

# A Novel Visualization Method for Biological Sequence Similarity

## Reports

Ed H. Chi<sup>1</sup>, John T. Riedl<sup>†</sup>, Elizabeth Shoop<sup>†</sup>, Phillip Barry<sup>†</sup>

<sup>†</sup> Department of Computer Science,

University of Minnesota,

4-192 EE/CSci Building,

200 Union St. SE,

Minneapolis, MN 55455

November 30, 1999

<sup>1</sup>Work done while at U of M.

## **Abstract**

Previously, we presented a system called AlignmentViewer that uses information visualization techniques to visualize similarities between a single DNA sequence and a large database of other sequences [3, 4]. In this paper, we extend, summarize, and describe the system using several interesting case studies. We present our comb glyph technique for visualizing alignments between sequences. In this paper, we also extend the original system by incorporating computational steering, and the visualization of differences between data sets. The case studies and the new extended system present our novel approach of extracting significant relationships in the biological data set.

## 1 Introduction

In an effort to gain insight into the function of uncharacterized sequences, molecular biologists routinely use similarity algorithms such as BLAST [1] and FASTA [11] to search the databases of known sequences for similarity to an unknown sequence. An *alignment* is a similar region between the input sequence and a sequence from the database. These algorithms produce alignment textual reports that are sometimes hundreds or thousands of pages long. The task of analyzing outputs from these algorithms becomes increasingly difficult as (a) the public databases grow larger, and (b) biologists gather more and more new sequence data. As databases grow larger from identifying new sequences, similarity reports will also increase in size because more hits will be found. For large-scale sequencing projects, the number of new sequences that must be analyzed is simply an overwhelming time-consuming task. An additional problem is that similarity reports are multi-variate, containing several pieces of information for each similar region found. With the value of the variables expressed as text, it is difficult to extract features or outliers from this data. Because biologists have had no way of efficiently sifting through the data in similarity reports, they often analyze the reports superficially, by keeping only the top scoring hits. Sometimes they are not able to analyze all of the reports, even though this data contains much valuable information. They need better methods for quickly and more thoroughly analyzing these reports.

Research in scientific and information visualization have shown that interactive graphics or visual presentations are effective methods that provide ways of interpreting large amount of scientific data [10, 2]. More recently, the advent of information visualization has led to new techniques for visualizing textual data [7, 12]. In our work, we wish to develop new information visualization techniques to aid the analysis problems associated with similarity reports. If this data could be presented visually, then scanning the data for existence of significant features or outliers would be more efficient.

Reported in two previous conference papers, we have developed and implemented a novel graphical representation for alignments contained in a sequence similarity search report [3, 4]. The representation visually reveals features that are difficult to find in textual reports, and opens new possibilities in the interpretation of similarity reports by enabling interactive investigation of the alignments. Our visualization system, called *AlignmentViewer* (AV), uses this representation, and is a result of a collaborative effort between computer scientists and molecular biologists, and is in daily use by biologists via our Web site. Our previous work gave technical details on the

system to visualization experts [3, 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 [8, 9, 13]. 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.

In this paper, we describe several extensions to the AV system, and we focus on the use of AV by molecular biologists by going through a detailed case study. We also summarize the findings in the previous papers. We have extended the AV system in two significant ways. (a) We extended the system with computational steering, enabling closed-loop analysis. (b) We extended the system to enable visualization of differences between alignment reports. As we will show, these are significant extensions that enable biologists more easily analyze sequence similarity reports.

The rest of this paper contains a brief overview of our visual representation with an introductory example in Section 2, case studies illustrating the features of our system in Section 3. Finally, we conclude with ideas for future improvements of the system. We also include a glossary at the end of this paper.

## 2 AlignmentViewer's Visual Data Representation

Figure 1 shows an example of the textual output given the BLAST similarity algorithm.

We developed an visualization application that helps biologists to analyze a similarity report using visualization techniques. In Figure 2 we show an example of how AlignmentViewer represents the similarity data from a similarity algorithm report. When the biologist opens a similarity report in AV, this is the initial view given. The biologists on our team chose the following important variables for this default view:

On the X-axis, we represent the input sequence in units of protein residues. In this example, the input sequence is a DNA sequence with 486 nucleotides, or about 162 residues. On the Y-axis, we represent the similarity score. Each alignment is a similar region between the input sequence and a database sequence, and it is represented by

Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Poisson Probability	
			P(N)	N
PEP_:pir S17695 S17695 chlorophyll a/b-binding protein...	+2	249	2.3e-20	1
PEP_:pir S07408 S07408 chlorophyll a/b-binding protein...	+2	247	3.6e-20	1
PEP_:pir S00442 S00442 chlorophyll a/b-binding protein...	+2	244	7.6e-20	1
PEP_:pir S40210 S40210 20K protein of CP24 precursor p...	+2	202	2.4e-15	1
PEP_:pir S33466 S33466 chlorophyll a/b-binding protein...	+2	200	3.8e-15	1
...				
PEP_:pir S01961 S01961 chlorophyll a/b-binding protein...	+2	89	0.037	1
PEP_:pir S29904 S29904 Light harvesting chlorophyll a ...	+2	88	0.045	1
PEP_:gi 12582 gp X68333 HHLHC_1 light harvesting chlor...	+2	88	0.045	1
PEP_:pir S30624 S30624 lhcII a/b binding protein typeI...	+2	84	0.046	1
PEP_:gi 17416 gp Z18205 ATTS0710_1 LHCII A/B BINDING P...	+2	84	0.046	1
PEP_:gi 405615 gp X61609 BNLHCB3B_1 LHC II Type III ch...	+2	66	0.064	2
PEP_:gi 460769 gp X75932 HSPLKSTK_1 Serine/Threonine p...	+1	59	0.13	2
PEP_:pir S34130 S34130 protein kinase plk-1 (EC 2.7.1....	+1	58	0.21	2
PEP_:gi 403458 gp L19559 HUMSTPK13_1 protein kinase [H...	+1	58	0.21	2
...				

