The Visual Comparison of Three Sequences

Kenneth P. Hinckley

Computer Science Dept.
Worcester Polytechnic Inst.
Worcester, MA 01609

Matthew O. Ward

Computer Science Dept.
Worcester Polytechnic Inst.
Worcester, MA 01609

Abstract

Scientists (particularly biologists) currently lack effective tools for comparing multiple sequences of numbers or symbols. This paper describes a method of visual comparison which provides the scientist with a new and unique tool to study the qualitative relationships between three such sequences. The program displays a three-dimensional shape containing the sequence similarities and differences, which manifest themselves as simple geometric shapes and colors that a human observer can easily detect and classify. The method presents all possible correlations to the user, giving it a considerable advantage over existing sequence comparison tools which only search for a programmed subset of all possible correlations. Thus, using this technique, researchers may detect sequence similarities which available analytic methods might completely overlook. The program can also filter out undesirable or insignificant correlations, letting the user focus full attention on the more interesting sequence relationships. The technique enjoys facile adaptation to a wide range of applications, including DNA and protein sequence analysis, speech analysis, signal processing, text comparison, and image analysis.

1 Introduction

The burgeoning needs of fields such as image processing, speech recognition, signal processing, text analysis, and molecular biology demand more effective sequence analysis tools. Molecular biologists are in particularly dire need of such tools. The domain of biological sequence data (including the sequences representing DNA, RNA, and protein macromolecules) has experienced explosive growth in recent years. GenBank, the Genetic Sequence Data Bank at Los Alamos National Laboratory, currently has over 35,000,000 bases catalogued in over 30,000 entries, and this size is currently doubling every 18 months [3, 5].

In sequence analysis one of the major goals is to characterize two or more sequences in terms of their similarities (often termed *homologies*) and their differences (caused by inserting or deleting subsequences or substituting one sequence element with another). It is our belief that the researcher can use the qualitative information presented through visual techniques in combination with the quantitative data produced by traditional methods to discover and investigate hitherto unknown phenomena found in sequence data.

In previous and ongoing research [10] at Worcester Polytechnic Institute, a visual technique for the comparison of two sequences known as the Correlation Image (CI) has been developed. This technique utilizes the basic precepts of Visualization to produce pictorial representations of pairwise sequence relationships. In CIs, each row corresponds to a distinct alignment of the two sequences and each pixel is set according to the relationship between the aligned sequence elements. The purpose of the technique is to simultaneously display all possible alignments between two sequences, thereby producing an image that can easily be scanned by the human eye for significance. Features such as homologies, insertions, deletions, substitutions, transpositions, and repeated subsequences manifest themselves as geometric entities in the image. Subsequent study [5] has shown the CI to be an adept tool for qualitative study of pairwise biological sequence correlations, and a system based on the concept, named *XSauci*, is now finding use at various centers for genetic research. The chief objective of the project described in this paper was to extend the CI concept to the simultaneous analysis of three or more sequences. As a primary goal, we wanted the multiple sequence solution to bear as many similarities as possible to the CI in order to provide upward compatibility for *XSauci* users, to build upon techniques developed for the CI, and to make use of performance analyses of the *XSauci* system.

2 Traditional Methods of Sequence Analysis

Traditionally, researchers have tackled the problem of sequence analysis using dynamic programming techniques [7], which are relatively easy to program and understand. In the comparison of a pair of sequences, the dynamic programming technique generates an M+1 by N+1 scoring matrix, where M and N are the lengths of the sequences to be compared. Any path traversing the matrix represents a possible analysis of the sequences; thus each cell of the matrix represents a possible partial sequence alignment. The path starts in the upper left corner of the matrix and proceeds to the lower right corner, with the understanding that each type of movement (horizontal,

CH3046-0/91/0000/0179/$01.00 © 1991 IEEE

vertical, or diagonal) to an adjacent cell corresponds directly to a type of difference (by insertion, deletion, or substitution) between the sequences. Each type of difference is assigned a weight, which generally depends on the type of sequences being compared. The optimal analysis for the given set of costs can be found by recursively calculating the minimum total cumulative distance associated with each cell of the matrix until the lower right hand corner is reached.

