An Interactive Visualisation for Investigating DNA Sequence Information

Paul Rutherford, Clare Churcher and John McCallum[†]

Applied Computing Mathematics and Statistics Group
P.O. Box 84, Lincoln University
Canterbury, New Zealand

Rutherp3/Churcher@Lincoln.ac.nz

[†] Crop and Food Research
Private Bag 4704, Christchurch
New Zealand

McCallumJ@crop.cri.nz

Abstract

[1]The sequence of nucleotides that makes up a DNA molecule encodes the characteristics of living things. Bioinformatics offers sophisticated methods to search for and compare nucleotide and protein sequences. Having found a sequence, however, it is useful to be able to quickly obtain an overview of its features. This can be difficult in the standard textual format used to represent sequences. In this paper we offer a simple facility whereby a scientist can interactively scan for recurring patterns in the sequence and investigate reading frames, variable regions, positions of particular codons and other features. It is intended that this be incorporated as a tool or component into public domain bioinformatics applications. The visualisations are useful for sequences up to 2000 base pairs.

Keywords: Information Visualisation, DNA sequences.

1 Introduction

A DNA molecule is made up of a long series of units called nucleotides. Each nucleotide has a variable component called a base. The base used comes in four forms, the chemicals: adenine, cytosine, guanine and thymine. For brevity, these are usually described as A, C, G, or T. As the base is the variable component of the nucleotide, it is used to describe the sequence.

For production of protein structures, DNA sequences are translated into amino acid sequences. During translation the DNA sequence is linearly grouped into sets of three nucleotides called codons. One codon encodes for one amino acid. There is one codon that signals the start of translation. Following this, the amino acid represented by a codon is attached to a growing amino acid sequence. Translation continues until a stop signal is encountered. The amino acid sequence makes up protein structures.

There are twenty amino acids. Some of these amino acids are encoded by more than one codon. Usually these will have the same first two bases and the third will vary. This is called a 'wobbly base'. Wobbly bases can be variable as they may change without affecting the final product.

So a DNA sequence has features that include its translated sequence, the percentage composition of the nucleotides G and C within it (GC content), and location of translation signals.

A DNA sequence is usually presented to a user as a series of characters that represent the bases making up the molecule (Figure 1).

```
   1 cgatttcatg gcttgtatgg gcgctttctt tggcgtggtt
  61 gctcttgatt gctgttgcca tatctatagc taggattctc
 121 aacagcgctt cttggtaacc ttccaagaac taaattgtat
```

Figure 1: Textual representation of a DNA sequence

Clearly this representation does not allow the reader to readily find features and patterns in the sequence. Attempts have been made at representing this information for protein sequences (Combet *et al*, 2000, Neshich *et al*, 2003, Hubbard *et al*, 2002).

There have been several applications of colour to amino acid sequences. Generally these visualisations will display the textual sequence listing with characters coloured according to some colour mapping. There is very little control and no dynamic interaction with them. ColorSeq, of the NPS@ project (Combet et al, 2000) is one such application. It allows the user to specify an amino acid sequence, a set of amino acids (letters) to be coloured, and an 'output-width'. It then displays the amino acid sequence truncated to lines of the specified length with selected letters coloured red, and unselected letters coloured black. This displays only one feature at a time.

STING (Neshich et al, 2003) gives more control for using colour in sequence listings and three-dimensional models of the protein sequences. Several colour mappings are provided. However, this program is more protein structure focussed. It is better suited for working with

known protein coding sequences unlike those intended for use with our visualisation, which are not yet understood. Features of the sequence, such as start and stop codons, are not displayed.

Ensembl (Hubbard et al, 2001) is a sequence database project that includes a visualisation for showing features of a sequence. It is intended for dealing with complete genomic sequences, so is designed for considerably larger sequences than we will work with. It shows a ruler across the display, representing the sequence, and below it are annotations to the sequence. Features are annotated using rectangles marking the region they occupy. Only one feature-type is displayed per row. This means to search for one feature-type you apply your attention to one row. Overall, it does well to provide a very condensed view of the sequence. It will, however, only be useful to display features, not discover them. Similar is the representation in GESTALT (Glusman, 2000), except that it makes some use of colour to distinguish types of features as well, so several may be combined in one row. It is still not sufficient for feature discovery though.