>PEP\_:pir|S17695|S17695 chlorophyll a/b-binding protein (clone pINEab 31) -  
Scotch pine>PEP\_:gi|20792|gp|X58516|PSLHAB1\_1 Type II chlorophyll a  
/b-binding protein [Pinus sylvestris]  
Length = 278

Plus Strand HSPs:

Score = 249 (86.5 bits), Expect = 2.3e-20, P = 2.3e-20  
Identities = 42/54 (77%), Positives = 49/54 (90%), Frame = +2

Query: 2 STPPEWLDGSLPGDFGFDPLGLSSDPDSLKWNVQAEIVHCRWAMLGAARDIHPQ 163  
+TPP WLDGSLPGDFGFDPLGL+SDP++LKW VQAE+VHCRWAMLGAA + P+

Sbjct: 80 NTPPPWLDGSLPGDFGFDPLGLGSDPETLKWVQAEIVHCRWAMLGAAGIFIPE 133

Score = 53 (18.7 bits), Expect = 0.75, Poisson P(2) = 0.53  
Identities = 8/11 (72%), Positives = 8/11 (72%), Frame = +1

Query: 190 LTIPSWYTXGE 222  
L PSWYT GE

Sbjct: 141 LNTPSWYTAGE 151

Figure 1: Excerpts of the textual similarity report for 10G8T7P

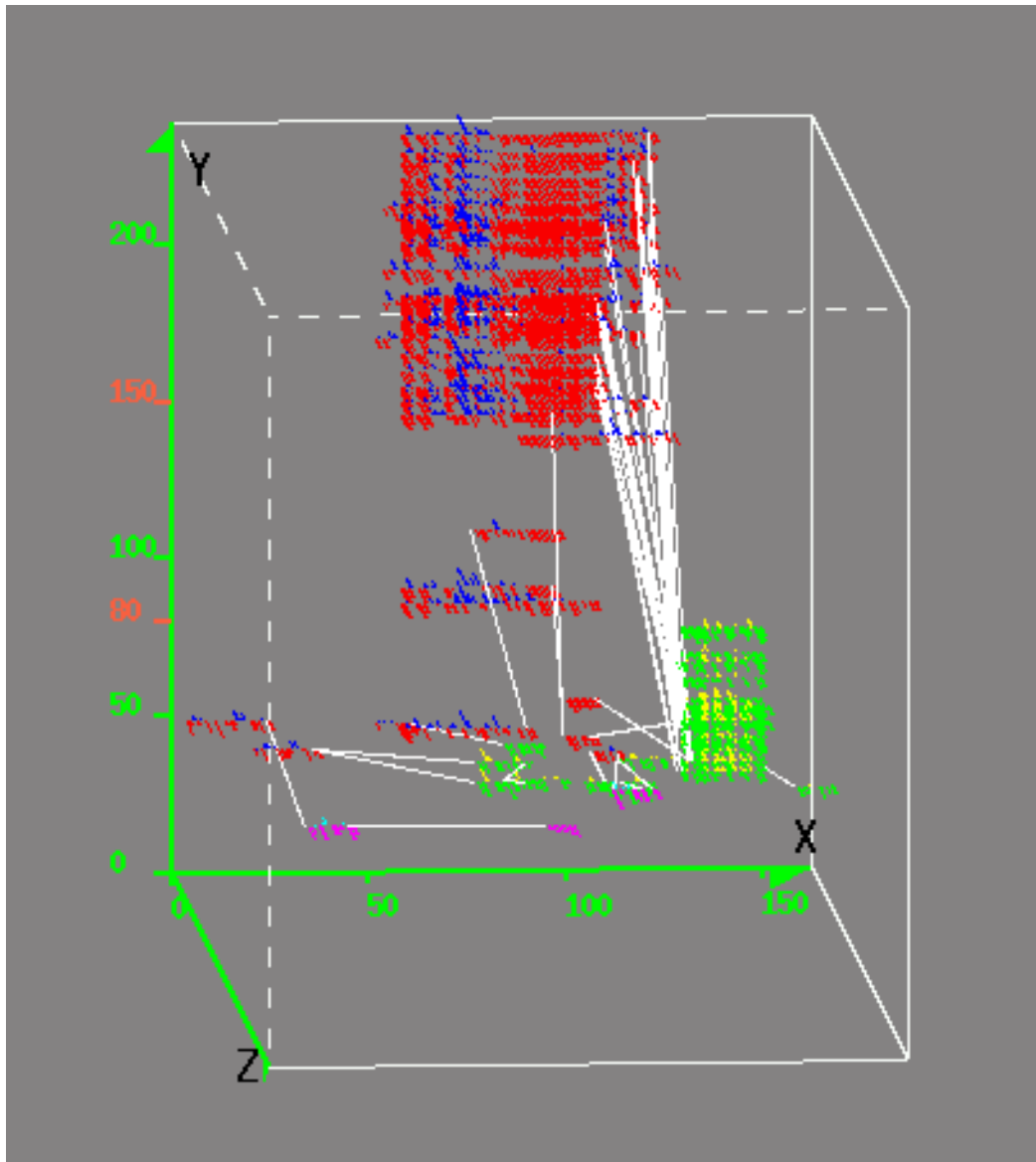


Figure 2: Similarity report for *Arabidopsis thaliana* sequence 11B11T7P represented in AlignmentViewer: X-axis is the position along the input sequence, Y-axis is the similarity score, and the Z-axis is the frame number of the alignment.

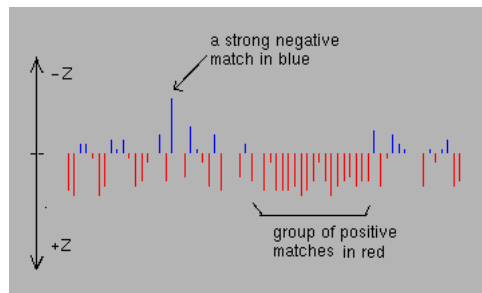
a “comb”-like object. There are 130 different combs in this figure, one for each alignment found. The length and position of the comb along the X-axis signifies the length and position of the alignment on the input sequence. For example, the top-most alignment has a score of 234 and aligns on the input sequence between 56 and 121. This representation allows a quick view of the score and position of all of the alignments.

AV depicts the frame number of each alignment. The frame number defines how the DNA sequence is translated into a protein sequence. DNA sequences are composed of a four letter alphabet (A, C, G and T for each of the four nucleotides). Three DNA bases encode one single protein residue (also called an amino acid), so there are 64 possible residue encodings. These 64 encodings represent the 20 fundamental residues with some redundancy. Protein sequences use a 20 letter alphabet for the 20 different residues. A DNA sequence can encode a protein sequence starting from the first, second, or third position. Biologists do not always know which position the encoding starts from. The starting point determines how bases are grouped into residues. DNA sequences are double stranded, with one strand named positive and the other negative. Biologists do not always know whether a particular sequence is from the positive or negative strand. Sequences from the negative strand must be translated into proteins in reverse (from right to left, as the sequence is usually written). Thus, there are six different ways of translating a DNA sequence into a single protein sequence: start from the first, second, or third nucleotide of the positive strand, or start from the first, second, or third nucleotide of the negative strand. The six ways of translation are called the reading frames of the sequence, and are labeled with frame numbers +1, +2, +3, or -1, -2, -3, respectively. Thus, when comparing a DNA sequence to a protein sequence, all six possibilities must be considered, and the frame number recorded.

