

Flexible Information Visualization of Multivariate Data from Biological Sequence Similarity Searches

Ed Huai-hsin Chi , John Riedl , Elizabeth Shoop , John V. Carlis , Ernest Retzel , Phillip Barry

Computer Science Department, University of Minnesota
Computational Biology Centers, Medical School, University of Minnesota

Abstract

Information visualization faces challenges presented by the need to represent abstract data and the relationships within the data. Previously, we presented a system for visualizing similarities between a single DNA sequence and a large database of other DNA sequences [6]. Similarity algorithms generate similarity information in textual reports that can be hundreds or thousands of pages long. Our original system visualized the most important variables from these reports. However, the biologists we work with found this system so useful they requested visual representations of other variables.

We present an enhanced system for interactive exploration of this multivariate data. We identify a larger set of useful variables in the information space. The new system involves more variables, so it focuses on exploring subsets of the data. We present an interactive system allowing mapping of different variables to different axes, incorporating animation using a time-axis, and providing tools for viewing subsets of the data. Detail-on-demand is preserved by hyperlinks to the analysis reports. We present three case studies illustrating the use of these techniques. The combined technique of applying a time axis with a 3D scatter plot and query filters to visualization of biological sequence similarity data is both powerful and novel.

Keywords: Information Visualization, Biomedical Visualization, Multimodal and Multidimensional Visualization, Applications of Visualization.

1 INTRODUCTION

Information visualization is becoming increasingly important as researchers discover different domains of multidimensional data. Various techniques have been developed to map multidimensional data to three-dimensional scenes, since data from different fields often require dramatically different visualization techniques. Recently, we presented a new technique for visualizing biological sequence similarity information [6].

Sequence similarity analysis is the comparison of a single sequence against known sequences kept in databases. Similarity analysis provides possible protein functions for the unknown input sequences, reducing the need for painstaking lab work [11]. Often, similarity reports include hundreds or thousands of *alignments* (matching segments between the input sequence and one of the database sequences); included with each alignment are measurements of how well the input sequence segment matches the database

sequence segment. The entire report can be hundreds or thousands of pages long.

Our earlier efforts produced AlignmentViewer (AV), which greatly improved biologists' ability to discover features in this information space. Our group has been using AV on a daily basis for the past 18 months. However, seeing the possibilities offered by information visualization, the biologists in our group became interested in many unrepresented variables that are either in the similarity report, or related to it. Examples of such variables include different similarity measures and the submission date of the matching database sequence.

These previously unrepresented variables exacerbate the intrinsic mismatch in dimensionality between the data and the visualization. Since the data is multidimensional and the graphical system is three-dimensional, we inevitably overload the dimensionality of the screen when we map the data to the screen. How should we choose which variables to map onto the dimensions we can display on the screen? We need to incorporate substantial new capabilities into AV that allow exploration of previously unrepresented variables. Since biological sequence similarity analysis is not an analysis of a physical model, our imagination is not constrained by an underlying physical model. The data are abstract and we are free to decouple and remap the variables to any of the spatial axes. Moreover, we add a time axis for supporting an additional variable. Certain variables, such as the submission date, map naturally onto the time axis. However, we may use the time axis to support animation over any variable. So the three spatial axes and the time axis all together map four variables at a time to the screen.

An additional problem is that the data set is not just multidimensional but also large, making screen real-estate precious. To reduce screen clutter, we introduce filters on each of the variables to further support analysis. Users can construct queries based on a range for each variable. The visual query filters provide an easy-to-use query interface for the information in the report.

In this paper, we present an implementation of the above techniques within the framework of AV. The contributions of this work are the application of a set of techniques that, although not entirely new when considered separately, provide a powerful interface to our multivariate data when combined:

We apply glyph techniques used originally in AV in the context of a larger visualization system. We demonstrate how glyphs can be used in conjunction with other projection techniques.

We decouple the variables from the axes, and allow the user to interactively choose mappings of variables to the axes. We demonstrate the usefulness of this interactive mapping with our data.

We use VCR-like animation controls to show the advantages of the added time axis. We demonstrate the addition of a time axis to a three-dimensional visualization system, raising the total number of dimensions to four.

email: echi@cs.umn.edu

Department of Computer Science, University of Minnesota, 4-192 EE/CSci Building, Minneapolis, MN 55455.

Computational Biology Centers, Medical School, University of Minnesota, Box 196, UMHC, 1460 Mayo Building, 420 Delaware St. S.E., Minneapolis, MN 55455

We apply visual query filtering techniques to each of the variables. We demonstrate how biologists can use these visual query filters to narrow their analysis.

We apply dynamic multivariate exploratory analysis to a new domain and demonstrate the benefits of a new system for viewing data from the field of molecular biology.

The remainder of the paper is structured as follows. In the next section, we present related work. In Section 3 we discuss the design of our new visualization system. Section 4 is devoted to case studies illustrating the features of the new technique. Finally, Section 5 contains concluding remarks.