CINEMA is another interactive DNA and protein sequence display (Parry-Smith *et al*, 1997). It allows visualisation and manipulation of sequences. CINEMA's purpose, however, is to develop alignments (finding regions of similarity) between sequences so will not help discover patterns. Nor will it display codon features.

These existing visualisations are only interactive in that the linear position in the sequence may be moved, but no folding or rearrangement of the sequence may be performed to discover other features such as patterns in the sequence.

We offer a facility that enables the user to interact with the display in order to facilitate the discovery of features and patterns in sequences of up to 2000 base pairs.

Our visualisations display overviews of base and amino acid sequences in an interactive framework that allows the user to zoom in on regions of interest and adapt the display to highlight different features. It also allows a user to explore different reading frames for a DNA sequence by displaying start and stop codons in each of the frames. The user can also interact with the display in order to discover patterns in either a nucleotide or amino acid sequence. Variable regions caused by wobbly bases are also indicated.

The visualisations were created using VTK (Schroeder *et al*, 1998), using a Tcl/TK interface (Ousterhout, 1994). The application currently reads sequence data in FASTA format (Pearson and Lipman, 1988).

In section 2 we describe how we represent a DNA sequence to see GC content. We discuss how enabling interaction with the display allows the user to quickly discover any patterns in the sequence. In section 3 we discuss variable regions of a DNA sequence. Section 4 looks at amino acid sequences. It explains how the user can investigate reading frames by seeing the location of start and stop codons. It also discusses ways to represent the different chemical nature of the amino acids. Finally in Section 5 we discuss how the ability of the user to interact with the visualisation allows exploration and the discovery of possibly overlooked areas of interest.

2 Representing a nucleotide sequence

The fours bases that make up a nucleotide sequence are represented by the letters ACGT. As shown in Figure 1, these are typically grouped in sets of ten to assist in locating a particular part of the sequence. An obvious way to represent the different nucleotides is by using colour[1]. Figure 2 shows the DNA segment in Figure 1 using coloured rectangles to represent the different nucleotides. One can clearly see the poly-A tail at the end of the sequence.

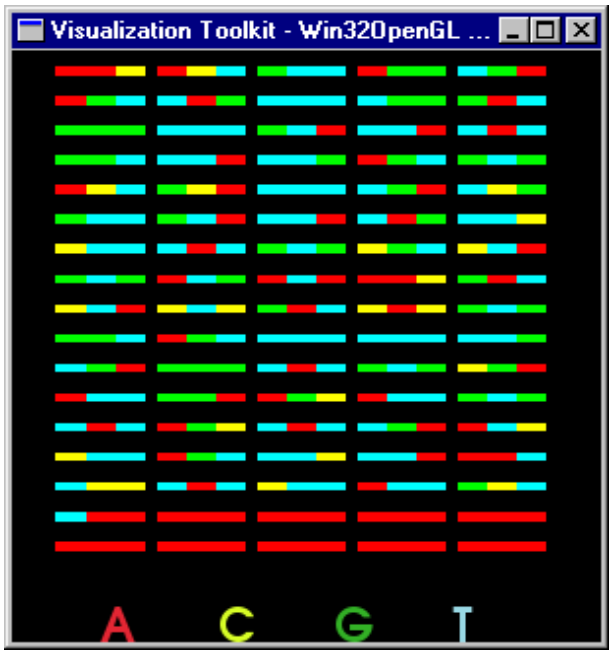


Figure 2: Colour representation of a DNA sequence

The height and width of the rectangles can be adjusted to allow the user to more conveniently gain an overview of different sized sequences. In addition it is possible to zoom in or pan to areas of interest for closer inspection.

One statistic of interest for biologists is the GC content (the ratio of G+C bases to A+T). Much of a DNA sequence is not used for coding and these areas are more likely to have a higher GC content. By altering the colouring of the nucleotides so that C and G are represented by the same (or similar) colours and similarly for A and T we are able to more easily see those areas which are likely to be coding regions. Figure 3 shows the same sequence as Figure 2 above in the new colouring scheme and we can see that the part of the sequence at the top of the display has a higher GC content than that at the bottom.

