

Visualization of Biological Sequence Similarity Search Results

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

[†] Computer Science Department,
University of Minnesota
4-192 EE/CSci Building,
Minneapolis, MN 55455
echi@cs.umn.edu

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

Abstract

Biological sequence similarity analysis presents visualization challenges, primarily because of the massive amounts of discrete, multi-dimensional data. Genomic data generated by molecular biologists is analyzed by algorithms that search for similarity to known sequences in large genomic databases. The output from these algorithms can be several thousand pages of text, and is difficult to analyze because of its length and complexity. We developed and implemented a novel graphical representation for sequence similarity search results, which visually reveals features that are difficult to find in textual reports. The method opens new possibilities in the interpretation of this discrete, multi-dimensional data by enabling interactive investigation of the graphical representation.

1 Introduction

Scientific disciplines are confronted with an increasing amount of data and few tools or techniques for extracting meaningful information from it. The questions of what to visualize and how to visualize it make dealing with large, multi-dimensional datasets one of the most important and exciting areas of scientific visualization today. In this paper, we present a novel representation for dealing with this kind of data in the context of molecular biology.

Molecular biologists study the function and evolutionary relationships of genes and proteins in cells. The genetic data is represented as sequential strings of letters, or *sequences*. Biologists use the large amount of known sequence data to help determine the function of new sequences. The similarity search algorithms developed for this purpose currently produce textual output.

Similarity reports for some sequences produce hundreds or thousands of pages of text. The information in these textual reports is proportional to the size of the databases of known sequences, which are growing rapidly. Ironically, as more information about possible function becomes avail-

able, the task of exploring similarity results and determining function becomes increasingly difficult. Visualization methods are needed for biologists to effectively explore the enormous amount of information available.

Similarity analysis information is multi-dimensional, because it contains several orthogonal pieces of information for each similar region found. This discrete, multi-dimensional data lacks a natural visual representation. The work presented here is the result of our efforts to determine a useful representation for the large dataset of similarity results produced for new sequences.

In this paper, we present Alignment Viewer (AV), a novel visualization tool for the large, discrete, and multi-dimensional dataset resulting from sequence similarity analysis. The contributions of this work are:

- We have determined the important features of the multi-dimensional data in this domain and have demonstrated new ways to graphically depict this data. These visualization techniques may be applicable to other domains that contain this type of data.
- AV applies interactive rotation, translation, and zooming of the graphical representation, enabling biologists to explore the data for features that are difficult to detect in the text.
- AV allows biologists to more readily determine differences between multiple similarity reports. This task was nearly impossible previously.
- AV is a successful tool currently in use by molecular biologists.

In the next section, we present some background in computational molecular biology and previous work in visualizing DNA sequences. In section 3 we discuss AV's method for representing sequence analysis data. Section 4

contains case studies of how AV is used in practice, illustrating the features of AV and demonstrating how it is useful to biologists for examining similarity data. Finally, in section 5 we present future work and concluding remarks.

2 Background and Motivation

Molecular biologists seek to determine the genes in the cells of organisms, the function of the proteins that these genes encode, and how these proteins are related evolutionarily across organisms. Genes, composed of DNA, are represented as discrete sequences of nucleic acids, also called bases. Proteins are represented as discrete sequences of amino acids, also called residues. Genes and proteins from different organisms are related through evolution, and share common functions. As molecular biologists discover new genes and the function of the corresponding proteins, their dataset of known information increases. This information is being cataloged in the form of DNA sequence databases for genes, and amino acid sequence databases for proteins.

The advent of DNA automated sequencing has triggered an enormous growth in the public databases. Large-scale sequencing projects now underway on many organisms are producing vast amounts of new DNA sequence data [13, 17]. GenBank, the primary source of DNA sequence data, contains roughly 250,000,000 nucleotides in 270,000 sequences, and is doubling every 1.3 years [4]. The largest protein sequence database is the Protein International Resource (PIR) [7]. PIR contains roughly 12,000,000 residues and 42,000 sequences, and is doubling every 2.4 years.