It is possible to extend this technique to the simultaneous analysis of three or more sequences by creating a scoring "cube" (analogous to the scoring matrix) for three sequences, or an n-dimensional scoring lattice for comparing n sequences [1, 2, 6, 8]. Computationally, however, the problem quickly gets out of hand: searching for a minimal path through an n-dimensional lattice leads to an exponential explosion in the number of computations required. Dynamic programming, however, still has the advantage of being a well-known and proven method, so researchers have channeled much effort into finding ways to modify the technique to accommodate larger numbers of sequences.

Most existing sequence analysis methods provide the researcher with a quantitative answer based upon some measure of distance between the sequences. Many algorithms attempt to use distance measures which are based upon models of sequence evolution, but such measures often lack theoretical foundation and are costly to compute, especially when multiple sequences are considered. Moreover, no quantitative sequence comparison method is able to discover sequence relationships which are currently unknown. Finally, traditional measures do not *describe* the relationship between sequences, rather they *quantify* it. Lost are the details of the location, type, and size of specific similarities and differences.

3 The Correlation Image Concept

The Correlation Image (or CI) is a visual technique for the comparison of two sequences [10]. More precisely, a Correlation Image is a matrix containing values which indicate the type and degree of match between pairs of sequence elements. These values can then be translated to colors and intensities for display. Although the CI is essentially a matrix, it is displayed as a parallelogram, since the most important types of similarities manifest themselves as eye-catching horizontal lines when this shape is used.

A Correlation Image is created by placing the sequences on horizontal and vertical axes, so that they define the edges of a matrix. The matrix is then filled with values, where the matrix entry *(x, y)* is determined by comparing element x of the first sequence with element y of the second sequence. Finally, the matrix is sheared to a parallelogram shape and displayed. The intensity of a pixel is set according to the degree of match between elements, with black indicating complete match and white indicating complete mismatch. The horizontally oriented sequence is referred to as the *stationary* sequence and the vertically oriented sequence is labelled the *floating* sequence. This is due to the manner in which CIs can be interpreted. The width of the CI corresponds to the length of the *stationary* sequence, and each successive row corresponds to sliding the *floating* sequence by one sequence element over the *stationary* sequence.

Table 1 summarizes some of the meaningful geometric entities (see Figure 1) which can be found in CI's (This information has been adapted from [10]). Sequence A refers to the stationary sequence and B refers to the floating sequence. Also note that an insertion in A is equivalent to a deletion in B and an expansion in A is the same as a contraction in B. A binary matching decision is used, so only black and white pixels are shown.

4 The Correlation Volume Concept

The fundamental goal of the project was to extend the Correlation Image (CI) concept to three dimensions. The extension, which we term the Correlation Volume (or CV), uses the x-, y-, and z-axes of a coordinate system to represent three sequences, thereby creating a three dimensional object which contains every possible sequence relationship.

The first step in creating a CV is to assign each of the three sequences an axis. Once this is done, the sequences define a three-dimensional matrix, or lattice, which will contain every element of every sequence compared to every element of the other two sequences. Each element *(x, y, z)* of the lattice is then assigned a value, according to the type and degree of match between the three sequence elements at that location; the values can be mapped to colors or intensities for display. The basic lattice shape is then skewed to produce a volume which can be interpreted in a manner similar to Correlation Images.

Once the lattice has been generated, the bottom right front corner of the lattice is "pulled" down (towards the x-z plane) 45 degrees and back (towards the x-y plane) 45 degrees. This transforms the lattice into a Correlation Volume, and the resulting three-dimensional object (a rhomboid prism) is displayed on the monitor (See Figure 3). The term "rhomboid prism" is meant to describe a parallelepiped whose six faces are all rhomboids.