AV presents reading frame information by using both colors as well as layers in the Z-direction. Comb objects in red and blue represent frame +1 alignments. Green and yellow combs represent frame +2 alignments, and magenta and cyan combs represent frame +3 alignments. The frame of an alignment is also encoded by putting the alignments in different slices on the Z-axis (not easily seen in this view). Frame +1 alignments are plotted in the first slice, with +2 and +3 in the second and third slice, respectively. AV represents negative frame alignments using a single white line (none in this figure), and users can switch between viewing positive or negative frame alignments. In Figure 2, we see that there are several long frame +1 alignments in the center of the sequence, and some short frame +2 alignments toward the end of the sequence. There are also a few low-scoring frame +3

alignments in the middle. The color coding of the frame enables us to determine which alignments are in which frame in any view.

Similarity algorithms such as BLAST compute the similarity score of each alignment [1]. For each pair of residues in an alignment, BLAST looks up the entry in a *substitution matrix* and gets the *residue pair score*, which is a measure of the match strength [6]. BLAST then sums all residue pair scores in the alignment to obtain the similarity score. The height of the comb on the Y-axis represents this similarity score.



```
>PEP_pir|JQ2217|JQ2217 myeloperoxidase (EC 1.11.1.7) precursor, anionic -
      Japanese aspen x large-toothed
      aspen>PEP_gi|217997|gp|D11102|POPPRXA1_1 peroxidase [Populus
      kitakamiensis]
      Length = 318
```

```
Score = 180 (92.5 bits), Expect = 2.6e-19, P = 2.6e-19
Identities = 32/65 (49%), Positives = 40/65 (61%), Frame = +1
```

```
Query: 169 FYDRSCPRLQTIVKSGVWRAFKDDSRIAASLLRLHFHDCFVNGCDGSILLNDSSEDFKGEK 348
      FY +CP +IV+ V A +D R A L R+HFHDCFV+GCDGSILL D+ E
Sbjct: 27 FYASTCPNVSSIIRGVVEQAAQNDVRLGAKLIRMHFHDCFVDGCDGSILLVDATGINSEQ 86
```

```
Query: 349 NAQPN 363
      + PN
Sbjct: 87 DEAPN 91
```

Figure 3: A single alignment represented using a comb object. A tooth in the negative Z-axis direction colored in blue, yellow, or cyan, is a bad pair-wise match, while a tooth in the positive Z-axis direction colored in red, green, or magenta is a good pair-wise match. Red and blue combs are frame +1 alignments (green and yellow—frame +2, magenta and cyan—frame +3).

The teeth on the comb represent the actual pair-wise matching of the alignment, and is why each comb is encoded using two colors. Each tooth on the comb represents a single pair-wise match (See Figure 3). If the

residue pair score is positive, then the replacement of the residues is considered likely and represented by the positive colors of each frame (red, green, and magenta). If the residue pair score is negative, the replacement is considered unlikely and negative colors are used (blue, yellow, and cyan). The length of the tooth also encodes the strength of the residue pair score, thus the “goodness” of the pair-wise match. If the residue pair score is positive, then the positively colored tooth points toward the positive Z-axis (the longer the tooth, the stronger the match). If the residue pair score is negative, then the negatively colored tooth points in the opposite direction (negative Z-axis). The comb thus encodes much of the composition of an alignment.

Alignments that result from the same database sequence are connected by a white line. Thus, we can determine if a sequence aligns with a database sequence in multiple regions, even when the alignments are in different reading frames.

This representation visually depicts the alignments in a single view that encodes much of the information in a similarity report. This view immediately conveys to the viewer the length of the sequence and the lengths of all of the alignments. It also depicts the overall strength and the composition of the alignments. Since this view compactly represents similarity reports, some higher level features can be extracted that otherwise would be difficult to discover in textual reports. For example, we can see the overall distribution of the alignments at different positions on the input sequence. We can also quickly gain insight into the overall distribution of the scores, and the number of alignments in each frame. By visualizing alignments all in one view, the biologists can quickly determine the correct frame of translation.

A feature of Figure 2 is particular noticeable—the color change of the alignments around residue 125 from red to green. A large group of red alignments (frame +1) are connected to another group of green alignments (frame +2). This color change suggests the possible existence of a *frame shift* at that position. A frame shift is when there is a “forgotten” base in the sequence. Remember that a white line between two combs means the two alignments hit the same database sequence. Therefore, the white lines that connects the two groups of alignments give further evidences for the possibility of a frame shift, most likely due to a lost base during sequencing. In our sequence data, we have identified many possible frame shifts using AV.

We can also see trends in different regions. There are regions where the matches are mostly positive, and other regions where the matches are mostly negative. Regions with many positive matches signals the possible

existence of a *conserved region* that can correspond to binding sites, or strong structural patterns such as an alpha helix. An *alpha helix* is a common structure of proteins, characterized by a single, spiral chain of amino acids. Short conserved regions with long positive teeth signals the possibility of a “motif”, since long positive teeth correspond to highly conserved residue pairs. A *motif* is frequently observed protein pattern that has biological significance.

These global features are difficult to extract from the textual report. For example, if a biologist wants to find trends or outliers using a textual report, she must correlate between different parts of the report for significant differences or commonalities. In comparison, our visual representation can be used to determine conserved regions in alignments and infer the relative composition of regions in alignments.

For other details that AV does not represent in graphical form, such as the description of the database sequence, we provide hyperlinks from the graphical representation to the actual textual report. Clicking on one of the combs using the right mouse button will bring the hypertext browser to the actual alignment. This “detail-on-demand” enables biologists to more easily navigate and explore all of the alignments in a single visual index. Biologists can use the visual representation as a visual index of all of the alignments.

We also added independent scaling (zooming) of each spatial axis. Each axis can be scaled independently. This enables ever more sophisticated analysis of the similarity results and easier “zooming in” on specific regions of interest.

For ease of comparison between reports, AlignmentViewer allows the subtraction of two different data sets, thus providing a way to visually subtract one visualization from another.

In our previous papers, we showed that our visual representation provides a good overall view of similarity reports, and enables the biologists to analyze large amounts of similarity reports effectively [3, 4].

### **3 Walk-Through of the Features using Case Studies**

In this section, we present the features of AV, and demonstrate the power of AV in the context of analyzing several sequences. We also describe the new features we have added: (1) computational steering, and (2) visualization of multiple data sets by looking at the differences between sequence similarity reports.