Traditionally, painfully detailed lab experiments are designed and carried out to determine the function of the proteins. This is still a relatively slow process. One method used to improve protein function determination is to search the databases of known sequences for similarity to an unknown sequence [1, 3, 8]. Similarity algorithms are a well-developed aspect of computational molecular biology research [12, 16], and employ dynamic programming and heuristic search techniques. BLAST [2] and FASTA [14] are the most popular database search algorithms in use today. These algorithms identify similar regions between an input sequence and all sequences in the databases of known DNA and protein sequences. These similar regions are called *alignments*. The results of these similarity searches allow biologists to formulate hypotheses on the possible functions of the query sequence, which are then tested in the wet lab. Identifying possible functions through database searches can dramatically reduce the number of experiments needed.

The difficulty in using similarity searches as a starting

point is that the time required to interpret the large textual reports increases as the size of databases increases. Fuchs remarked that as more sequence data are gathered “data interpretation is likely to become the time-limiting factor in genome analysis [6].”

Previous work in biological sequence visualization concentrated on *single sequence representations*, which are alternatives to the DNA alphabet. The H-Curve, a 3D curve defined iteratively, was suggested by Hamori and Ruskin to represent a long DNA sequence [9]. H. Jeffrey developed another iterative method, the chaos game representation [10]. Wu extrapolated the work of Hamori and Jeffrey and presented a third iterative method called W-Curves [19].

While single sequence representations can find interesting features of individual sequences, such as repetition of nucleotides, 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. Therefore, they do not seem appropriate for finding similarities between sequences. Moreover, while single sequence representations are very good for viewing small amounts of sequence data, they are simply not designed for large datasets.

Instead of devising a new method for representing DNA sequences, we focus on representing the large amount of data from sequence similarity algorithms in a more comprehensible manner. Our motivation is to develop new visualization techniques that enable biologists to discover sequence relationships that are very difficult to determine from textual reports.

3 Alignment Viewer’s Representation

Alignment Viewer (AV) is a result of a collaborative effort between computer scientists and molecular biologists. This collaboration enabled us to determine the most important features of the output of similarity search algorithms. We then designed the representation to enhance these features graphically.

First, AV depicts positional information. Each alignment is between a subsequence of the query sequence and a subsequence of a sequence from the database. The position and length of an alignment identify a region of similarity between two sequences. In the X-axis, AV plots the *absolute position* in the query sequence. An alignment is signified by a comb-like object, such as the ones shown in Figure 1. The beginning, end, and the relative length of each object correspond to the beginning, end, and the length of the alignment.

Second, AV depicts the *frame number* of each alignment. The frame number defines how the DNA sequence

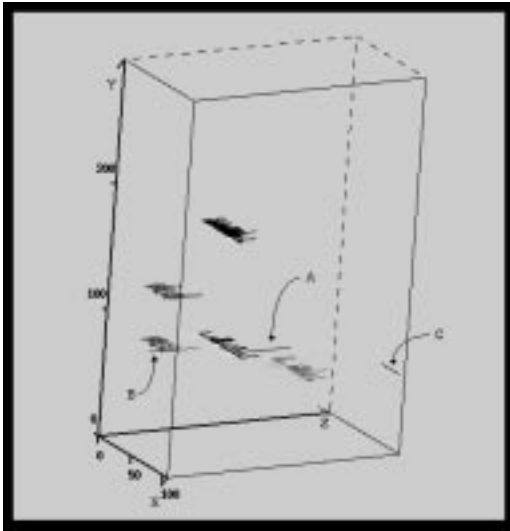


Figure 1: Several alignments represented in AV

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 *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. The starting point determines how bases are grouped into residues. DNA sequences are double stranded, with one strand named positive and the other negative. 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 from the first, second, or third nucleotide of the negative strand. These 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. Biologists do not always know the starting position of the encoding, or which strand a sequence came from. 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 using both colors and layers in the Z-direction. AV puts all +1 frame alignments in the first layer, +2 frame alignments in the second, and +3 frame alignments in the third. The fourth

layer is reserved for all negative frame alignments; these alignments appear only as white lines. The white line labeled C in Figure 1 is a negative alignment in the fourth layer. The user can choose to reverse the negative and positive frames so that the negative frames appear in the first 3 layers and with color, and the positive frames in the 4th layer as white lines. This separation between the positive and the negative frames is because usually the biologist is interested in only positive frame alignments, or only negative frame alignments.

The frame number is also encoded by color. Color actually has a dual purpose, encoding both frame number and residue pair score (explained later). So each frame is coded with two colors. The +1 frame alignments uses red and blue, +2 alignments green and yellow, and +3 alignments magenta and cyan.