If the lengths of the three sequences are defined as a, b, and c, and if the coordinates of a single entity of the CV are defined as *(cvx, cvy, cvz)*, then the equations which translate the lattice entry *(x, y, z)* to the CV shape are as follows:

$$\begin{aligned} cvx &= x \\ cvy &= (a + c - 1) - x + y - z \\ cvz &= a - x + z \end{aligned}$$

The geometric entities within a CV formed by *pairwise* sequence relationships can be interpreted in exactly the same manner (if viewed properly) as a corresponding CI. This is true because one-correlation thick slices of the CV contain all of the information contained within an analogous CI. For example, vertical slices (one correlation thick) of the CV which are parallel to the y-z plane contain enough information to generate the CI for sequences B and C. By taking

Table 1: Sequence features and their manifestation in a Correlation Image.

Relationship	Representation in CI
equivalent	solid horizontal line across entire CI (Fig. 1a)
substitution	broken horizontal line (Fig. 1b)
insertion in A	downward and diagonally shifted horizontal line (Fig. 1c)
deletion in A	upward shifted horizontal line (Fig. 1d)
transposition	upward then downward shifted horizontal lines (Fig. 1e)
expansion in A	downward diagonal line (Fig. 1f)
contraction in A	vertical line (Fig. 1g)
repetition in A	downward shifted horizontal lines (Fig. 1h)

slices of the CV in other directions, CI's for sequences A and B or sequences A and C can be obtained.

The geometric patterns seen with three-way sequence relationships are arranged in three dimensions, not just two. This means that no single view of a CV is sufficient to identify the sequence relationships present. The views where geometric entities in CV's most resemble those seen in CI's are the bottom (B), the front-tilted (FT), and the front-left-tilted (FLT) views. The front-tilted view is a front view of the CV rotated 45 degrees around the z axis, while the front-left-tilted view is a front view of the CV rotated -45 degrees (to the left, or clockwise) around the y axis and 45 degrees around the z axis.

Table 2 summarizes the geometric patterns which manifest themselves in the CV for different types of sequence relationships. Sequence relationships are stated in terms of changes to sequence A, with the understanding that similar changes in the other sequences could be detected as similar patterns from different views. A complete list (with images) can be found in [4].

To facilitate the viewing of the CV, we implemented the following interactive capabilities:

View Selection. The program provides commonly used views in a pull-down menu. The user has the capability to specify arbitrary views, since it is useful to occasionally fine-tune the default views, and because arbitrary viewing helps to reveal exact spatial arrangements of three-way correlations when the image is inspected from unusual viewpoints.

Exposing Buried Correlations. In any sizable CV, correlations buried deep within the volume will not be visible from any viewing angle. A subsequence viewing capability was implemented to allow the user to work with portions of each of the three sequences (that is, portions of the CV volume), and thereby circumvent the problem.

Filters. Ideally, filtering removes any insignificant "noise" while retaining all information of interest to the user. The use of filtering entails a certain risk, however, since filtering algorithms may mistakenly remove important sequence relationships along with the "noise". Nonetheless, filtering does offer some advantages, especially in sequences where elements are drawn from small alphabets (such as DNA and RNA sequences), since the CV's produced by such sequences tend to contain many incidental, isolated correlations which can be safely removed. As a compromise, a filtering capability was included with the CV implementation, but with a warning: filtering may jettison valuable information along with the "useless" clutter. The program currently only filters three-way correlations, and furthermore, it only recognizes groups of correlations which represent matching subsequences; thus, sequence relationships such as expansions, compressions, and texture patterns may be obscured. Filtering is performed by eliminating points belonging to diagonal lines shorter than a user-specified tolerance.

Types of Correlations Displayed. During typical usage, the user only wants to view a limited subset of all possible correlations within a CV. All useful combinations of correlations can be selected from a pull-down menu.

Projections. The program allows the user to generate views of the CV using both perspective and parallel projections. Displaying the CV using a perspective projection adds a sense of depth to the image, while the parallel projection helps the user determine the precise geometric relationships within an image.

Color Use. A final feature allows the user to modify colors used for different types of correlations. This feature is typically only used when generating CI's from a CV. The various color rules are menu-selectable.