### 3.1 Closed-loop Analysis via Computational Steering

Let us demonstrate how a biologist analyzes a sequence with AlignmentViewer. First the biologist opens the application, and chooses a sequence using a file selection box. The sequence used for this walk-through is a 546 base sequence (10G8T7P, GenBank T04685) from *Arabidopsis thaliana*. If a BLAST or FASTA similarity report exists, then AlignmentViewer finds and loads the dataset. If a similarity report for this sequence is not available, AV asks the biologist how to process the input sequence by popping up the steering control panel, as shown in Figure 4.

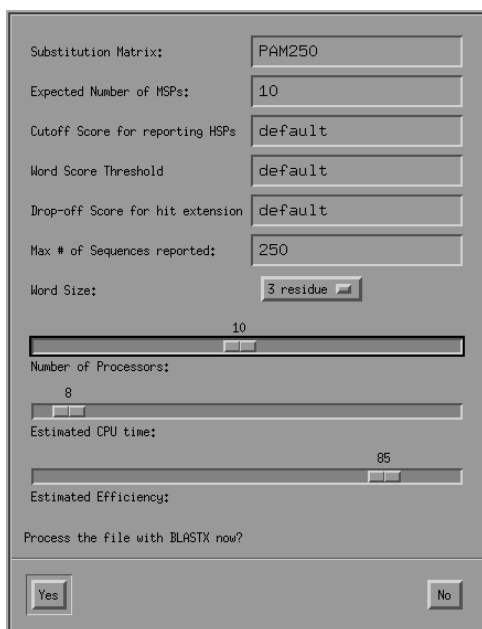


Figure 4: AlignmentViewer's Steering Control Panel: This sequence requires 10 processors to get the response time under 20 seconds on a SGI Challenge XL machine. The efficiency is estimated to be 89 percent.

The initial input parameters to the BLASTX algorithm are shown on the top half of the control panel, and the machine steering control parameters are shown on the bottom half. AlignmentViewer predicts that the computation requires 69 seconds using one processor. We obtain these performance prediction by establishing a parallel performance model of the BLASTX algorithm, and then using the model to extrapolate the running time under different situations [5]. By doing the computation on multi-processor machines, we can significantly reduce the running time. Suppose that the biologist wants the response time to be shorter, say 8 seconds. She drags the

response time slider to 8 seconds. The system computes the required number of processors (10 processors on a SGI Challenge XL), and the utilization of computational resources (85 percent efficiency).

In the past, the process of running the similarity algorithm and then using AlignmentViewer to analyze the data is a tedious cycle. In this new feature, we close the loop of this analysis cycle by allowing the user to run the similarity algorithm directly in AlignmentViewer and then analyzing the result immediately. Closed-loop analysis enables the biologist to perform multiple “what-if” scenario analyses without having to rerun the algorithm outside of the system, waiting for it to finish, parsing the results, and then finally feeding the data back into the visualization application. This convenience allows the biologists to focus on the analysis instead of the details of running the algorithm.

Because biologists often run similarity algorithms on the same sequence with different parameters, our system incorporates the parallel BLAST algorithm directly with the visualization application. Biologists directly invoke the algorithm with different parameters and see the visual representation of the results immediately. By utilizing parallel computation, AV computes the results in interactive time. Thus, biologists are allowed to “steer” the computation and the analysis by changing the parameters of the algorithm.

### **3.2 Variable to Axis Mapping**

Each alignment is associated with twelve variables, which are shown in Figure 5<sup>1</sup>. 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 5. The four columns of radio buttons indicate which variables are mapped to the X, Y, Z, and time axes, respectively.

Since the position variable might not always be mapped onto a spatial axis, we need an adjustment to the representation. If the position variable is not mapped onto any of the three spatial axes, then the alignment is represented simply as a single point. Each point is still colored using the positive color of that frame. Thus, AV is not confined to the representation as explained in Section 2.

Figure 6 shows a 3D scatter plot of three measures of similarity for the 10G8T7P sequence. Percent identities, similarity score, and P-value are mapped to the X, Y, and Z axes, respectively. In general, as similarity score

---

<sup>1</sup>The value for these variables are not always provided directly by every similarity algorithm. However, AlignmentViewer computes the missing values as necessary.



Figure 5: Choosing the axis mapping using the axis mapping panel. The four columns of radio buttons indicate which variables are mapped to the X, Y, Z, and time axes, respectively.

increases on the Y-axis, we expect the percent identities to increase. Therefore, we expect the alignments to fall mostly on the diagonal from the origin to the top right corner. 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 decreasing as expected. However, there are two lines of points (marked as A and B) that extend to the right without corresponding increase in score—these alignments have high percent identities but low scores. This is a significant feature extracted from the report.

This feature gives biologists the flexibility of exploring the different possible correlations between the variables of a similarity report using variable to axis mapping. This mapping increases the biologists' ability to find interesting features in the similarity report as shown in the above example. In [4], we explored this feature in more detail and demonstrated the usefulness of the ability to map any of the variables to any of the axis in the representation.

### 3.3 Time Axis and Animation

Any variable can be represented on the “temporal axis” by using the variable to axis mapping. To control the time axis, AV uses the VCR-like control panel shown in Figure 7. 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 step size between successive animation frames. 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. The buttons marked as “*Accumulative Play*” provide the same functionality<sup>2</sup>. The buttons marked as “*Sliced Play*” provide a different kind of animation. 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.

The VCR-like buttons provide easy-to-understand controls for first-time users, and the other controls allow more sophisticated types of animation. This design enables biologists to view both increasing or decreasing sets of alignments, or slices of information.

Next we animate the visualization for 10G8T7P using length as our time axis. Several animation frames

---

<sup>2</sup>These Accumulative Play buttons do not control the sliders, which we acknowledge as potentially confusing