We encoded the frame information using both layers and color so frame information is visible from all viewing angles. For example, when viewed down the Z-axis, the layer information is lost if only the Z-coordinate is used to distinguish frame numbers. On the other hand, if only color is used, objects toward the front would mask out objects near the back when viewed down the X-axis. By separating the frames into layers, the amount of information encoded in each layer is reduced.

AV can also assign the colors dynamically. In this mode, the frame with the highest score is assigned red/blue, with green/yellow and magenta/cyan assigned to the 2nd and 3rd highest scoring frames, respectively. Dynamic color assignment is a crude way of guessing the correct reading frame for a given sequence.

Third, AV depicts the *similarity scores* along the Y-axis. The search algorithms produce statistical measures of similarity. These statistical techniques rely on the *substitution matrices* pioneered by Dayhoff et al. in [5]. Each matrix entry measures the likelihood of one amino acid replacing another in a sequence due to genetic mutation and natural selection. As some amino acids are functionally similar to other amino acids, some substitutions are favored biologically more than others.

Similarity algorithms use substitution matrices to compute the *similarity score* of an alignment. For a given pair of residues in an alignment, an algorithm like BLAST looks up the entry in the matrix and gets the *residue pair score*. This residue pair score is then a measure of the strength of the match. To compute the actual score of the alignment, BLAST sums all the residue pair scores in the alignment. The location on the Y-axis of an alignment represented by a comb object corresponds to its similarity score. For example, in Figure 1, the alignment labeled B

has a score of 95.

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).

Each of the residue pair scores is also encoded by the length of the tooth—the stronger the value of the residue pair score, the longer the tooth. For example, the line labeled A in Figure 1 has a residue pair score of 17, and the line labeled B has a score of -4.

We implemented AV on Sun and SGI workstations running X windows. Commercial visualization engines might have served our needs, but using an existing package would still have involved a fair amount of development. We also wanted to integrate AV with our development of a DNA sequence analysis database system [15]. Furthermore, we wanted to distribute our application to molecular biologists, who might not have easy access to commercial visualization packages.

In summary, AV represents the length and position of alignments, the frame numbers, the similarity scores, and the residue pair scores. These elements of the textual report are essential to the data analysis, and demonstrate the multi-dimensional nature of sequence similarity data. By collaborating with biologists working on sequencing projects, we ensure that the above representation is understandable and effective in the analysis of sequences.

4 Case Studies of Alignment Viewer

We now discuss an AV visualization of the *Human Immunodeficiency Virus* (HIV), and a visualization of a sequence from a well-studied plant called *Arabidopsis thaliana*, commonly known as mustard weed. These case studies have been chosen because they are of interest to the molecular biologists in our research group and because they illustrate features of AV.

We took a section of the HIV sequence and ran the BLAST algorithm against GenBank and PIR. The sequence (GenBank Sequence K02012) has 5362 bases, which translates to about 1787 residues. The BLAST textual report is roughly 3200 pages, and contains a total of 6692 alignments to the GenBank database. Each alignment in the text report looks like the one shown in the right hand side of Figure 5, where the text report has been parsed and converted to hypertext by our analysis engine [15]. We filtered out all alignments to copies of HIV, and 1867 alignments remain, which is still roughly 800 pages of printed text. Analyzing this amount of data in a textual report is prohibitive.

Figure 2 is AV output for the same report. The graphical view condenses 800 pages of text into one screen of information. The left hand side is a 3D view, while the right hand side is a 2D projection. The positions and relative lengths of the alignments provide a quick summary of where alignments are located along the query sequence. By rotating the 3D figure, the user can immediately see that there are no negative frame alignments, since there are no white lines in the fourth layer. This example shows that using a single color for all negative frame alignments allows the user to immediately determine whether the query sequence is positive or negative.

The color guide on the right hand side of the picture shows dynamic color assignment. We see a large region of green/yellow +3 frame alignments toward the front of the sequence from about 1 to 500, and a second large region of red/blue +2 frame alignments from about residue 500 to 1450.