2 RELATED WORK

Much has been done in the field of multivariate visualization, so our treatment of related work is illustrative rather than comprehensive. We'll focus in this section on work most closely related to ours, including techniques used in dynamic multivariate statistics systems, additional techniques for displaying high dimensional data, and other biological sequence visualizations.

Statisticians have investigated the display and exploration of multidimensional data. Common techniques used in systems such as PRIM-9 [18] and MacSpin [9] include two and three dimensional scatter plots, the ability to rotate 3D displays dynamically, simple animation capabilities, and the ability to mask or mark subsets of the data. Many of these packages center around a concept called projection pursuit, which is the ability to step through different projections. In many cases, very high dimension pointsets are considered. Projection pursuit techniques ignore the semantics of the different variables, and treat every axis equally. Because exhaustively enumerating all possible projections is prohibitive when the dimensionality is large, "grand tour" algorithms have been developed to automatically choose sequences of different projections [3, 7].

In the visualization community, a number of techniques have been used for displaying high dimensional data, including glyphs, worlds-within-worlds, and parallel coordinates. In the use of glyphs, two or three variables of a datum are often used to position a small marker representing the datum, while a number of other variables are encoded by the marker's size, color, etc. In worlds-within-worlds, a point in 3D is first specified, then a second smaller frame is displayed at this point. A surface can then be drawn using a new coordinate system within the second frame [10]. Parallel coordinates lays out major axes in parallel with each point represented by a line connecting each axis [14, 15]. The use of these and related techniques usually involves interaction. For example, a user controlled probe may indicate a location at which a world-within-world frame is then displayed to provide more information. Or a user may specify a range in the parallel coordinate technique to filter out lines outside that range. Another example, from user interface research, uses two dimensional scatter plots in conjunction with interactive query filters in [1]. Dynamic interactive capabilities have been found to be essential in exploring high dimensional data, such as the network visualization system in [4].

Previous work in biological sequence visualization concentrated on *single sequence representations*, which are alternatives to the DNA alphabet. The H-Curve, W-Curve, and chaos game representation are iterative methods for representing a long DNA sequence [12, 16, 19].

Our work is both similar to and different from these related projects. Our system has many similarities to the dynamic multivariate statistics packages. For example, our new system uses scatter plots in some situations, allows different projections of the data, continues to allow interactive geometric transformations of

the scene, and incorporates some, albeit different, data filtering capabilities. It differs from most such systems in that it still uses the glyph representation in many situations and has more emphasis on the use of animation for data display and exploration than most statistics packages. Furthermore, while both our system and certain statistics systems allow easy user defined mappings of variables to axes, our variables have strong semantics associated with them and our users are likely to know which mappings are most useful in given situations. We therefore did not make use of techniques like projection pursuit that treat different variables equally.

Our work also shares some of the characteristics of the higher dimensional visualization techniques mentioned. We use scatter plots and visual query filters similar to [1], and found the technique extremely useful for sequence data. Similar to worlds-within-worlds, each of our points opens up to another world in certain mappings, which is described by a glyph shaped like a comb. Similar to the parallel coordinates method, we use sliders to filter and construct queries. While our technique has similarities with each of the above methods, there are some obvious differences. Our system makes a significant departure from [1] by using glyphs and incorporating an additional time axis, thus introducing animation as an additional tool for correlation between variables. The worlds-within-worlds method has been successfully used for examining point information that are dense in multidimensional space, such as points on a hyper-surface or values in a vector field [10]. However, our multivariate data are more sparse, requiring a different approach. No obvious method exists for modifying the parallel coordinate technique to depict the alignment itself, since there can be hundreds or even thousands of matching positions in the alignment. Our employment of the glyph technique makes the alignment itself and its associated data visible. Rather than applying these higher dimensional visualization techniques individually, our system combines these different techniques, providing a simple but powerful set of tools for exploring data. Moreover, because of the interaction between different tools, their combined use provides more capabilities than if we had applied the tools independently.

While our system visualizes biological sequence information, it differs significantly from the single sequence representations mentioned above. While such representations can find interesting features of individual sequences, they are difficult to use for comparing sequences. Comparison of two sequences would involve detailed visual inspections of a pair of three-dimensional curves or 2D plots. Further, while single sequence representations are valuable for viewing small amounts of sequence data, they are simply not designed for large datasets.

Given the complexity of our multivariate data, our motivation is to combine various techniques into a single method that enables biologists to discover relationships that would otherwise be difficult to discover due to the dimensionality of the data.

3 DESIGN

In this section, we present the design of our system. We describe the variables comprising our data, then present the representation we have developed for mapping the variables to the axes. We then describe the time axis and its added benefits.

3.1 Variables

Our multivariate data come from alignment reports generated by biological sequence similarity algorithms. Each report consists of an input sequence and many alignments. Each alignment is a match between a subsequence of the input sequence and a subsequence of a sequence from the database. Thus, an alignment indicates a region of similarity between two sequences. Such alignments are

the basic elements in our visualization system. Associated with each alignment is the following information:

Position. The position in the input sequence where the alignment starts.