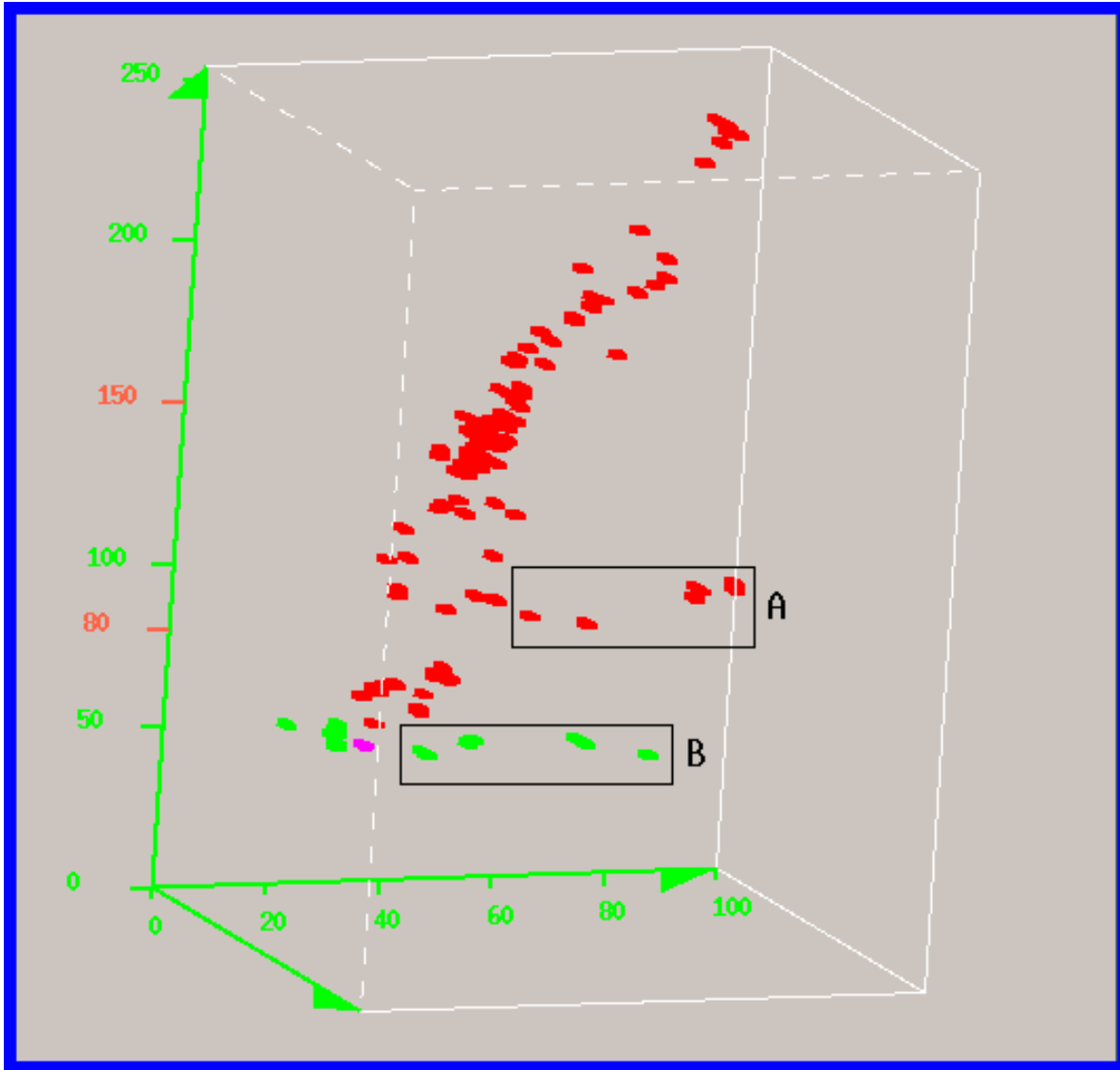


Figure 6: 3D scatter plot of *Arabidopsis* sequence 10G8T7P: the X, Y, Z axes are percent identities, score, and P-value, respectively. Two lines of points (marked as A and B) represent alignments with high percent identities but low scores.

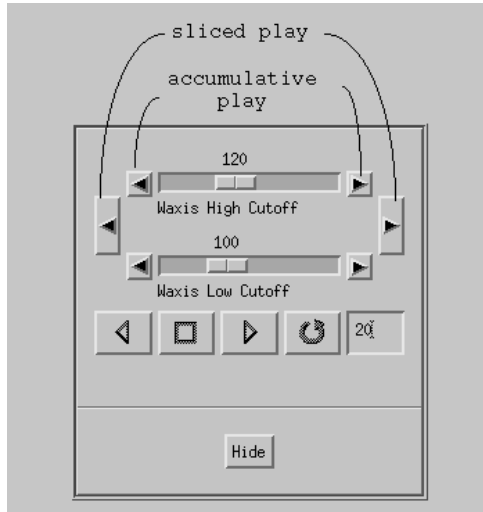


Figure 7: Time axis animation control panel. The buttons marked as *Accumulative Play* and the buttons near the bottom control the high cutoff slider with reverse, pause, forward play, and loop capability. The text entry box near the bottom specifies the step size between successive animation frames. Pressing the buttons marked as *Sliced Play* increases or decreases the high and low cutoffs simultaneously.

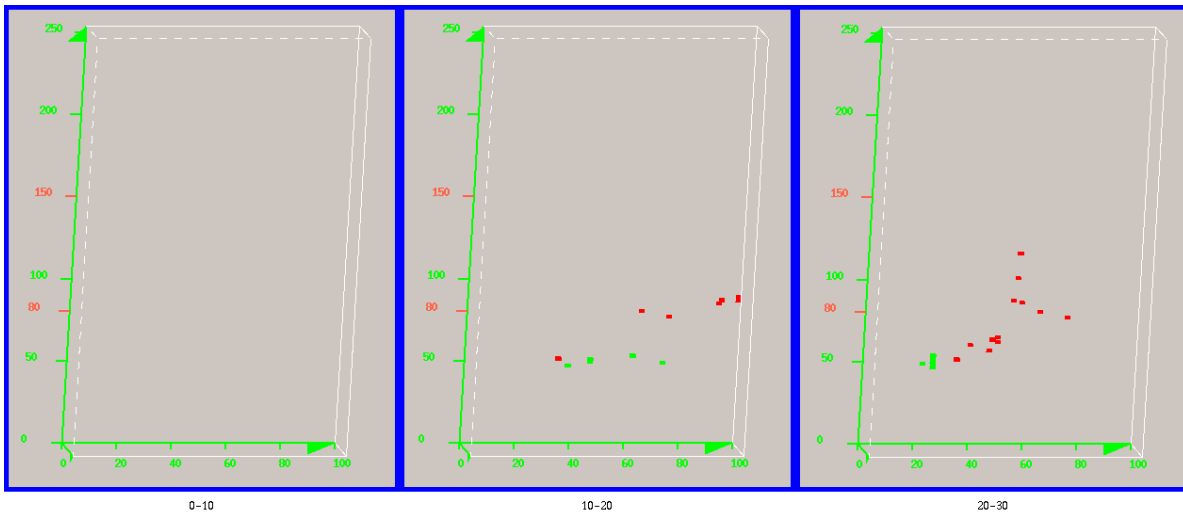


Figure 8: Using 3D scatter plot with an additional time axis to animate 10G8T7P. The X, Y, Z, and time axes are percent identities, score, P-value, and length, respectively. The left, middle, and right snapshots are animation frames representing length 0–10, 10–20, and 20–30, respectively.

appear in Figure 8, showing that the stray points correspond to very short alignments. This accounts for the low scores even though the percent identities are high. Combs and points in AV are hyper-linked 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 short regions that have been preserved with little change over evolution, presumably because their existence is important to the function of the protein.