Sometimes a single DNA sequence encodes proteins in two slightly overlapping frames. The complicated mechanism that makes this possible is called a *frame shift* [18]. Biologists are interested in frame shifts, especially since they are very important to the function of the sequence and are difficult to discover. In this example, the first green/yellow region seems to encode one protein in one frame, and the second red region seems to encode a different protein in another frame. The color changes suggest the occurrence of a frame shift. Indeed, in HIV, the so-called *gag* protein in the first region is known to overlap with the *pol* protein in the second region [18]. The detection of the frame shift in the text report would require looking through 800 pages of text. AV's representation makes this phenomenon stand out immediately.

In Figure 3, the zoom feature in AV has enlarged the *gag* protein region. The two regions A and B correspond to large concentrations of positive residue pair scores. Remember a positive residue pair score corresponds to a strong match. Biologists are interested in identifying *conserved regions*, which are regions that have been preserved over evolution, and are less likely to change due to mutation and selection. Such a region usually plays an important role in the function of the sequence. The large concentration of positive scores suggest that the two regions A and B are conserved.

We also see in Figure 3 a region labeled C where negative residue pair scores are abundant. These negative scores suggest a region where residues are weakly conserved, since there are more differences between the query and the database sequences. Most of these alignments are to the *gag* protein of the *Simian Immunodeficiency Virus* (SIV), a

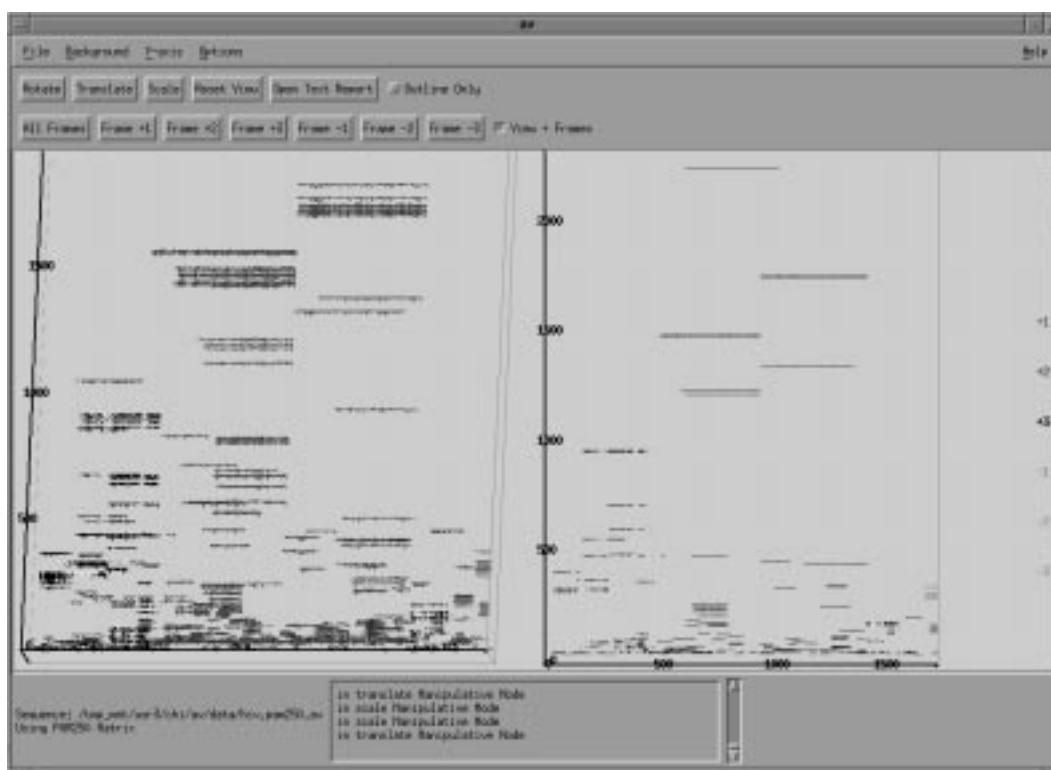


Figure 2: AV visualization of *Human Immunodeficiency Virus* (HIV) BLAST PAM250 output

distant relative of HIV that infects monkeys. Thus, region C is likely to correspond to a segment where HIV and SIV are different biologically. Since SIV cannot infect humans, biologists are very interested in such differences.

The ability to interactively zoom into the relevant region was important in the identification of these conserved regions. Zooming away from the image makes it easy to detect broad similarity features. Zooming in makes details clearer and simplifies the identification of regions A and B as local regions of high similarity to other sequences.

