Colour merging for the visualization of biomolecular sequence data

Mark Alston and Colin G. Johnson
Computing Laboratory
University of Kent
Canterbury, Kent, CT2 7NF
England
C.G.Johnson@kent.ac.uk

Gary Robinson
School of Biosciences
University of Kent
Canterbury, Kent, CT2 7NJ
England

Abstract

This paper introduces a novel technique for the visualization of data at various levels of detail. This is based on a colour-based representation of the data, where "high level" views of the data are obtained by merging colours together to obtain a summary-colour which represents a number of data-points. This is applied to the problem of visualizing biomolecular sequence data and picking out features in such data at various scales.

1. Introduction

In recent years the science of *bioinformatics* has come to prominence. This is driven by the easy accessibility of large amounts of biological data, e.g. through the human genome project. One way to cope with the complexity of bioinformatic data is via visualization techniques. This paper gives a brief introduction to the main problems in bioinformatics, surveys the role of visualization in bioinformatics and then introduces a new method for the visualization of DNA and protein sequences based on merging colours. Applications are described in finding features of protein sequences and in public understanding of science.

2. A quick trip through bioinformatics

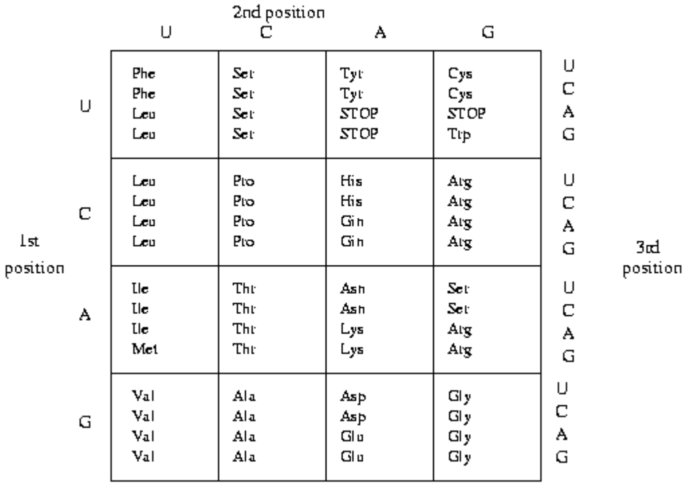
In recent years vast amount of information about biological systems has been obtained by various *sequencing* projects which extract the core information content from DNA and proteins found in an organism. Understanding the role of DNA and proteins in the functioning, development and evolution of organisms is the core concern of modern molecular biology.

Proteins are the molecules which play the largest role in the functioning of the body. Many of the structures in the body are made out of various kinds of protein molecules, and other kinds of proteins are involved in carrying energy around the body and in communicating chemical signals between organs. The set of protein molecules encompasses many complex, diverse chemical structures—yet at the same time the basic structure of protein molecules has a basic simplicity. Whilst proteins are complex three-dimensional structures, they are made from a "one-dimensional" molecule which is folded many times to produce a complex three-dimensional structure. The meaning of *one-dimensional* here is that the molecule is constructed of a linear strand of basic units, with no branching. There are twenty of these basic units: they are known as *amino acids*.

The order of these amino acids are specified by DNA molecules in the nucleus of the cell. DNA is another linear chain molecule, however there are only four basic units in DNA (usually abbreviated to C,A,T and G). The process of protein production from DNA is in two parts. Firstly strands of a DNA-like molecule called RNA are copied from the DNA strand (transcription), and these are carried to a molecular machine in the cell called a ribosome, which translates these into protein strands. The order of amino acids along the protein strand is specified by the order of the bases on the DNA strand. Three bases on the DNA strand code for one amino acid on the protein strand, according to the code given in figure 1 (the amino acids all have three-letter abbreviations).

When this reaches the DNA triple which codes for *stop*, the process stops and releases the protein strand. Interactions between the components of the strand cause it to fold up into a complex three dimensional structure (an example is given in figure 2). Understanding the relationship between the sequence of amino acids and the final structure is one of the most complex open questions in bioinformatics. A *gene* is the DNA which encodes for one protein molecule.

In order to understand these DNA and protein structures a large amount of information has been measured about var-



1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure 1. DNA to amino acid coding.

ious organisms. Much of this information is in the form of *sequences*. A DNA or protein sequence is an ordered list of the basic units which make up a particular gene. Sequencing projects are one of the main sources of information in bioinformatics; the most well known sequencing project is the Human Genome Project, which sequenced the DNA in humans. Another kind of information which feeds into bioinformatics projects is information about the 3-dimensional structure of proteins, obtained for example by X-ray diffraction or NMR spectroscopy. Databases of these various kinds of data, in standard file-formats, are a major resource for bioinformatics.