This example uses the approach of exploring general trends and outliers in a graphical output to identify interesting features in the data. The animation example here shows a significant finding for this sequence report. The reason that there are many short alignments with high percent identities is because the sequence has many “motifs” in the database. Biologists can use this animation capability to see correlations between four different variables.

### **3.4 Visual Query Filters**

Even though our system allows only four variables to be represented on the screen, the user can specify a filtering range for any variable using visual query filters. These filters are simple sliders for a high and low cutoff of each variable. In Figure 9, we have chosen to view only alignments that start from position 0 to 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. This reduces the clutter of visual information for large reports and allows users to narrow in on important subsets of the data.

Suppose we are interested in the alignments in the center of Figure 2. We set the position filter to include only alignments no longer than 75 residues with starting position between 48 and 98. Figure 10 shows the result of filtering 11B11T7P’s report to obtain these alignments. After filtering some alignments, we can further examine these set of alignments by constructing animations, 3D scatter plots, or even further filtering. As another example, we used visual query filters to obtain the comb isolated in Figure 3.

Biologists can explore the data interactively in real-time, seeing the result of the constructed filter as they change the value of the high and low cutoffs. Since larger reports will be the norm in the future, the above features will become very important. With a particular kind of alignment in mind, biologists can use visual query

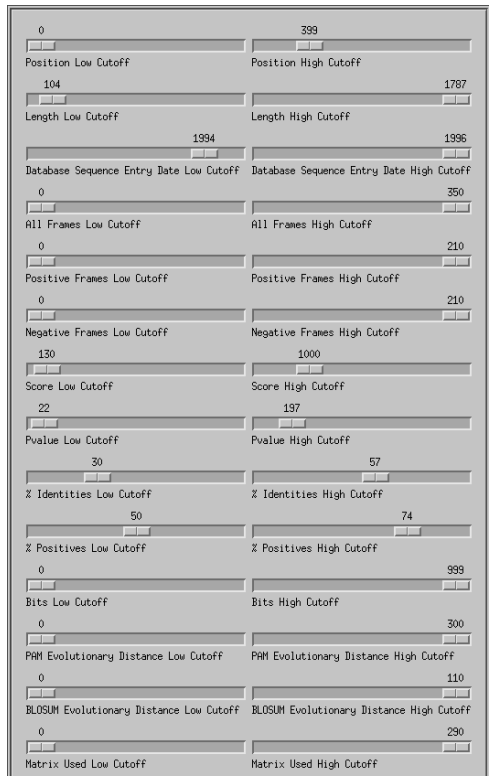


Figure 9: Visual query filters for a simple query. We have chosen to view only alignments that start from position 0 to 399, with the length greater than 104 but less than 1787, etc.

filters to focus on subsets of the data and do more detailed analysis.

### 3.5 Subtraction and Addition between Datasets

Biologists also need a way to compare and contrast different reports generated using different parameters, especially since they can generate them so quickly using an interactive steering system. How does a report generated using a specific set of parameter compare to another report? AlignmentViewer solves this problem by treating similarity reports as sets of alignments that can be compared using algebraic set operations. Thus, reports can be subtracted and added to each other, and AV shows the result of these comparisons visually.

We can compute data set differences in several different ways: (1) Most commonly, if an alignment  $X$  from one report hits the database sequence  $A$  and an alignment  $Y$  from a different report hits the same database sequence  $A$  and the regions of the  $X$  and  $Y$  alignments overlap, then they are considered to be the same. This is one possible equivalence operator. Other possible definitions include (2) an exact operator, where position of the alignments must match exactly, and (3) an equal-name operator, where two alignments are considered equivalent if they correspond to the same database sequence.

In Figure 11, the biologist is interested in similarity reports on the same sequence, but with different parameter values for the substitution matrix (left and center of the Figure). In the past, the user would run the process off-line. Instead, the biologist have the convenience of a closed-loop analysis by running the similarity algorithm directly in the system. To choose different parameters, the biologist uses the steering control panel described earlier. In this figure, we show the visualization of another Arabidopsis sequence (172C2T7, GenBank H36421), computed using two different substitution matrices: PAM60 and PAM250.

Comparing the results using PAM250 and the results using PAM60, we see many visual similarities. However, we still find it difficult to visually contrast the two images to find the exact differences. The result of subtracting the PAM250 visualization from the PAM60 visualization is shown in the right of Figure 11. The Figure shows some alignments that were found by the computation using the PAM60 matrix, but not found by PAM250.

To determine the alignments in the difference set, we use the visual hyperlinks by clicking on a few of these alignments. The results of the analysis suggests that the extra alignments found by the PAM60 matrix were “motifs” of a protein called “peroxidase.” An alignment that was found by PAM60 but not by PAM250 is shown in Figure 12.

Subtracting the dataset in the opposite direction (PAM250-PAM60) results in an empty difference set. Thus, all alignments that were found by PAM250 are also found by PAM60.

We demonstrated the ability to visually compare two similarity reports by using set operations that compute differences between the visualizations. This enables biologists to more closely understand the how varying the parameters of the similarity algorithm affects the results of the computation.

#### **4 Conclusion and Future Work**

We have developed a method for graphically representing the information contained in similarity algorithm search reports. In this paper, we have summarized our findings [3, 4] and extended the method by incorporating new features: (1) We showed the usefulness of this visual representation using several case studies. (2) We showed how AlignmentViewer, a system designed with this visual representation, enables the biologists to correlate between many different variables contained in similarity reports. (3) We demonstrated the usefulness of the ability to map any of the variables in a similarity report to any of the axes in the representation. (4) We demonstrated how animation or a time axis can be used as an additional axis for correlating the variables. (5) We demonstrated how visual query filters can be used to narrow down and select alignments for display and analysis, especially in a large report. (6) Detail-on-demand is provided interactively via hyperlinks in the visual representation.

The new features we added include computational steering and visualization of differences between data sets. (7) We closed the analysis cycle by allowing the algorithm to run interactively in the application. Within the application, we enable the biologists to choose different parameters of the algorithm, and then see the search result immediately. (8) We also demonstrated the ability to visually compare two similarity reports by using simple algebraic set operations. We showed how a biologist uses these operations to find interesting differences between algorithm runs.