5 Examples

To demonstrate some of the features of the program, we performed a simple evolutionary analysis of 5S ribosomal RNA sequences from various organisms. The study, shown in Figures 7a-f, allows one to easily infer the evolutionary distance between human 5s rRNA and that of the other organisms studied (rat, iguana, chicken, silkworm, fruit fly, and rye). The study indicates that humans, rats, and iguanas are all at about the same evolutionary "step," with chickens trailing slightly behind, and furthermore, that humans have some similarity to silkworms and fruit flies, but have almost nothing in common with rye.

Some less obvious conclusions might also be derived from Figure 7. One could hypothesize from Figs. 7a-c that 5s rRNA only rarely mutates in the course of evolution, since human, rat, chicken, and iguana 5s rRNA are all nearly identical. Even a comparison of

Table 2: Sequence features and their manifestation in a Correlation Volume.

Relationship	View	Representation in CV
equivalent	all	horizontal line across entire image (Fig. 2a)
substitution	all	broken horizontal line (Fig. 2b)
insertion in A	FT	broken horizontal line (Fig. 3a)
insertion in A	B	broken horizontal line with non-overlapping vertical shifts (Fig. 3b)
insertion in A	FLT	broken horizontal line with overlapping vertical shifts (Fig. 3c)
transposition in A	FT	solid horizontal line (Fig. 4a)
transposition in A	B	similar to CI (Fig. 4b)
transposition in A	FLT	similar to CI, except some vertical overlap (Fig. 4c)
repetition in A	FT	broken horizontal line (Fig. 5a)
repetition in A	B	diagonally shifted, non-overlapping horizontal lines (Fig. 5b)
repetition in A	FLT	diagonally shifted, overlapping horizontal lines (Fig. 5c)
expansion in A	FT	diagonally stacked horizontal lines (Fig. 6a)
expansion in A	B	solid diagonal lines stacked vertically (Fig. 6b)
expansion in A	FLT	solid diagonal lines aligned horizontally (Fig. 6c)

vastly differing organisms, such as the comparison of human, rye, and fruit fly in figure 7f, yields some interesting results, since the regions of similarity which appear (despite the immense evolutionary distance of the organisms being compared) might indicate "core" or fundamental segments of the 5S rRNA which perform basic biological functions necessary in all viable organisms.

6 Conclusions

There has been a rapidly expanding need for sequence analysis tools in recent years. There have been many attempts to develop quantitative algorithms to deal with this problem, but they suffer a common flaw: they cannot discover unsuspected sequence relationships, since they have not been programmed to search for such relationships. The human visual perception system, on the other hand, can detect unusual or unexpected patterns in images. While Correlation Images perform admirably when comparing pairs of sequences, they cannot deal with multiple sequences. This project was initiated to research the possibility of extending the Correlation Image to multiple sequences. It has resulted in the development of the concept of a Correlation Volume, a three dimensional shape which contains all possible sequence relationships between three sequences. The subsequent implementation of the Correlation Volume concept revealed that the geometric patterns produced by three-way sequence relationships bear many similarities to the patterns seen in Correlation Images. The program, and the initial results obtained from it, demonstrate the validity of the Correlation Volume technique, its readiness for use by those involved in analyzing sequence data, and the need for further exploration of the technique's capabilities.

There is much potential for additional research on Correlation Volumes. Some avenues are based on improving the flexibility of viewing. One idea is to add the capability to view diagonal slices of the CV; this feature would make the search for three-way homologies buried within the volume much simpler. The addition of an "observer pod" which the user could "fly" into the volume to seek out interesting homologies might also prove useful. Another idea is to draw the user's attention to significant homologies by rendering them as surfaces. Finally, a future study of the CV program might look to add quantitative data (i.e., results produced by dynamic programming or other quantitative sequence alignment algorithms) to the qualitative data presented, thereby making the CV program an integrated tool for the study of multiple sequences. Beyond the three-dimensional Correlation Volume, there is the possibility of an n-dimensional Correlation Volume; it may be the case that the general principles of Correlation Images and Correlation Volumes can be extended to even more sequences.

References

[1] Altschul, S. F., Lipman, D. J., "Trees, Stars, and Multiple Biological Sequence Alignment", *SIAM J. Appl. Math*, 49 (1989), pp. 197-209.