Frame, or frame number. The frame number defines how the DNA sequence is translated into a protein sequence. DNA sequences are composed of a four letter alphabet, which are used to encode the sequence of *nucleotides* (also called a *base*). Three DNA bases encode one *protein residue* (also called an *amino acid*). A DNA sequence can encode a protein sequence starting from the first, second, or third position. The starting point determines how bases are grouped into residues. When comparing a DNA sequence to a protein sequence, each encoding is tried separately.

Length. The length of the alignment measured in residues.

Entry Date. The submission date of the database sequence for the alignment into the database.

Similarity Scores and Residue Pair Scores. Similarity algorithms such as BLAST compute the similarity score of each alignment [2]. For each pair of residues in an alignment, BLAST looks up the entry in a *substitution matrix* and gets the *residue pair score*, a measure of the match strength. A positive entry corresponds to a good match, and a negative entry corresponds to a bad match [8]. BLAST then sums all residue pair scores in the alignment to obtain the similarity score.

P-value. The Poisson P-value for an alignment measures the statistical probability that an alignment could have occurred by chance. Because of its large range, P-value is commonly represented by the negative logarithm. Thus, an alignment with P-value of $1e^{-45}$ is represented by 45.

Percent Identities. The percentage of *exact matches* to the total alignment length.

Percent Positives. The percentage of *positive matches* to the total alignment length.

Bits. The amount of information in the alignment measured in “bits” using information theory [17].

PAM Evolutionary Distance. Different substitution matrices allow different degrees of mismatches and mutations. These matrices are either experimentally or theoretically derived. The *PAM (Point Accepted Mutations)* matrices use a rough measure of how many generations of evolution it would take to mutate one sequence into another [8]. For example, the PAM120 matrix allows fewer mutations than PAM250. We can obtain a rough estimate of the evolutionary distance of the alignments by statistically recomputing and normalizing the similarity scores obtained by using different PAM matrices [17].

BLOSUM Evolutionary Distance. The evolutionary distance measured using the BLOSUM matrices, which are experimentally derived from sequence data. In contrast to the PAM matrices, a low number signifies a large evolutionary distance [13].

Matrix Used. The number designating the substitution matrix used for the alignment report. For a single input sequence, AV has the ability to read in multiple alignment reports that were computed using different substitution matrices.

The PAM and BLOSUM Evolutionary Distances are not given by the alignment report, rather AV computes these as needed. The entry date is also not given in the report, but is retrieved from a separate database as the report is read into the system.

For an alignment, in addition to the above variables, there is the matching vector itself. This vector, represented by an array of integers, contains the residue pair scores of the matches starting from the first matching position. Therefore, an alignment can be viewed as a twelve dimensional point (for the above twelve variables), plus a matching vector.

3.2 Variable to Axes Mapping

The new AlignmentViewer uses three spatial axes and one temporal axis. Any of the above variables can be mapped to any of the four axes using the dialog box in Figure 1. All variables appear on the right side of the box. The four columns of radio buttons indicate which variables are mapped to the X, Y, Z, and time axis, respectively.



Figure 1: Choosing the axis mapping

When the *position* variable is mapped onto a spatial axis, an alignment appears as a comb-like glyph. The position of the comb in the 4D space is specified by the combination of the four mapped variables, with the comb extending along the position axis. We can imagine that the point opens up into a comb. An example is shown in Figure 2. The beginning, end, and relative length of the comb correspond to the beginning, end, and the length of the alignment. The comb teeth represent the matching vector. Each residue pair score in the matching vector determines the length of the corresponding tooth on the comb—the stronger the residue pair score, the longer the tooth. For example, line A in Figure 2 has a residue pair score of 17, and line B of -4 . The tooth color has a dual purpose, encoding both frame number and the sign of the residue pair score. The frame number is encoded by assigning each frame with two colors. The +1 frame alignments use red (positive score—match likely) and blue (negative—match unlikely), +2 alignments

green (positive) and yellow (negative), and +3 alignments magenta (positive) and cyan (negative).

If the variable *position* is not mapped to a spatial axis, an alignment appears as a single point with the matching vector glyph not shown. The result is therefore a 3D scatter plot. Since the matching vector specifies the alignment's composition, the glyph is important in identifying specific regions or trends in the alignment. Therefore, biologists often map the position variable to a spatial axis, resulting in the glyph representation.