[1] If you are viewing a printed copy of this document you will not fully appreciate this. See this paper on-line at http://www.invis.canterbury.ac.nz/

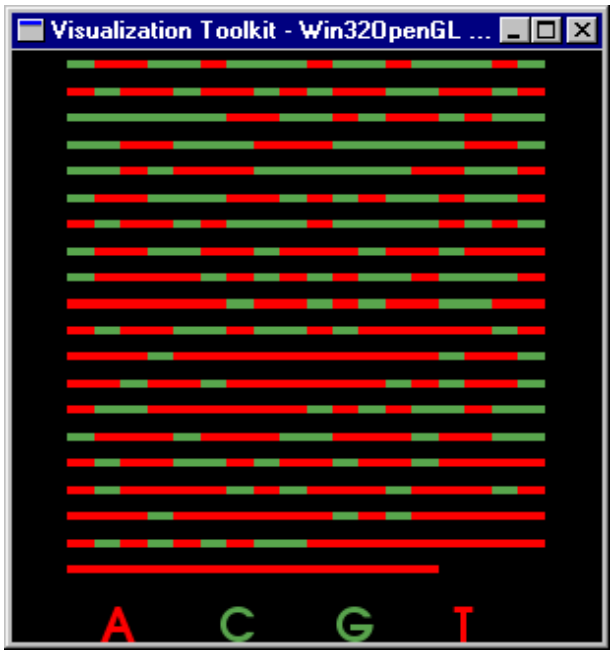


Figure 3: Representing: Representing AT/GC content of a DNA sequence

2.1 Finding codon patterns in a DNA sequence

Figures 2 and 3 display the nucleotides as a continuous sequence running from left to right. However it is groups of three of bases that are of interest to biologists. Each group of three bases, or codon, translates to an amino acid. For example CGA translates to the amino acid Arginine. Having established a reading frame (see Section 4.1) it is useful to display the sequence broken up into its codons (Figure 4).

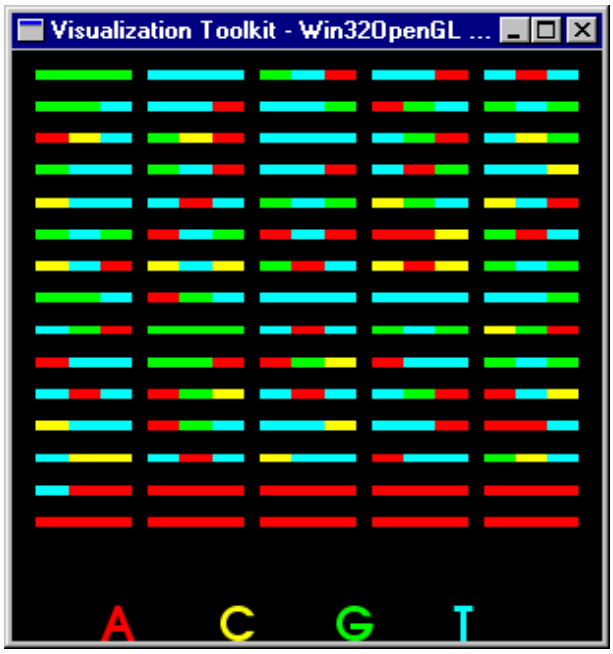
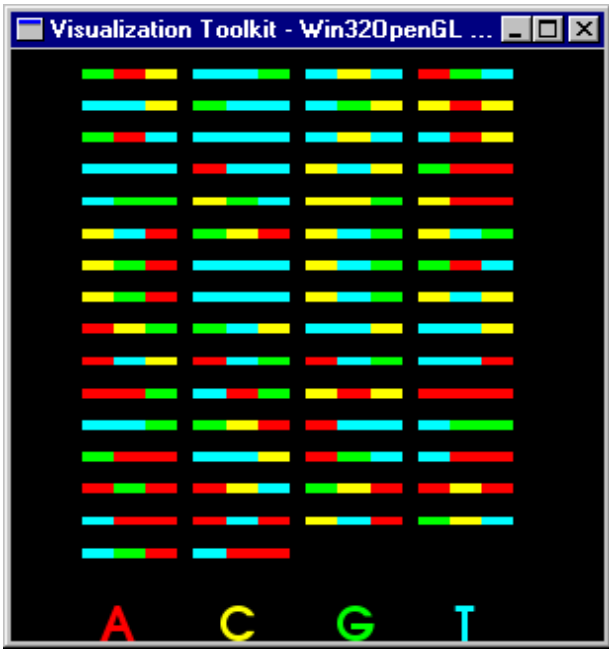