[2] Carrillo, H., Lipman, D., "The Multiple Sequence Alignment Problem in Biology", *SIAM J. Appl. Math*, 48 (1988), pp. 1073- 1082.

[3] Core, N. G., Edmiston, E. W., Saltz, J. H., Smith, R. M., "Supercomputers and Biological Sequence Comparison Algorithms", *Computers and Biomedical Research*, 22 (1989), pp. 497-515.

[4] Hinckley, Kenneth P., "The Visual Comparison of Three Sequences", *Worcester Polytechnic Institute Major Qualifying Project Report*, May 1991.

[5] Hopkins, Reginald D., "Nucleic Acid Sequence Analysis Using Correlation Images", *Worcester Polytechnic Institute Major Qualifying Project Report*, May 1990.

[6] Lipman, D. J., Altschul, S. F., Kececiouglu, J. D., "A Tool for Multiple Sequence Alignment", *Proc. Natl. Acad. Sci.*, USA, 86 (1989), pp. 4412-4415.

[7] Sankoff, D., Kruskal, J. B., Eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, Addison-Wesley, New York, 1983.

[8] Smith, H. O., Annau, T. M., Chandrasegaran, S., "Finding Sequence Motifs in Groups of Functionally Related Proteins", *Proc. Natl. Acad. Sci.*, USA, 87 (1990), pp. 826-830.

[9] Tufte, E. R., *Envisioning Information*, The Graphics Press, Cheshire, CT, 1990.

[10] Ward, M. O., Adams, D. S., "Nucleotide Sequence Analysis Using Correlation Images", *Proceedings of the First IEEE Conference on Visualization in Biomedical Computing*, 1990, pp. 49-56.

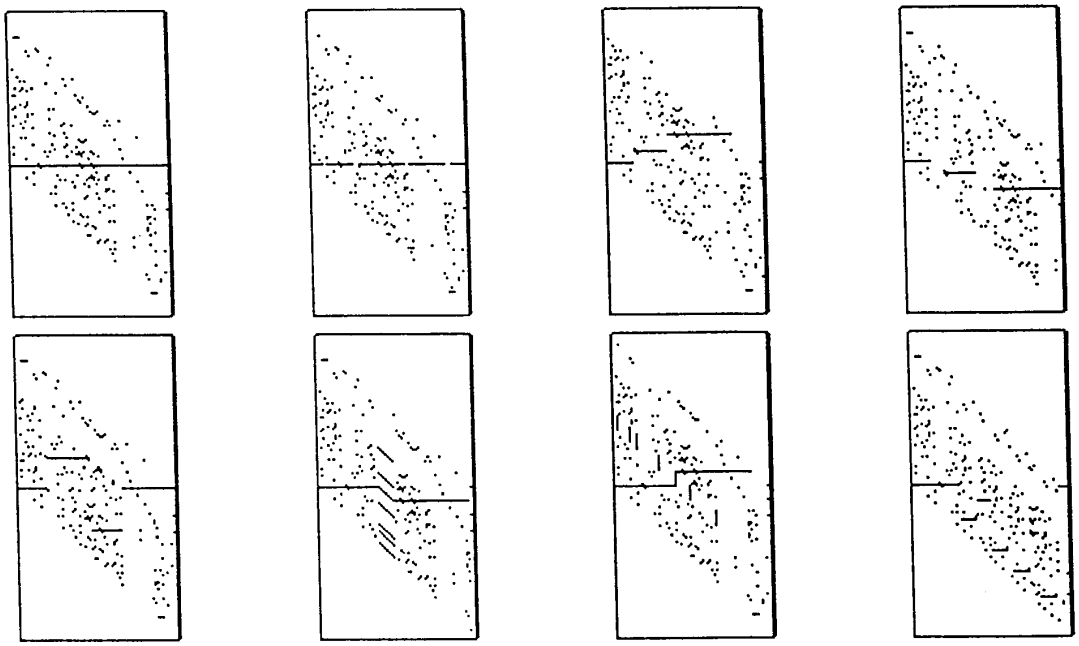


Figure 1: Geometric patterns found in Correlation Images: equivalence (a), substitution (b), insertion (c), deletion (d), transposition (e), expansion (f), contraction (g), repetition (h).