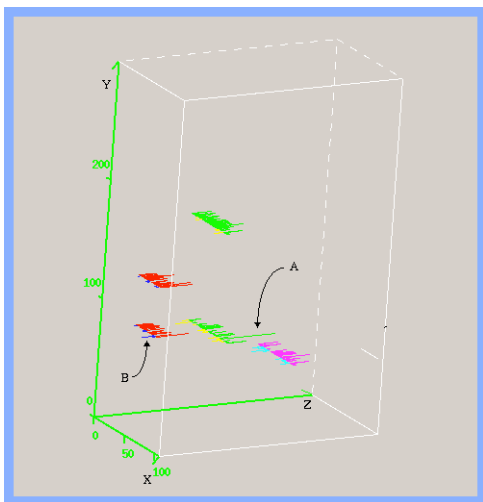


Figure 2: Several alignments represented in AV

The glyph concept in the system is a significant departure from other multivariate analysis systems that handle only point sets. This departure is caused by the fundamental difference in the underlying dataset — our data has an additional matching vector. A question for further consideration is whether we can use other high dimensional presentation techniques profitably to represent still more variables, or whether the current glyph/scatterplot representation is more useful for the biologists.

The visualization created by the old AV [6] can be recreated by mapping *position*, *score*, and *frame* to the X, Y and Z axes, respectively. For example, Figure 2 shows this mapping, and the alignment labeled B has a score of 95.

In using the original AV, biologists noticed that certain useful views were impossible to construct because the three spatial axes need to be scaled independently. For example, the user might want to zoom in on the position axis, but show the entire score axis. Thus, an additional enhancement is independent scaling of the X, Y and Z axes.

3.3 Time Axis and Animation

We need a visualization model that is understandable by the biologists—the users of the system. The addition of the time axis is useful because of its familiarity.

There was an initial bias to map only temporal variables such as *entry date* or *evolutionary distance*¹ to the time axis. There may also be a tendency not to map the temporal variables to a spatial axis. In AlignmentViewer, these restrictions are removed so any variable can be mapped onto the time axis or a spatial axis. Thus,

¹Evolutionary distance is not a direct time measure, but is a measure of the generational distance between alignments, which is a time representation in a broad sense.

a variable may have spatial or temporal quality, but we are not restricted to represent them in a certain way on the screen. So, for example, *score* can be mapped onto the time axis.

To control the time axis, AV uses the VCR-like control panel shown in Figure 3. This provides a familiar, easy to use interface controlling a number of different capabilities. The buttons near the bottom control the high cutoff slider with reverse, pause, forward play, and loop capability. The entry box to the right of these buttons specifies the stepsize between successive frames. In Figure 3, when a user presses the forward play button, the high cutoff slider increases by 20 each frame. We called this “Accumulative Play,” as the results accumulate on the screen as the high cutoff increases. Reverse accumulative play and looping are also possible.

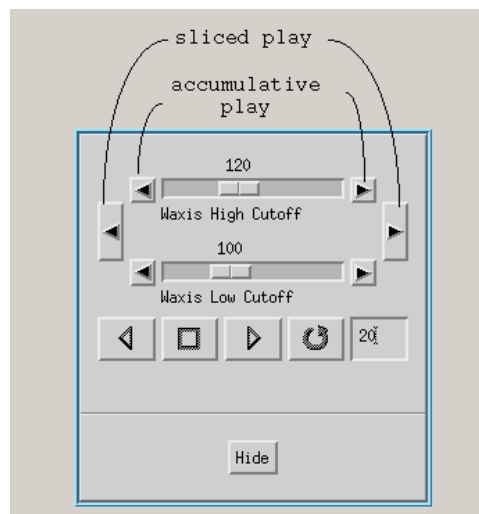


Figure 3: Time axis animation controls

The buttons flanking the sliders are more complex. The smaller buttons control forward play and reverse play on the respective sliders. So there are actually two buttons that both do forward play. The two buttons are aliased to the same function because first-time users could understand the VCR-like buttons more easily. The bigger buttons that straddle the two sliders provide “Sliced Play”. Pressing the forward sliced play button increases the high and low cutoffs simultaneously. Thus, the user is presented with “slices” of the data with respect to the time axis.

Initially our design did not allow variables to be doubly mapped onto two different axes. We found that this was actually a limitation, and thus the system allows mapping the same variable to two different axes. As an example where this might be useful, if the same variable was mapped to the X axis as well as the time axis, the user could then view successive slices of the data along the X axis using the animation controls.

3.4 Visual Query Filters

Many high dimensional exploratory statistical packages have the capability to mask or mark a subset of the data (see, e.g. [9]). These techniques reduce clutter by filtering, which makes navigation through the information space easier. The animation capabilities of our system provide one type of filtering, but we also provide additional slider-based filtering. Even though our system allows only four variables to be represented on the screen, we allow the user to specify a filtering range for any variable. This is done by the use of simple sliders for a high and low cutoff. In Figure 4, the user has specified to view only alignments that start from position 0 to position 399, with the length greater than 104 but less than 1787.

The user can change the range on any of the twelve variables dynamically, and see the result immediately. Biologists can explore the information space interactively in real-time, seeing the result of the constructed query as they change the value of the high and low cutoffs.