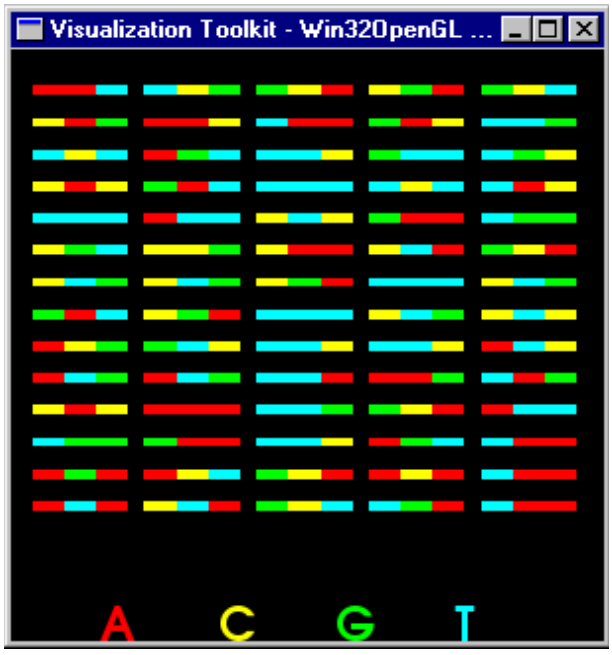
Figure 4: Separating the codons of a DNA sequence

Repetitive sequences at nucleotide and amino acid level are of functional and practical interest to the biologist. By adjusting the number of nucleotides in a row, patterns with different lengths of repeat can become apparent. In Figure 5 we see how adjusting the number of nucleotides in each line reveals different patterns. In the third column of Figure 5a (rows 6,7,8) we see a codon repeated every fourth codon. When the width of the display is changed so these no longer fall under each other (Figure 5b), the pattern disappears. However this width reveals a different repeat every fifth codon at the bottom right of the display.

By providing interactive adjustment of the width of the display we enable the user to quickly scan for patterns that might repeat at a different scales.



5a



5b

Figure 5: Discovering repeat patterns in a DNA sequence

3 Variable regions of a DNA sequence

As well as seeing patterns of repeating codons there are other features of a nucleotide sequence that are of interest to biologists. These include variable regions and the position of start and stop codons.

For production of protein structures, DNA is transcribed into RNA. For our purposes the RNA can be considered essentially the same as the DNA except that the base Thymine (T) is replaced by Uracil (U). Each triplet of bases codes a particular amino acid, however there is some degeneracy in the coding. For example the amino acid Valine is coded for by GUU, GUC, GUA and GUG. With G and U as the first two bases the third base does not make any difference to the amino acid produced. This is the case with the third base in many amino acids and is sometimes referred to as a "wobbly" base.

Variable regions are of interest when considering mutations and they also can be a means of identifying individuals. On the other hand some laboratory procedures require specific sequences so are intolerant to the presence of a variable region.

Being able to identify variable regions in a sequence is therefore of interest to scientists. Our visualisation can be adapted to show these regions by using a different shape to represent a "wobbly" base. This is shown in Figure 6 where wobbly bases are represented by a circle.

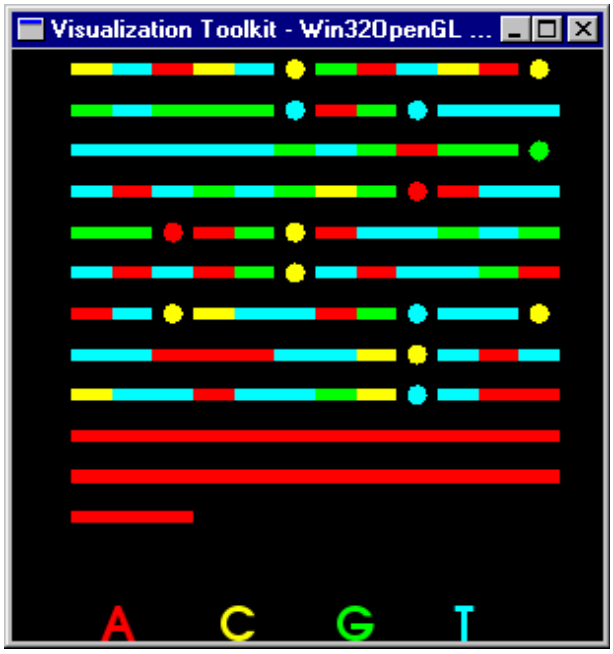


Figure 6: Representing variable regions

4 Amino Acid Sequences

As well as investigating the DNA sequence and its individual bases it is also useful to see the result of the translation of triplets or codons into their respective amino acids. The amino acids can be represented similarly to the representation of bases in Figure 2. Each symbol now represents an amino acid or a specific combination of 3 bases. The display is consequently compressed to a third its original size.

