innovative query-by-example tool that uses an easily drawn sketch of a time-series profile to retrieve similar profiles, with similarity defined by Euclidean distance [18]. This approach is simple and intuitive, but accurate sketches may be difficult to draw, and query constraints, including the similarity threshold, are not adjustable. Spotfire's Array Explorer 3 [16] supports graphically editable queries of temporal patterns, but the result set is generated by complex metrics in a multidimensional space. This potent approach produces useful results, but users may wish to constrain result sets more precisely. Spotfire also includes integrated support for a variety of clustering algorithms.

Algorithmic tools for working with microarray time series data sets have addressed difficulties caused by missing data and differences in experimental time scales that might require time warping [1,4]. The combination of these tools with TimeSearcher's interactive visualization might be an interesting area for future work.

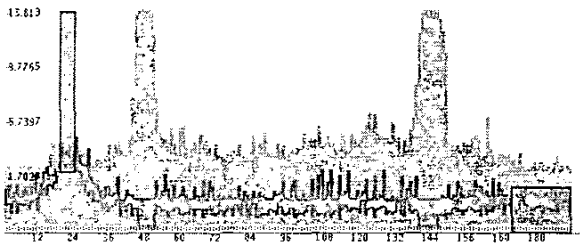

Figure 5: Timebox query aimed at finding pentamers with higher frequencies at a specific region within introns (the branch site) and lower frequencies elsewhere within introns.

6. DISCUSSION AND CONCLUSIONS

The identification of interesting transitions is a key element of analysis of the microarray and genomic data sets described above. For example, identification of genes with characteristic increases or decreases in expression levels is necessary for finding genes that respond to certain stimuli. Similarly, the candidate regulatory splice sequences were those that had high frequencies at one position and low frequencies at another, requiring a conjunctive query.

Timeboxes and TimeSearcher are particularly well suited to support these queries. As timeboxes are drawn directly on a graph space that is also used for plotting data, the queries are easily interpreted at a glance. Complex queries containing multiple timeboxes provide visual feedback that illustrates the pattern defined by the query. The graph envelopes drawn directly on the two-dimensional query space provide additional feedback that can aid the process of creating queries and interpreting result sets.

The power of the timebox model lies in its flexibility. For queries involving identical constraints over w adjacent attributes, a single timebox of width w can be used to express all of the constraints. This represents a substantial improvement over single-attribute query widgets, which would have required manipulation of w individual widgets to specify the same number of parameters. When desired query constraints vary from one attribute to the next, multiple boxes can be combined to specify a complex, conjunctive query (Figures 2, 3, and 5).

As an interactive, dynamic query tool, timeboxes can assist analyses of microarray and oligomer count data sets by providing rapid feedback that links the results of queries to the query criteria. Together with TimeSearcher's graph envelope and data envelope overviews that provide high-level summaries, these queries can help biologists understand data sets. Timebox queries that describe patterns of interests may be interesting results in themselves, possibly providing parameters that might be used with algorithmic approaches.

The design and functionality of TimeSearcher has been influenced by our needs in examining data sets similar to those discussed above. In addition to providing preliminary validation of the timebox model, this work has led to numerous suggestions for extensions to the query model.

For example, Variable Time Timeboxes (VTTs) can be used to specify queries requiring that values remain in a given range for a given amount of time, occurring within some larger range.

These queries might be used to find genes that have peak expression levels for three consecutive measurements anywhere in a window containing 5 time points. VTTs have proven useful in the construction of queries that separate two classes of profiles in a larger data set [10]. Future work will involve incorporation of VTTs and other extensions into TimeSearcher.

ACKNOWLEDEGMENTS

The first author is supported by a fellowship from America Online. Thanks to Steven Salzberg from The Institute for Genomics Research for providing the nucleotide sequence data.

REFERENCES

[1] J. Aach and G.M. Church Aligning Gene Expression Time Series with Time Warping Algorithms. *Bioinformatics* 17(6): 495-508.

[2] Z. Bar-Joseph, G. Gerber, D.K. Gifford, and T.S. Jaakkola. A new approach to analyzing gene expression time series data. In *The Sizth Annual International Conference on Research in Computational Molecular Biology,* 2002.

[3] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf Computational Geometry: Algorithms and Applications, Springer-Verlag: Berlin, 2000.

[4] J. DeRisi, V. Iyer, and P. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278: 680–686, 24 October 1997.

[5] M.B .Eisen, P.T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci USA* 9514863-14868.December, 1998.

[6] W.G. Fairbrother, R.F. Yeh, P.A. Sharp, and C..B. Burge Predictive identification of exonic splicing enhancers in human genes. *Science 297*, 1007-1013, 9 August 2002.

[7] J. van Helden, J., B. Andre, B. and J. Collado-Vides Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology 281*(5), 827-842, September 4, 1998.

[8]. H.S. Hochheiser and B. Shneiderman Visual specification of queries for finding patterns in time series data. In K.P. Jante and A. Shinohara, editors, *Proceedings of Discovery Science 2001*, Lecture Notes in Artificial Intelligence 2226, 441-446. Berlin, 2001. Springer

[9] N.S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. Banavar, and N. Federoff Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Nat. Acad. Sci USA* 97(15):8409-8414. 18 July 2000.

[10] E. Keogh, H. Hochheiser, and B. Shneiderman. An Augmented Visual Query Mechanism for Finding Patterns in Time Series Data. *Proceedings of Flexible Query Answering Systems 2002,* Lecture Notes in Artificial Intelligence 2522, 240-250. Berlin, 2002. Springer.

[11] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, and S.G. Davidson. A gene expression map for *Caenorhabditis elegans Science* 293:2087-2092.

[12] U. Ohler, U., and H. Niemann, Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends in Genetics 17*(2), 56-60, Feb 2001.

[13] S. Salzberg. Personal Communication, 2002.

[14] J. Seo and B. Shneiderman Understanding hierarchical clustering results by interactive exploration of dendrograms: A case study with genetic microarray data *IEEE Computer, 35*(7), 80-86, July 2002.

[15] C. G. Simpson, G. Thow, G. P. Clark, S. N. Jennings, J. A. Watters and J. W. S. Brown. Mutational analysis of a plant branchpoint and polypyrimidine tract required forc onstitutive splicing of a mini-exon. *RNA* 8:47-56. January 2002

[16] Spotfire. http: //www.spotfire.com.

[17] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *Proc. Nat. Acad. Sci USA* 96:2907-2912, March 1999,

[18] M. Wattenberg. Sketching a graph to query a time series database. In *Proceedings of the 2001 Conference Human Factors in Computing Systems, Extended Abstracts*, pages 381–382, Seattle WA, March 31- April 5 2001. ACM Press.

[19] K. P. White, S.A. Rifkin, P. Hurban, and David Hogness. Microarray Analysis of *Drosophila* Development During Metamorphosis. *Science* 286:2179-2184. 10 December 1999.