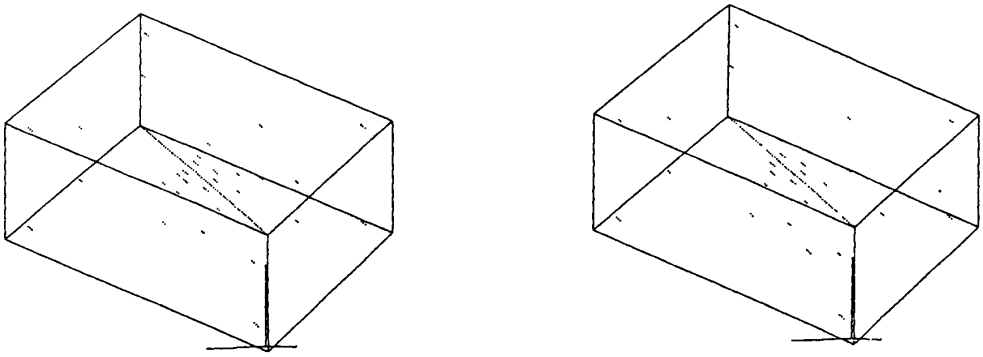


Figure 2: View of CV for equivalent sequences and sequences with substitutions. These figures show perspective projections of CVs using a filter length of 1, an x rotation of 337.5 degrees, and a y rotation of 45 degrees.

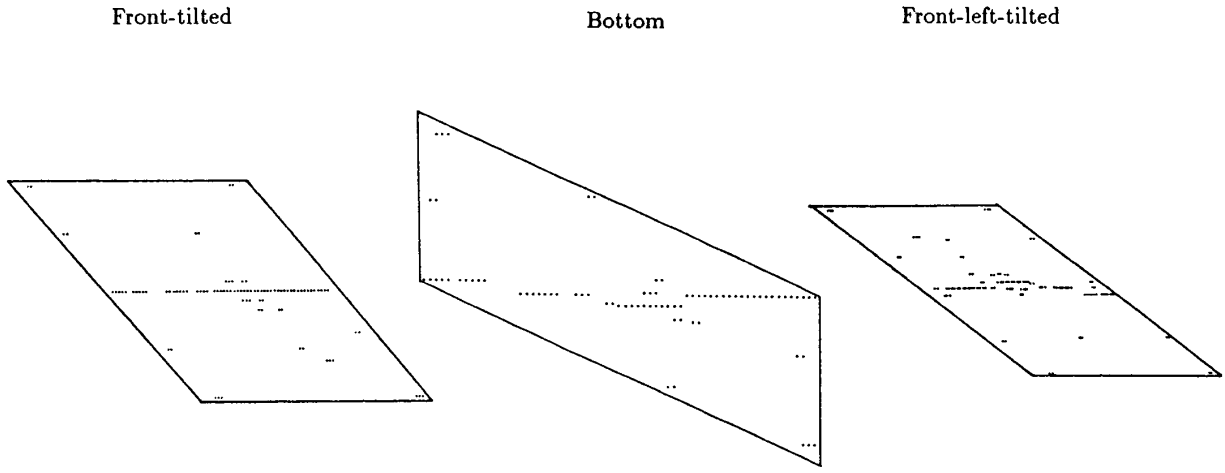


Figure 3: Three views of CV with insertions and deletions in one of the sequences using a parallel projection and a filter length of 1.