4.1 Reading Frames

Before we can see the amino acids we need to establish a reading frame for the DNA sequence. The reading frame will not necessarily start at the beginning of a sequence. There are, for our purposes, three possible ways to construct the sets of triplets depending on where we start. Only one of these will be the correct frame for providing coding information.

Some triplets of bases represent special codons (start and stop) which indicate the beginning and end of a region that contains information. The codons AAA, AGA and GAG always represent a stop codon. The codon AUG can represent the start of a coding region in some circumstances but can also code for the amino acid Methionine. (Purves *et al*, 1995)

The correct reading frame will therefore likely contain a lengthy region of amino acids unbroken by stop codons and with a start and stop codon at either end. We use shapes to distinguish the start and stop codons as shown in Table 1.

Featureless amino acid	
Possible start codon	
Stop codon	
Amino acid containing a wobbly base	

Table 1: Representing types of amino acid

The shape for a stop codon is distinctive and stands out clearly as can be seen in Figure 7. Because the start codon can translate to an ordinary amino acid in many circumstances its symbol is less obvious but able to be located.

The user can now interactively switch between different reading frames looking for one with a region lacking stop codons. Having found one it is then possible to look more closely to see if there is a suitably placed start codon.

Figure 7a shows a likely reading frame with an uninterrupted region at the beginning while 7b shows an unlikely frame with stop codons scattered throughout.

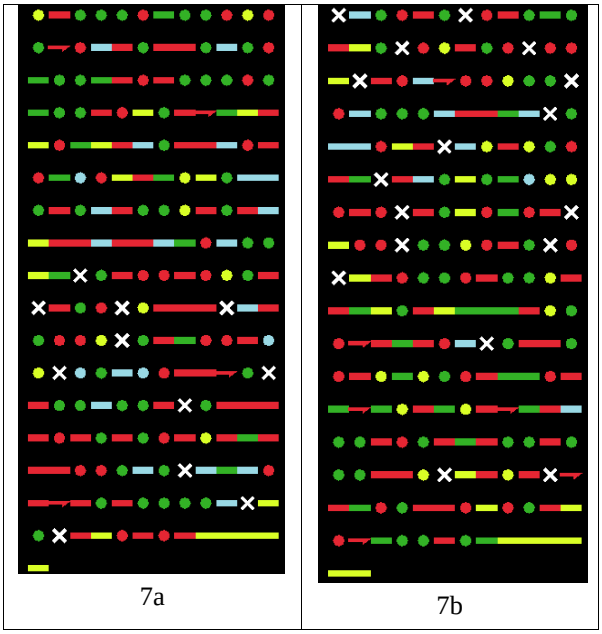


Figure 7: Representing features of an amino acid sequence in different reading frames

Being able to quickly swap between frames is a useful way of investigating reading frames.

4.2 Representing amino acids

The shapes, especially the stop codons, are very evident in Figure 7. The colours also contain information. There are 20 different amino acids, too many to be easily distinguished using colour. However it is possible to group the amino acids into various groups depending on their chemical properties. In Figure 7 the different colours represent amino acids that are hydrophobic, hydrophilic, acidic or basic (Table 2).

Group	Amino acids	Colour
Hydrophobic	A V L I P M F W	Red
Hydrophilic	G S T C Y N Q	Green
Acidic	K R H	Yellow
Basic	D E	Blue

Table 2: Showing the amino acid groups and their representative colours

5 Interacting with the visualisations

While the displays described so far give much information about the DNA or amino acid sequences, it is the interaction by the user which sets these apart from many existing applications.

We have already mentioned how changing the number of bases or amino acids displayed on a row can reveal patterns with different repeat lengths. The success of this depends on how easily the user can make the adjustments. The control window, shown in Figure 8, allows the user to adjust sliders and get immediate feedback.