As described above, our representation emphasizes conserved regions. In scanning *Arabidopsis* sequences using AV, our research group found many sequences with possible conserved regions. Figure 4 shows one of these sequences. The two red bands labeled A and B are quite obvious. Region A is apparent from almost every angle, whereas region B is more visible when viewed down the Z-axis as shown. The ability to rotate and view the visualization from different perspectives was crucial to the identification of this region. The alignments in the A and B regions of this report are to a variety of binding proteins and ribonucleoproteins, suggesting that conserved regions A and B might share a common function with these other

database sequences.

AV uses a fat line approach to ensure rapid rotation and zooming. During a mouse-driven user interaction, such as rotation or zooming, each alignment is reduced to a single fat line; when the user releases the mouse button, the full-feature visualization is displayed. The number of lines drawn during a single rotation in the example from Figure 2 is 1857 using fat lines compared to 123,229 for the full image.

Sometimes detailed information about a particular alignment is needed. Biologists must then consult the textual report. AV projects the 3D representation onto the XY-plane as shown in Figure 2. The user can translate, zoom, and rotate both the 3D and 2D representations. The 2D projection was created to allow direct interaction with a single alignment. When a user clicks on an single alignment, the hypertext document browser showing the text report [15] will jump to the correct place and display the detailed information about that alignment. Figure 5 shows this interaction during the analysis of the HIV similarity report. The details of the alignment labeled B is shown in the right half of the figure. The hyperlinks from the graphical output to the textual output ensure that all information is accessible.

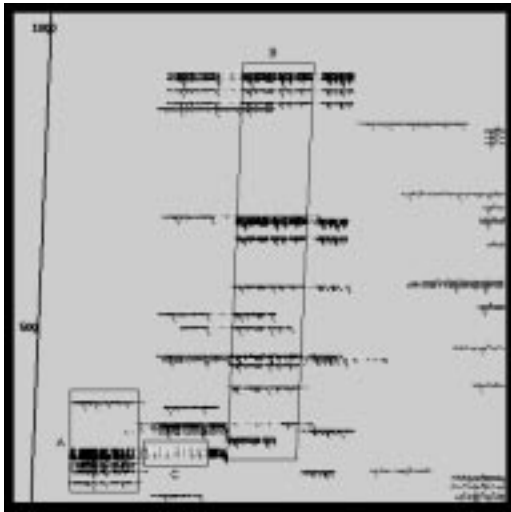


Figure 3: Closeup view of the HIV *gag* protein region

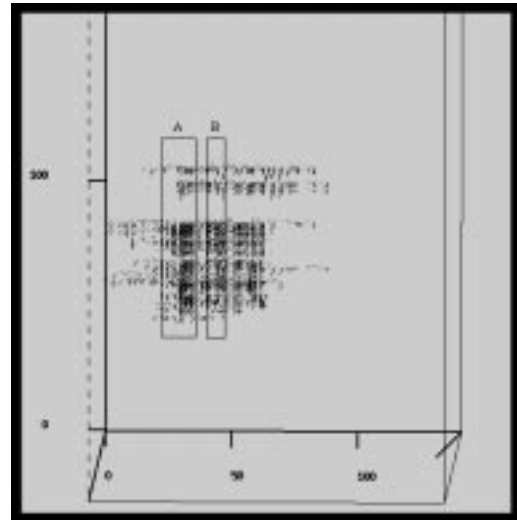


Figure 4: Arabidopsis 128I14T7 AV visualization

Finding all information related to a particular alignment becomes easier because AV provides a visual index to all of the alignments in a report, whereas in the past users had to search through pages of text to find an alignment.

When the user chooses the alignment, a curve plot also appears on the screen as shown in Figure 5. The plot, which we call the *alignment matrix curve*, shows the different similarity scores from different substitution matrices. Different substitution matrices allow different degrees of mismatches and mutations. These matrices are either experimentally or theoretically derived. The *PAM (Point Accepted Mutations)* matrices [5], use a rough measure of how many generations of evolution it would take to mutate one sequence into another. Thus, the PAM scale corresponds roughly with different evolutionary distances, where a low number signifies a close evolutionary distance. For example, the PAM120 matrix allows fewer mutations than PAM250.