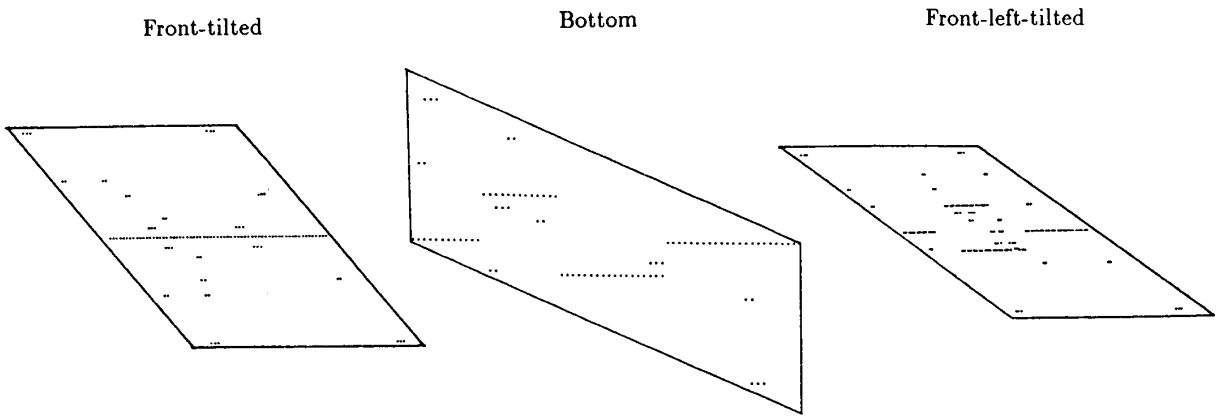


Figure 4: Three views of CV with transposed subsequences in one of the sequences using a parallel projection and a filter length of 1.

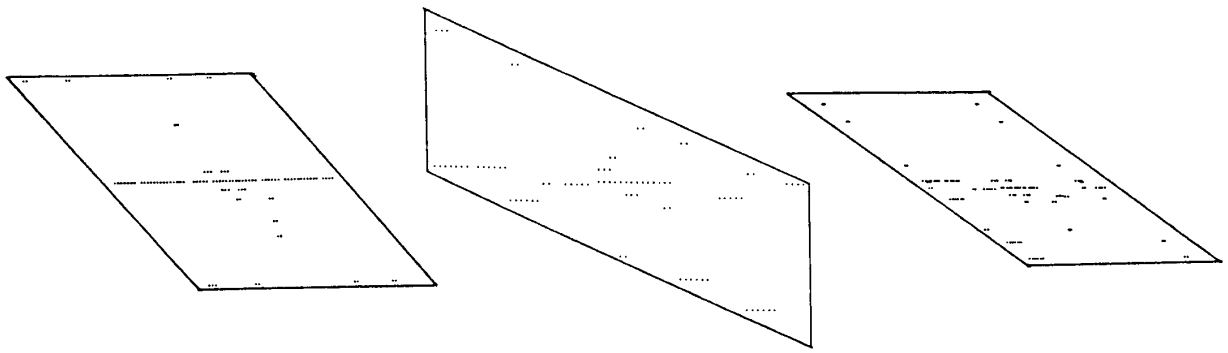


Figure 5: Three views of CV with repeated subsequences in one of the sequences using a parallel projection and a filter length of 1.

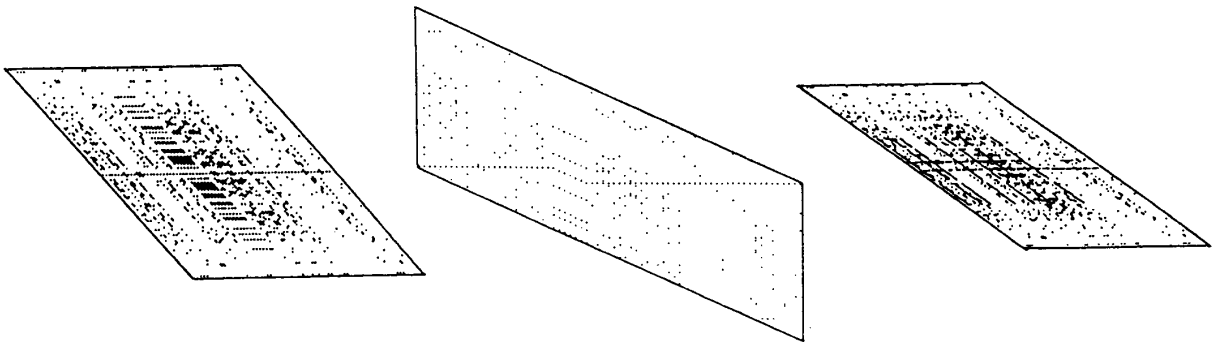
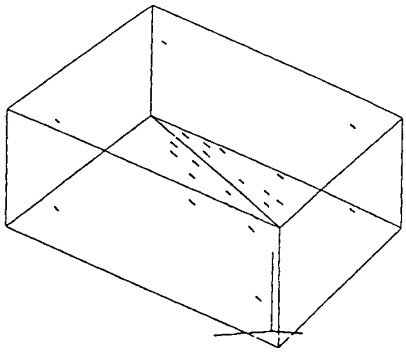
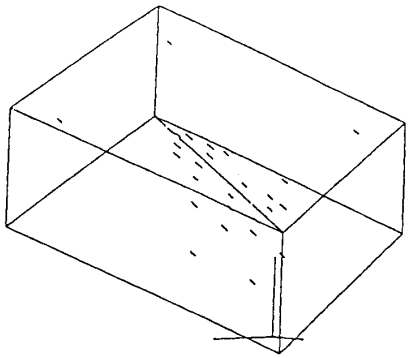