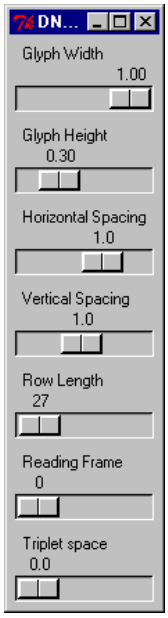
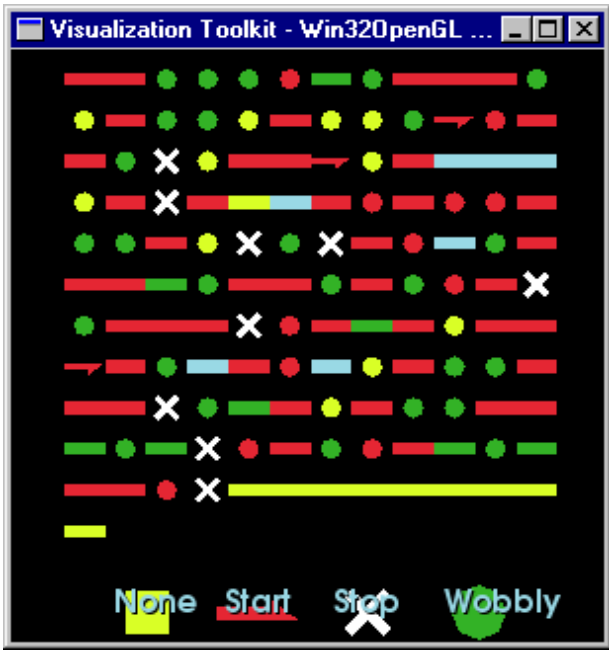


Figure 8: User control window

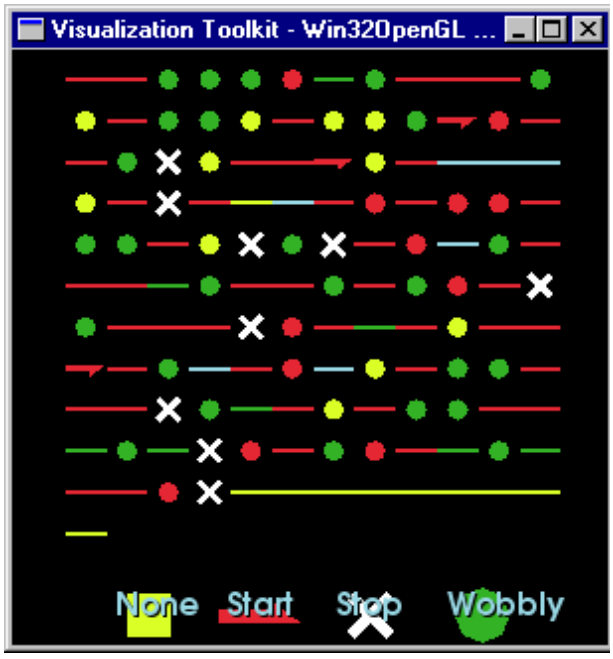
This allows the user to quickly adjust the height and width of the symbols, the spacing between each symbol and between the rows, to add spaces between codons in a DNA sequence, adjust the number of symbols on a row and investigate the three reading frames. The feedback is immediate. The same interactions are provided for the Protein Sequence Viewer, except for adjustment of space surrounding codons as this grouping exists only in DNA sequences.

Some examples of the effects of adjusting the display parameters are shown in Figures 9 and 10.

The height of the symbols in Figure 9b is smaller than in Figure 9a and this has the effect of emphasising the features and reducing the emphasis on the chemical properties of the amino acids. In Figure 10a we see the entire sequence of amino acids while in Figure 10b we have separated the glyphs and row spacing and zoomed in to allow closer investigation of the features.