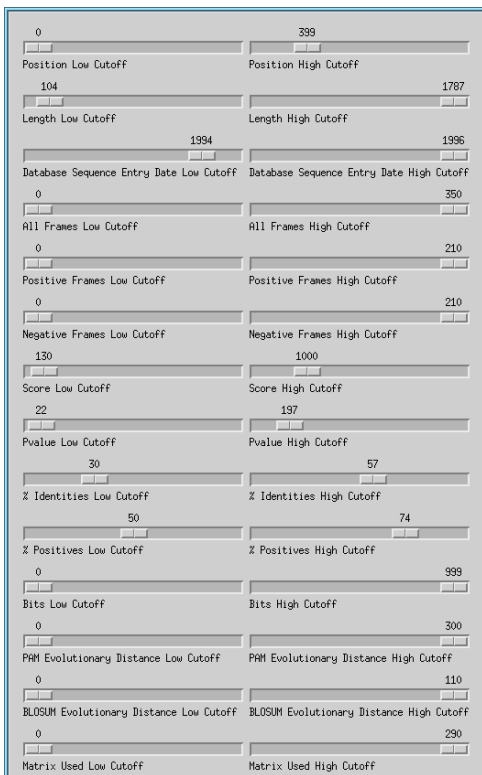


Figure 4: Visual query filters for a simple query

This capability allows biologists to analyze different groups of alignments from an alignment report. Suppose a biologist wants to analyze an alignment report of a sequence 1000 residues long that contains many alignments along different regions of the sequence, and wants to view two regions of alignments separately. To analyze the first group, she simply filters out all alignments except the first region, say positions from 0 to 500. Then she can further examine these set of alignments by constructing animations, 3D scatter plots, or even further filtering. After examining this group, she can reset the filters to encompass only alignments between positions from 500 to 1000 for further analysis.

Our exploration with AV led to the implementation of a new system that allows the arbitrary mapping of variables to the four axes, where the mapping is specified by the user interactively. The user can animate the visualization or filter the result by building visual queries. In the next section, we present some interesting uses of this new system, and demonstrate how the decoupling of variables from the axes enables new types of analysis.

4 CASE STUDIES

The enhanced capabilities described in the previous section provide significantly more power in exploring the information space. In this section, we illustrate these enhancements using sequences that are important in molecular biology and of interest to biologists in our research group.

4.1 Similarity Analysis of a Plant Sequence

We first analyze a DNA sequence from the well-studied plant *Arabidopsis thaliana* (mustard weed). Figure 5 shows a 3D scatter plot of three measures of similarity for *Arabidopsis* sequence 10G8T7P. Percent identities, similarity score, and P-value are mapped to the X, Y, and Z axes, respectively. In general, as similarity score increases on the Y-axis, we expect the percent identities and the P-value to increase. Therefore, we expect the alignments to fall mostly on the diagonal from the origin to the top right back corner.

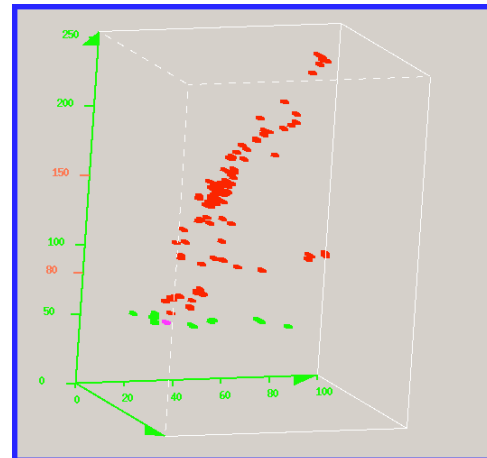


Figure 5: 3D scatter plot of *Arabidopsis* sequence 10G8T7P: the X, Y, Z axes are percent identities, score, and P-value, respectively

As the score increases, we do see percent identities increasing in general. In another rotated view that is not shown, the scatter plot also shows P-value increasing as expected. However, there are two lines of points, one red and one green, that extend to the right without corresponding increase in score — these alignments have high percent identities but low scores.

Using the visual query filters dynamically, we notice one particular variable affects those stray points the most — alignment length. We then animate the visualization using length as our time axis. The animation, several frames of which appear in Figure 6, shows the stray points correspond to very short alignments. This accounts for the low scores even though the percent identities are high.

Glyphs and points in AV are hyperlinked with the actual alignment report in hypertext HTML format. Clicking on the points reveals the stray points correspond to alignments containing a light-harvesting complex chlorophyll binding protein. The short alignments with high percent identities correspond to “motifs” that are highly conserved in the binding protein. Motifs are regions that have been preserved with little change over evolution, presumably because their function is important to the survival of the organism.

This example uses the common approach of exploring general trends and outliers to identify interesting features in the data. Moreover, the visual query filters aid in finding these features, because the user can interactively explore the correlation between variables.