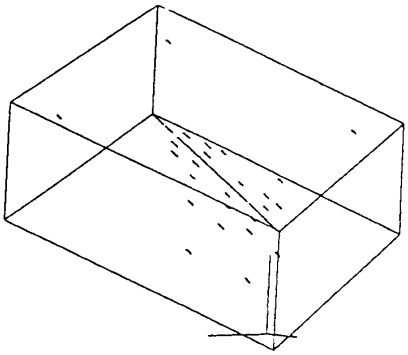
Figure 6: Three views of CV with an expansion in one of the sequences using a parallel projection and a filter length of 1.



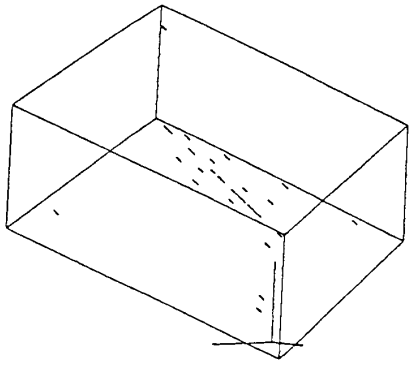
Human - Iguana - Rat



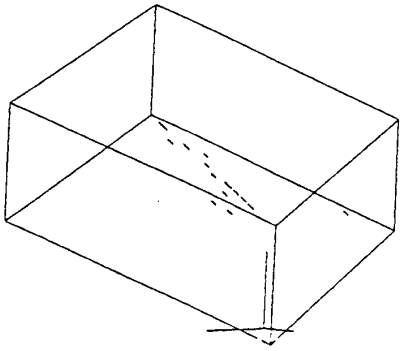
Human - Chicken - Rat



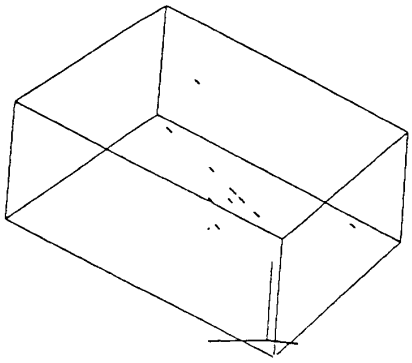
Human - Chicken - Iguana



Human - Silkworm - Iguana



Human - Silkworm - Fruit Fly



Human - Rye - Fruit Fly

Figure 7: Evolutionary study of 5S ribosomal RNA sequences. The GenBank Locus and Accession number for each of the sequences are as follows: Rat(Loc:RATRRA, Acc#:K01594); Iquana (Loc:IGURRAA, Acc#:M10817); Chicken (Loc:CHKRRA, Acc#:X01309); Silkworm (Loc:BMORRAX. Acc#:K03316); Fruit Fly (Loc:DRORRA, Acc#:J01854); and Rye (Loc:RYERRA, Acc#:M10818). All images were generated with a filter length of 4.