9a

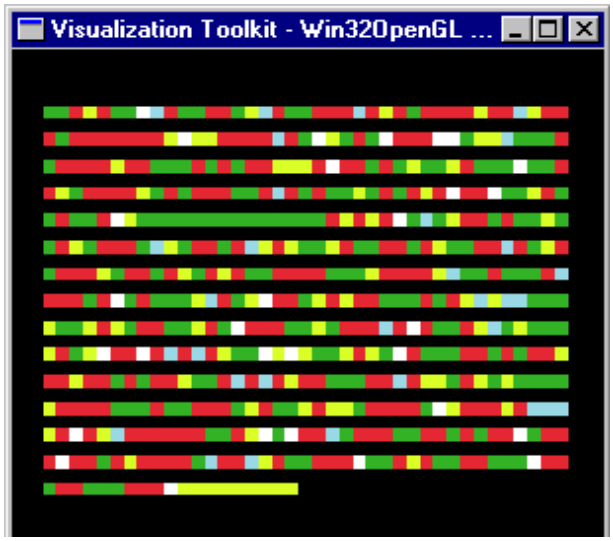


9b

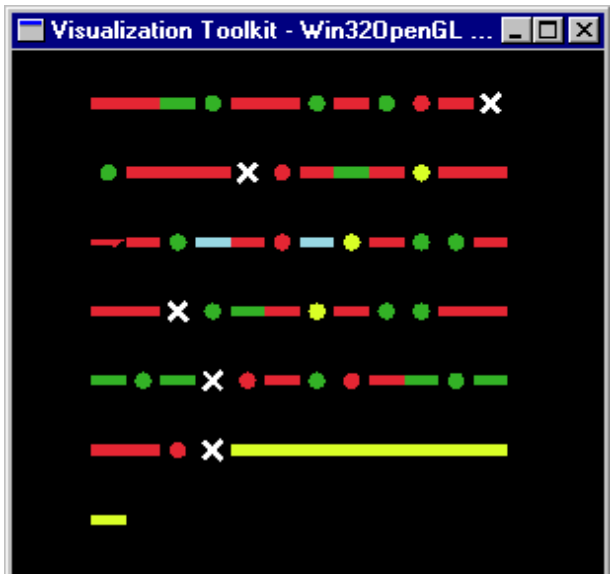
Figure 9: Emphasising features by reducing glyph heights

6 Conclusion

We have provided an interactive application that takes DNA sequence data in FASTA format and provides a visual representation of the information. The visualisation itself is very simple but the ability to interact enables users to discover patterns in DNA and amino acid sequences and to investigate possible reading frames. In addition information about the bases, the chemical nature of the amino acids and the presence or absence of variable regions are clearly visible. It is intended that this application be incorporated as a tool or component into public domain bioinformatics software



10a



10b

Figure 10: Overview versus detail

7 Acknowledgements

The authors wish to acknowledge the assistance of the New Zealand Foundation for Research, Science and Technology (C02X0203)

8 References

Combet, C., Blanchet, C., Geourjon, C., & Deléage, G. (2000). NPSA: Network Protein Sequence Analysis. *Trends In Biochemical Sciences* **25**(3):147-150.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, E., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., & Clamp, M. (2002), The Ensembl genome database project. *Nucleic Acids Research*. **30**(1):38-41.

Neshich, G. Togawa, R.C., Mancini, A.L., Kuser, P.R., Yamagishi, M.E.B., Pappas, G. Jr, Torres, W.V., Fonseca e Campos, T. , Ferreira, L.L., Luna, F.M., Oliveira, A.G., Miura, R.T., Inoue, M.K., Horita, L.G., de Souza, D.F., Dominiquini, F., Álvaro, A., Lima, C.S., Ogawa, F.O., Gomes, G.B., Palandrani, J.F., dos Santos, G.F., de Freitas, E.M., Mattiuz, A.R., Costa, I.C., de Almeida, C.L., Souza, S., Baudet, C., & Higa, R.H. (2003). STING Millenium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Research* **31**(13): 3386-3392.

Ousterhout, K. K. (1994). *Tcl and the Tk Toolkit*. Addison-Wesley Publishing Company, Massachusetts.

Parry-Smith, D.J., Payne, A.W.R., Michie, A.D., and Attwood, T.K. (1997), CINEMA – a novel Colour INteractive Editor for Multiple Alignments, *Gene*, **211**(2), GC45-56

Pearson, W. R., & Lipman, D. J. (1988). Improved Tools for Biological Sequence Comparison. *Proceedings of the National Academy of Science. U.S.A* **85**(8):2444-2448.

Purves, W. K., Orians, G.H., & Heller, H.C. (1995), Life: The science of biology, 4th ed., W.H. Freeman and Co., Utah.

Schroeder, W., Martin, K., & Lorensen, W. (1998). *The Visualisation Toolkit: an object oriented approach to 3D graphics*, 2nd ed.. Prentice-Hall, New Jersey.

Stoesser, G Wendy Baker, Alexandra van den Broek, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Francesco Nardone, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara and Robert Vaughan (2003), The EMBL Nucleotide Sequence Database: major new developments, *Nucleic Acids Research* **31**(1) 17-22**.**