4.2 Similarity Analysis of HIV

The visual query filters can aid enormously in constructing interesting queries for a large alignment report. In our previous paper [6], we analyzed the *Human Immunodeficiency Virus* (HIV) sequence, and found an interesting region containing alignments between HIV and the *Simian Immunodeficiency Virus* (SIV). Here we demonstrate how a similar analysis can be performed with greater precision using visual query filters.

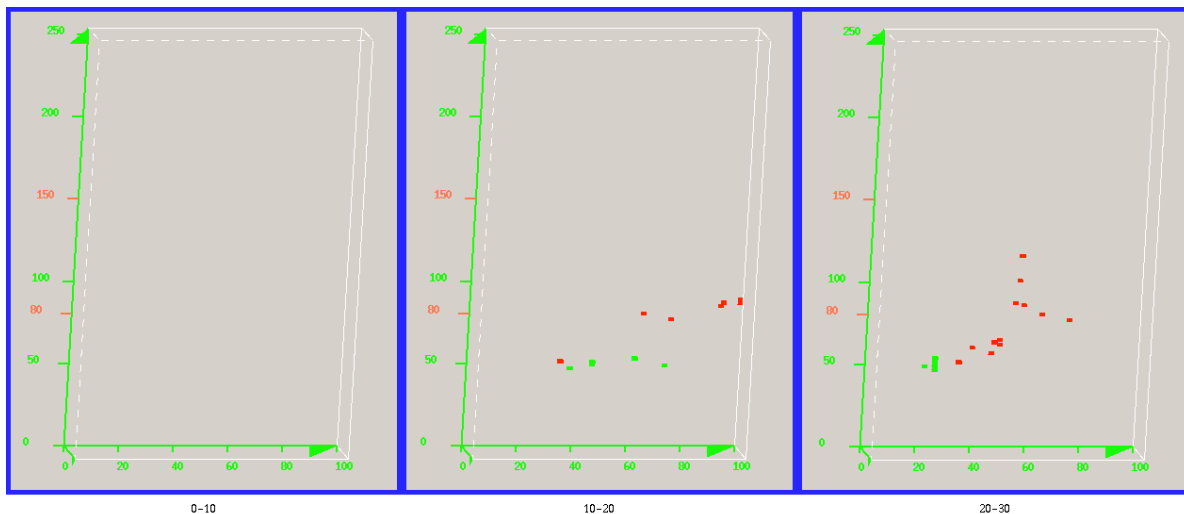


Figure 6: Using 3D scatter plot with an additional time axis to animate 10G8T7P. The X, Y, Z, and time axes are percent identities, score, P-value, and length, respectively. The left, middle, and right snapshots are frames representing length 0–10, 10–20, and 20–30, respectively.

In addition to isolating that region, suppose we are interested in alignments that are longer than 100 residues, because we want to filter out short motifs and look for a long alignment based on weak similarity. So we know that the alignment should not be an exact replica, but rather a loose identity. Thus, we take the number of matching positions (percent identities) to be between 30 and 70 percent, and the number of acceptable substitutions (percent positives) in the alignment to be between 40 and 75 percent. In addition, we last examined this sequence in 1994, and want to look at new alignment information only. Lastly, we are looking for a relatively strong hit, so the score should be at least 130, and the P-value should be better than 10^{-22} .

The visual query filter constructed for the above statement is shown in Figure 4 in the design section. The position, score, and P-value are mapped to the X, Y, and Z axes, respectively. The result appears in Figure 7. Instead of a scatter plot of points, we see that each point is expanded into a glyph, and most short cluttering alignments have been filtered out, making any subsequent navigation around the information space faster and focusing our attention on the data of interest.

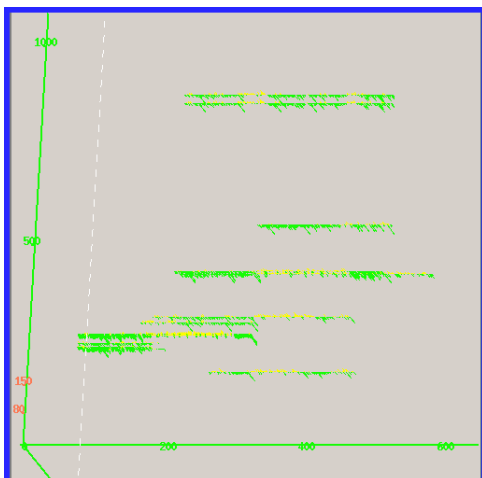


Figure 7: A visual query filter result showing alignments between HIV and SIV.

The glyphs in this figure are useful because they show large concentrations of positive colors that may indicate conserved regions. Conserved regions are important because they suggest where structural motifs or functional domains could be present. Remember that green is used as a positive color, representing a good substitution between two residue pairs. The alignments have green regions flanking large yellow regions. These alignments are to SIV sequences, thus marking a region of difference between the HIV and the SIV sequence. Since SIV is similar to HIV but cannot infect humans, biologists are very interested in such differences.