The HIV BLAST result in Figure 2 uses PAM250. This means the similarity algorithm was most sensitive to alignments near a PAM distance of 250. Biologists do not always know which matrix is more advantageous to use for a particular sequence.

In the matrix curves of Figure 5, the similarity scores were recomputed and normalized statistically using different matrices [11]. The X-axis on the curve plot is the evolutionary distance measured in PAM, and the Y-axis is the renormalized score computed using a particular PAM matrix. The alignment matrix curve for different alignments peaks at different distances. For example, the alignment

labeled A peaks at PAM60, whereas the alignment labeled B peaks at PAM120. These peaks provide estimates of the evolutionary distances of the alignments. This curve was added to AV to help biologists determine which substitution matrix to use.

Most of the alignment matrix curves from the HIV PAM250 report peaked mostly around 40-120 PAM. Since PAM250 is most sensitive for alignments at 250 PAM, and less sensitive for alignments at 60 PAM, these peaks might suggest that the initial choice of PAM250 is too insensitive for most of the alignments in this report. After looking at these matrix curves and their peaks, we ran BLAST with PAM60. Figure 6 shows the result. There was a noticeable decrease in the number of alignments from 1867 to 1516. The major features were retained, and the amount of low scoring alignments near the bottom of the plot (score \approx 50) decreased significantly.

The PAM60 textual output is also over 800 pages, and the comparison between these two outputs with different matrices is not feasible using only the textual outputs; however, the visualization overviews provided by AV concisely summarize the textual data, making the above comparison possible.

Our research group has a database of over 21,000 sequences and their analysis reports. Instantaneous access to all of the data is as critical in the analysis as the data interpretation. Each hypertext report has a static AV plot that is rotated to a fixed position. The static image then serves as a hyperlinked icon to the actual AV input file. Since AV is integrated with the World Wide Web, any user with AV in-

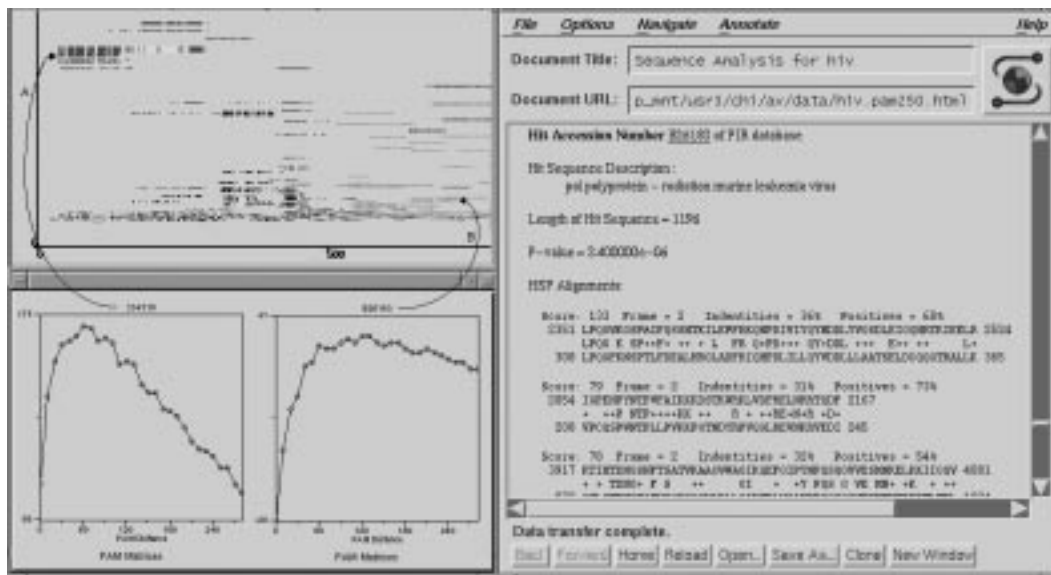


Figure 5: HIV BLAST PAM250 Alignment Viewer visualization (Alignment Matrix Curves and Hyperlinks)



Figure 6: HIV BLAST PAM60 AV visualization

stalled on her workstation can click on the image icon and interact with the AV visualization directly.