There are many possible directions for further work. We believe other visualization techniques can be employed with more powerful and flexible filtering techniques. An additional possibility is to explore ways to visualize multiple search reports simultaneously. Although AV is already available on multiple platforms (SGI and Sun machines), we would like to explore the possibility of using a cross-platform environment such as Java applets.

Some animations and other related information on AlignmentViewer can be found at its home page (<http://www.cs.umn.edu/~echi/av.html>). AlignmentViewer is in daily use by the biologists, and a total of 45,000 visualizations can be found in the similarity reports of plant genome sequences at our project's home page (<http://www.cbc.umn.edu/ResearchProjects/Search/index.html>).

In conclusion, we believe that the field of computational molecular biology could greatly benefit from the use of information visualization techniques to extract and data-mine the enormous amount of knowledge acquired. We hope that our work has filled a gap in biologists' need for better tools.

### **Acknowledgments**

This work has been supported in part by the National Science Foundation under grants BIR-940-2380 and CDA-941-4015. 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] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- [2] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufman, San Francisco, CA, 1999.
- [3] 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.
- [4] E. H. Chi, J. Riedl, E. Shoop, J. V. Carlis, E. Retzel, and P. Barry. Flexible information visualization of multivariate data from biological sequence similarity searches. In *IEEE Visualization '96*, pages 133–140,477. IEEE CS Press, 1996.
- [5] E. H. Chi, E. Shoop, J. Carlis, E. Retzel, and J. Riedl. Efficiency of shared-memory multiprocessors for a genetic sequence similarity search algorithm. Technical report, University of Minnesota, 1996.
- [6] 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.

- [7] J. Dill and N. Gershon, editors. *Proceedings of the Symposium on Information Visualization '97*. IEEE CS Press, 1997.
- [8] 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.
- [9] H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
- [10] B. McCormick et al. Visualization in scientific computing. In *Computer Graphics*, volume 21. ACM Press, November 1987.
- [11] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85:2444–2448, 1988.
- [12] G. Wills and J. Dill, editors. *Proceedings of the Symposium on Information Visualization '98*. IEEE CS Press, 1998.
- [13] 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.

## 5 Glossary

**Alignment:** A biologically similar region between two sequences.

**Amino Acid:** The basic building block of a protein.

**Base:** See definition for Nucleotide.

**BLAST:** A well-known sequence similarity algorithm.

**Conserved Region:** A region of a protein or nucleotide sequence that is conserved over evolutionary time, which are generally presumed to have biological significance.

**FASTA:** A well-known sequence similarity algorithm.

**Frame number:** The frame number determines the way a DNA sequence is translated into a protein sequence—the 1st, 2nd, or 3rd nucleotide of the positive strand of the DNA sequence, or the 1st, 2nd, or 3rd nucleotide of the negative strand.

**Glyph:** A icon or symbol designed with features that represent data variables.

**Motif:** A frequently observed protein pattern that has biological significance. A motif is generally a conserved region.

**Nucleotide:** An element of a DNA sequence, sometimes also called a *base*. Three nucleotides code for a single amino acid.

**Percent identities:** The percentage of exactly matched positions in an alignment.

**P-value:** A measure of an alignment's significance, and is the Poisson probability of an alignment being statistically significant. A low P-value represent a good alignment.

**PAM distance:** Point-Accepted Mutations is a measure of evolutionary distance, which is a rough measure of how many generations of evolution it would take to mutate one sequence into another.

**Similarity Score:** A measure of the amount of similarity in an alignment. It is the sum of all residue pair scores in an alignment.

**Substitution matrix:** A two dimensional matrix representing the likelihood of one amino acid replacing another. These matrices are the foundation of statistical techniques for finding alignments.