4.3 Animating HIV's Similarity Data History

Our final example illustrates how AlignmentViewer's animation capabilities can be used to show the history behind the investigation of the HIV sequence. The idea is to study the entry dates of sequences related to HIV.

Figure 8 uses AV's accumulative play feature, showing the information gathered on HIV from 1979 to the present. Position, score, frame, and entry date are mapped onto the X, Y, Z, and time axes, respectively. The glyphs represent the composition of each alignment horizontally, since position is mapped onto the X axis. The animation moves smoothly over all years, but only certain representative frames appear in this figure. The representative frames begin with 1979 (the date of the first database entries). AIDS was first diagnosed in 1981, and the frame for that year provides evidence of this discovery. Research between 1982 and 1984 revealed HIV to be the causal agent of AIDS, and an increase in the number of alignments occurred in 1985. A larger increase in the number of alignments occurred in 1986 when HIV gained international attention. Still another large jump happened in 1989, shortly after the U.S. government provided large grants for HIV research. The first automated sequencing machines were also put on the market around 1989. The last frame shows the current state of information related to HIV, displaying all 6692 alignments to the GenBank database. The HIV textual report for these alignments is roughly 3200 pages. In [6], we demonstrated how AV revealed high-level features in the HIV report that were previously hard to find. We also showed how the detailed information on each alignment is either represented by the comb glyph or provided by hyperlinks to the original textual report.

This animation provides a glimpse into the history of informa-

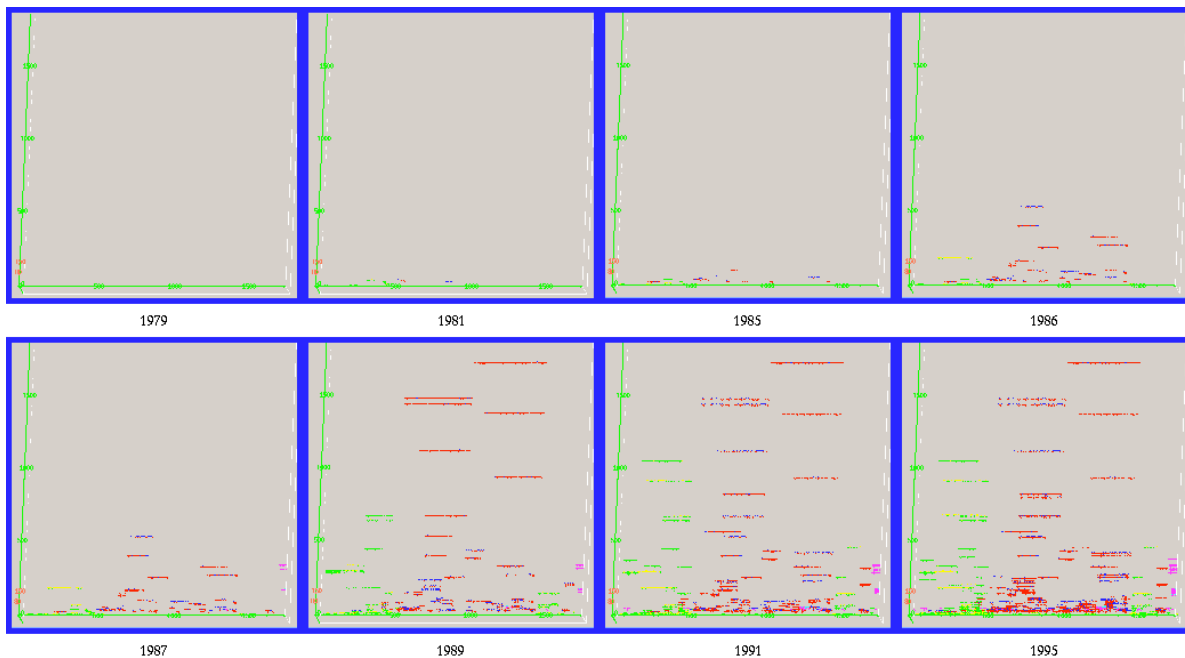


Figure 8: HIV's alignment report animated with position, score, frame, and entry date mapped onto the X, Y, Z, and time axes, respectively. The amount of sequence information related to HIV grows rapidly. ('81: diagnosis of AIDS, '82 to 85: HIV linked to AIDS, '86: HIV gained international attention, '89: large grants supporting HIV research.)

tion related to HIV. It also illustrates the growth of the sequence databases. GenBank [5], the primary repository for DNA sequence data, contains roughly 463,800,000 bases in 686,000 sequence records as of February 1996, and is doubling every 14 months.

We demonstrated how our system visualizes the history of HIV and the growth of the accumulated sequence data related to HIV as a short animation. These techniques have proven useful in our analysis, and provide evidence of the power of the combined technique.