In this section, we showed several examples of the use of AV. In particular, we showed how to identify a possible frame shift by looking for a color change. We also showed how to find possible conserved regions by scanning for large concentrations of positive colors that correspond to strong matches. AV allows the user to examine the data closely using zoom, translate, and rotate. These

features also helped in the discovery of conserved regions. The visualization is hyperlinked with the textual report to provide the most detailed information. AV also provides alignment matrix curves to estimate the evolutionary distance of an alignment between two sequences. This additional piece of information can be used to estimate which substitution matrix to use next. These features of AV have proven useful in the analysis of sequence similarity.

5 Conclusion and Future Work

The results of sequence similarity algorithms are large, discrete, and multi-dimensional, which lead to difficult and ineffective analysis of textual reports. Although the positional, compositional, and similarity information are contained in these reports, it is difficult to determine the biological significance of the alignments. For example, it is difficult to determine whether two alignments are in the same region, or whether there are regions where two sequences share similar conserved composition.

AV provides a novel data representation and visualization method for the output of BLAST, one of the most popular alignment algorithms. The graphical representation provides a visual index to the output. AV encodes alignment data by using features such as colors and layers for different frames, and allows interactive zooming, translation, and rotation. The fat line technique provides real-time feedback during interaction. The alignment matrix curves present an estimate of the evolutionary distance of an alignment.

AV visualizes high level information contained in the similarity reports. The advantages of this approach are that AV offers a concise view of the global features and eases the interpretation. The potential disadvantage of omitting the details in the textual reports is mitigated by hyperlinks from the visualization to the text report. AV is used as a visualizer for data contained in these hypertext reports, in the same way external viewers are used to view images on World-Wide Web (WWW) documents. Our research group manages data from several genome sequencing groups, including 15,000 Arabidopsis sequences from Michigan State University. A total of 21,000 AV visualizations can be found in the similarity reports of plant genome sequences located on our WWW site (<http://lenti.med.umn.edu>). The WWW documents provide easier access and distribution of the data to the biology community. AV is in use on a regular basis by biologists in our research group.

We plan future enhancements. The capability to choose different colors interactively is required for different output media. Also, the information encoded by the Y-axis can be other biologically meaningful similarity measures besides similarity score. We also plan to develop techniques for viewing the output of multiple algorithms as well as techniques for viewing several reports simultaneously.

As depicted in the case studies, AV greatly reduces the amount of time required to review and analyze the BLAST output. By visually presenting all the information from an alignment report, AV frees molecular biologists from the drudgery of the text reports, and enables new types of analysis.

References

- [1] M. A. Adams. Sequence identification of 2375 human brain genes. *Nature*, 355:632–634, 1992.
- [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] S. F. Altschul, M. S. Boguski, W. Gish, and J. C. Wooten. Issues in searching molecular sequence databases. *Nature Genetics*, 6:119–129, 1994.
- [4] D. Benson, M. Boguski, D. Lipman, and J. Ostell. GenBank. *Nucleic Acids Research*, 22(17):3441–3444, 1994.
- [5] 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.
- [6] R. Fuchs, P. Rice, and G. N. Cameron. Molecular biological databases — present and future. *Trends in Biotechnology (TIBTECH)*, 10:61–65, 1992.
- [7] D. G. George, W. C. Barker, H.-W. Mewes, F. Pfeiffer, and A. Tsugita. The PIR-International Protein Sequence Database. *Nucleic Acids Research*, 22(17):3569–3573, 1994.
- [8] W. Gish and D. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3:266–272, 1993.
- [9] 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.
- [10] H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.
- [11] 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.
- [12] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.
- [13] M. V. Olson. The human genome project. *Proc Natl Acad Sci U S A*, 90(10):4338–4344, 1993.
- [14] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85:2444–2448, 1988.
- [15] E. Shoop, E. Chi, J. Carlis, P. Bieganski, J. Riedl, N. Dalton, T. Newman, and E. Retzel. Implementation and testing of an automated EST processing and analysis system. In *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, volume 5, pages 52–61. IEEE CS Press, 1995.
- [16] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
- [17] Tom Newman et al. A Summary of Methods for Accessing Results from Large-Scale Partial Sequencing of Anonymous *Arabidopsis* cDNA Clones. *Plant Physiology*, 106:1241–1255, 1994.
- [18] W. Wilson, M. Braddock, S. Adams, P. Rathjen, S. Kingsman, and A. Kingsman. HIV expression strategies: Ribosomal frameshifting is directed by a short sequence in both mammalian and yeast systems. *Cell*, 55(6):1159–1169, 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.