**Residue:** An amino acid, which is a basic element in a protein sequence.

**Residue Pair Score:** An entry in a substitution matrix.

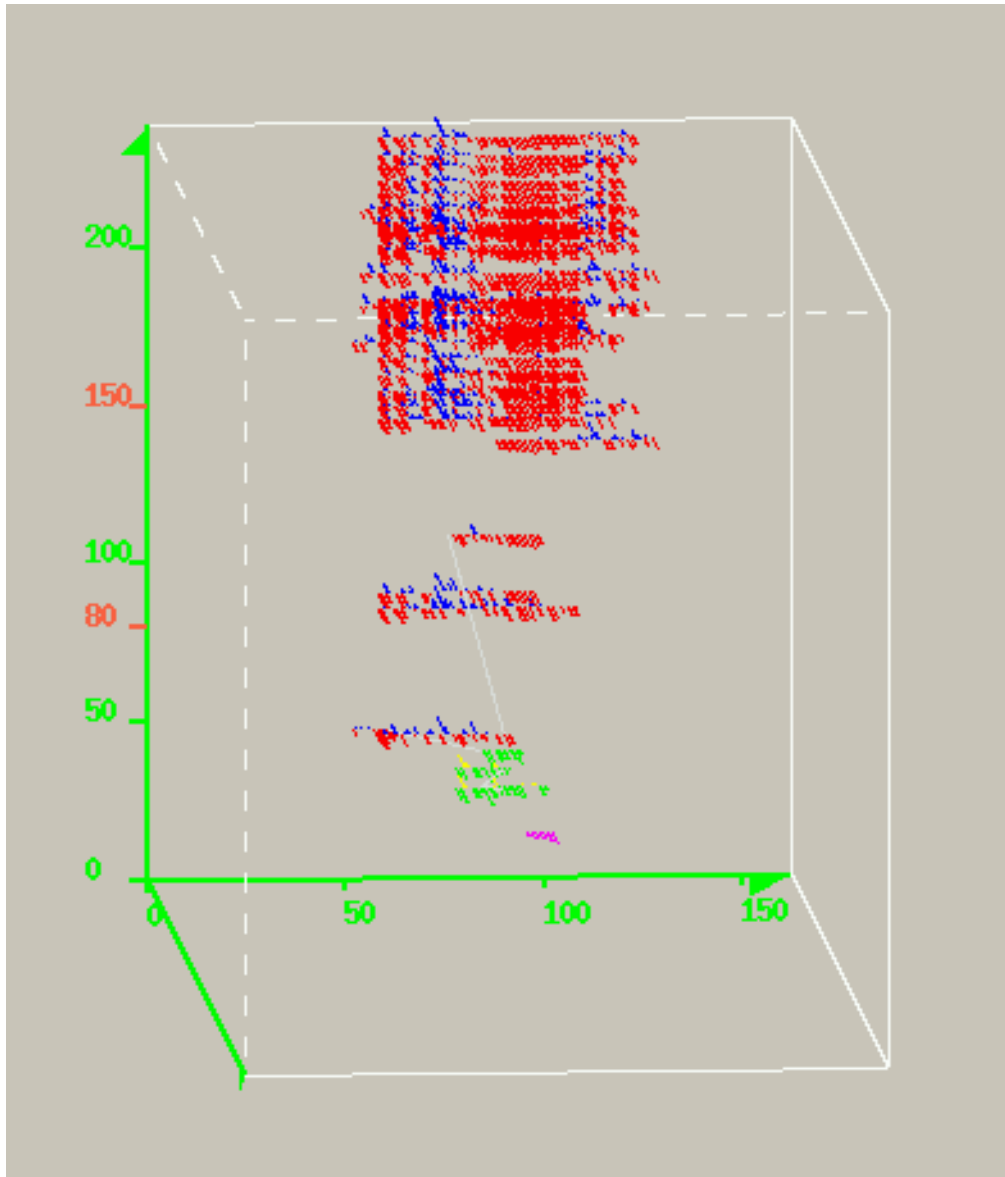


Figure 10: Visual query filters enable biologists to focus the analysis on a group of alignments near the center of sequence 11B11T7P.

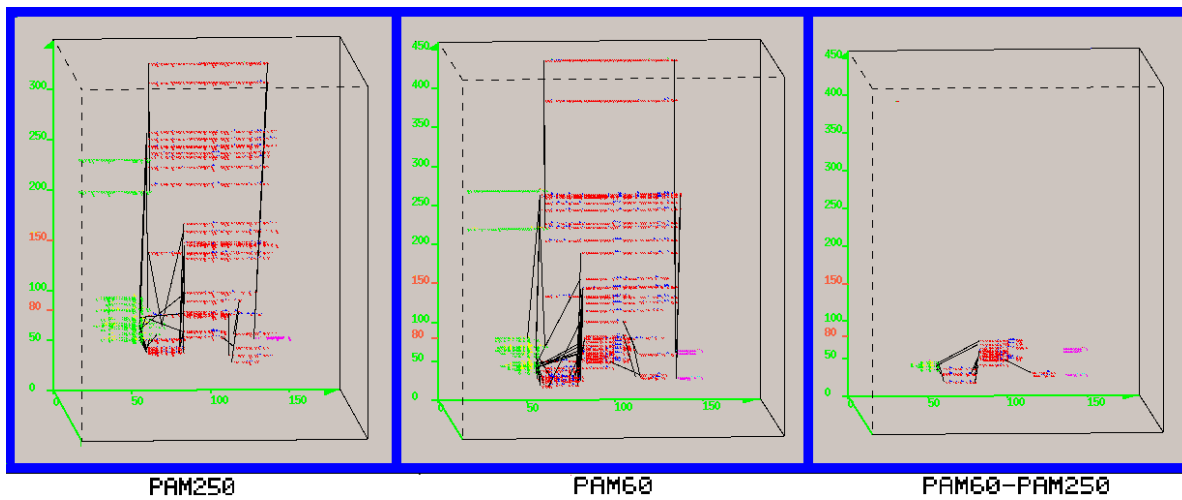


Figure 11: The visual result of subtracting the PAM250 dataset from the PAM60 dataset. Leftmost and center are AV's representations of sequence 172C2T7's similarity report. Rightmost is the result of performing a set difference operation between these two reports.

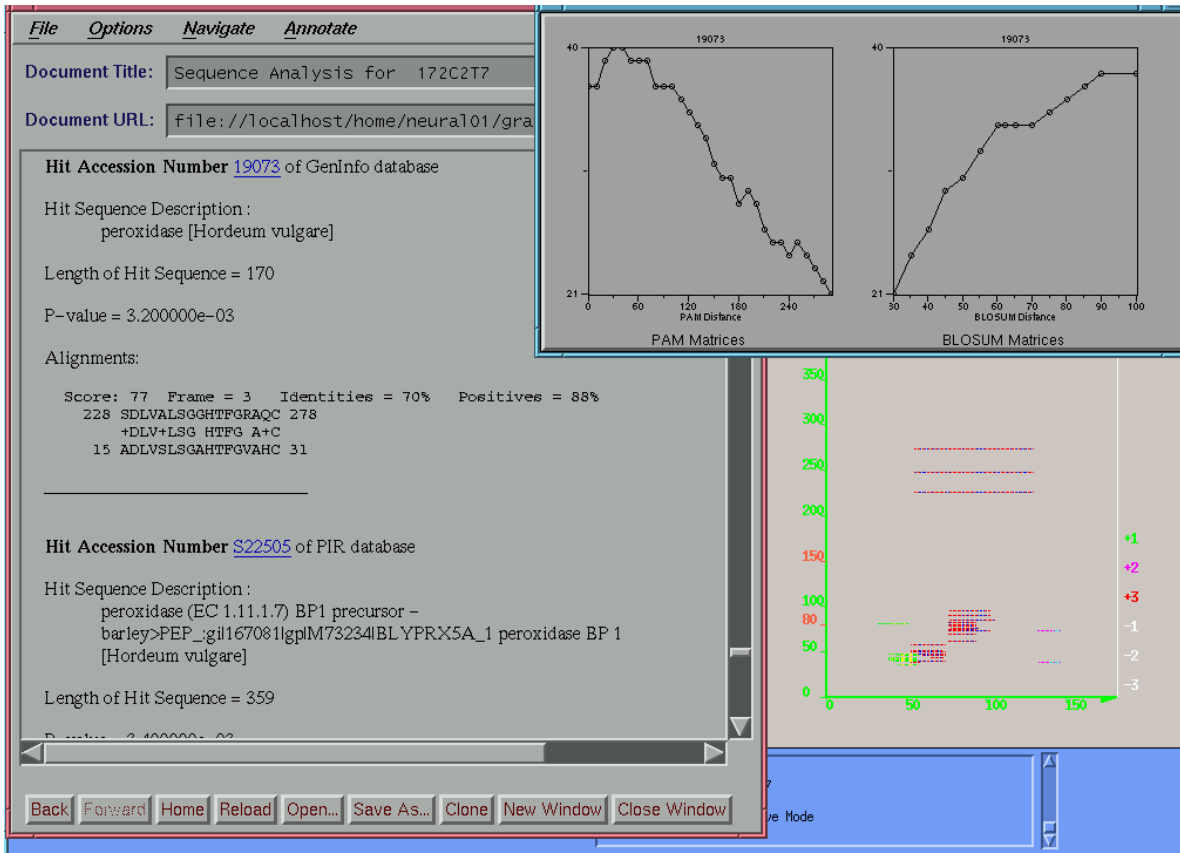


Figure 12: When the user picks an alignment, AlignmentViewer shows the detail of the alignment by telling the HTML browser to jump to the correct alignment in the HTML report.