5 CONCLUSION AND FUTURE WORK

Information visualization confronts abstract data by graphically representing relationships between pieces of data, and visually bringing out features inherent in the data. We often encounter multivariate data, but with only a few dimensions to explore on screens with limited resolution, we must think of creative ways to depict the data. Previously, we presented a new technique for representing information contained in biological sequence similarity search reports that maps three of the most important variables—*position*, *score*, and *frame*—to the screen [6].

Seeing the potential of information visualization, the biologists on our research team asked for ways to visualize other information in the data. Responding to this need, we developed a technique that allows the user to arbitrarily map any of twelve variables onto the X, Y, Z, and time axes. We showed how biologically significant features can be investigated with the additional time axis. Moreover, we showed how the user can develop simple queries on this data using visual query filters, and how the filtering reduced the clutter in the information space. The case studies showed the power of the combined technique in finding, extracting, exploring, and analyzing features that were hard to find previously. We showed how the added time axis and visual query filters provided an effective way to analyze a plant sequence. We also showed the history of the HIV sequence by visualizing the increase in HIV related sequence information over the years using AlignmentViewer. The combined approach of arbitrary mapping of variables to axes, animation, and

visual query filtering makes exploration in the information space more productive.

The HIV history animation and other related information on AlignmentViewer can be found at AV's home page (<http://www.cs.umn.edu/~echi/av.html>). In addition, AlignmentViewer is in daily use by the biologists, and 45,000 AV visualizations can be found in the similarity reports of plant genome sequences at our project's home page (http://lenti.med.umn.edu/general_cdna/wais_search.html).

There are many possible directions for further work. One possibility is to link the database to the visualizer directly to explore different interfaces between the visualizer and the database. A second possibility is to use more powerful and flexible filtering techniques. An additional possibility is to explore ways to visualize multiple search reports simultaneously.

We have extended AlignmentViewer with a set of powerful visualization techniques, based on feedback from the biologists. By decoupling the variables and allowing these variables to map to any of the three spatial and the one temporal axis, AlignmentViewer enables molecular biologists to study the intricate relationships between similar sequences.

Acknowledgments

This work has been supported in part by the National Science Foundation under grants BIR 940-2380 and CDA 9414015. We wish to thank members of the Arabidopsis sequencing group at Michigan State University and the genomic database group at the University of Minnesota for their advice and suggestions.

References

- [1] C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proc. ACM CHI '94 Conference*, pages 313–317, 1994.

- [2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [3] D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Statistical Computing*, 6(1):128–143, 1985.
- [4] R. A. Becker, S. G. Eick, and A. R. Wilks. Visualizing network data. *IEEE Transaction on Visualization and Computer Graphics*, 1(1):16–28, 1995.
- [5] D. Benson, M. Boguski, D. Lipman, and J. Ostell. GenBank. *Nucleic Acids Research*, 22(17):3441–3444, 1994.
- [6] E. H. Chi, P. Barry, E. Shoop, J. Carlis, E. Retzel, and J. Riedl. Visualization of biological sequence similarity search results. In *IEEE Visualization '95*, pages 44–51. IEEE CS Press, 1995.
- [7] S. Crawford and T. Fall. Projection pursuit techniques for visualizing high-dimensional data sets. In *IEEE Visualization '90*, pages 94–108, 1990.
- [8] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3*, chapter 22, pages 345–352. National Biomedical Research Foundation, 1978.
- [9] A. W. Donoho, D. L. Donoho, and M. Gask. Macspin: Dynamic graphics on a desktop computer. In W. Cleveland and M. McGill, editors, *Dynamic Graphics for Statistics*, pages 331–352. Wadsworth and Brooks/Cole, Belmont, Calif., 1988.
- [10] S. Feiner and C. Beshers. Visualizing n -dimensional virtual worlds with n -vision. *Computer Graphics*, 24(2):37–38, 1990.
- [11] W. Gish and D. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3:266–272, 1993.
- [12] E. Hamori and J. Ruskin. H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry*, 258(2):1318–1327, 1983.
- [13] S. Henikoff and J. Henikoff. Performance evaluation of amino acid substitution matrices. *Proteins: Structure, Function, and Genetics*, 17:49–61, 1993.
- [14] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization '90*, pages 361–378, 1990.
- [15] A. Inselberg and B. Dimsdale. Multi-dimensional lines I and II: Representations and proximity applications. *SIAM J. on Appl. Math.*, 54(2):559–596, 1994.
- [16] H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
- [17] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87:2264–2268, 1990.
- [18] J. W. Tukey, J. H. Friedman, and M. A. Fisher. PRIM-9, an interactive multidimensional data display and analysis system. In W. Cleveland and M. McGill, editors, *Dynamic Graphics for Statistics*, pages 91–110. Wadsworth and Brooks/Cole, Belmont, Calif., 1988.
- [19] D. Wu, J. Roberge, D. J. Cork, B. G. Nguyen, and T. Grace. Computer visualization of long genomic sequences. In *IEEE Visualization '93*, pages 308–315. IEEE CS Press, 1993.