A number of questions can be approached using bioinformatic data. One of the most important classes of questions concerns the relationships between DNA sequences or protein sequences. If we put two protein sequences side-by-side, can we identify regions of similarity? If so, do these indicate a common history for these two proteins, or is it more likely that the two converged to a similar structure due to the need to provide similar function? Given a number of seemingly related DNA sequences, can we determine the probability that they have a common ancestor, and determine the likely order of branching? Given a protein structure, can we determine qualitatively different regions within the strand which could map to different functional domains in the final folded structure? Given a DNA or protein sequence and a corresponding 3-dimensional protein structure, can we relate features of the sequence to features of the structure, or predict the structure from the sequences?

Detailed introductions to bioinformatics can be found in a number of books and web tutorials [1, 2, 6].

3. Visualization techniques in bioinformatics

A large number of packages are available for the analysis of bioinformatics data, and many of them incorporate some way of interacting visually with the data.

One aspect of visualization is viewing the 3-dimensional structure of protein molecules. A number of programs facilitate this, and an example, viewed using the *Swiss pdb-viewer* program, is given in figure 2. An innovative recent approach to the visualization of such structures is the use of 3D-printer technology [11, 12]. This produces 3-dimensional plastic models of the molecules. These have been successfully applied in the investigation of “lock-and-key” type matching between molecules. Another innovative way of visualizing these is by using stereopsis pictures [5], similar to the system used to give the illusion of depth in random-dot stereograms.

Another aspect is the display of predicted relationships between various organisms. The difference between DNA sequences can be used to predict the order in which species will have branched off from each other over evolutionary time. Presenting this data in an intuitive way is an interesting challenge, which draws heavily on work in graph drawing [4].

Perhaps the greatest challenge in bioinformatics visualization is the display of superficially simple sequence data. The previous examples in the section have concerned structures which have a natural structure to be visualized; in sequence display visual information has to be put in to emphasize appropriate features.

One feature of sequences which can be highlighted using visualization techniques is similarity between similar regions along a chromosome. For example the *MapViz* program [9] allows the user to select regions along a DNA strand and dynamically creates graphs to show which parts of the strand have some similarity to that region.

One method which is commonly used in the display of protein sequences is to break the set of twenty amino acids into a number of subsets so that substitution of one amino acid for another is more likely within the subsets than between subsets. A number of chemical properties of the amino acids can be used to create such a set, as can statistical analysis of proteins which are slightly different from each other in sequence but known to belong to the same lineage. Over evolutionary time mutations will sometimes cause a change in the DNA which leads to a change in the protein; however changes which make radical differences to the structure of the protein are less likely to be preserved by natural selection, as such changes are likely to disrupt the functioning of the organism. Therefore the most likely substitutions are between amino acids with similar chemical characteristics; a hydrophobic amino acid will probably substitute for another hydrophobic one, for example. These

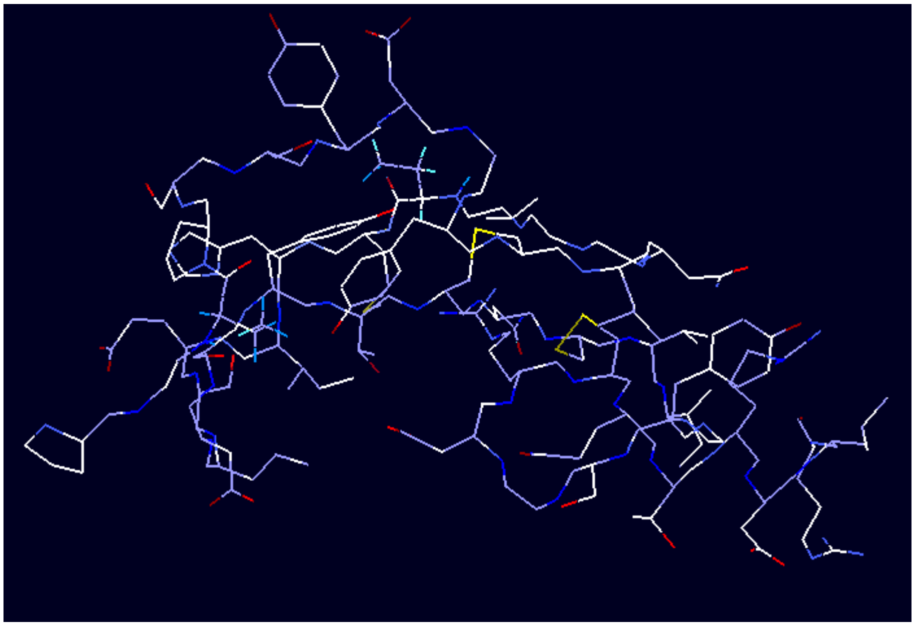
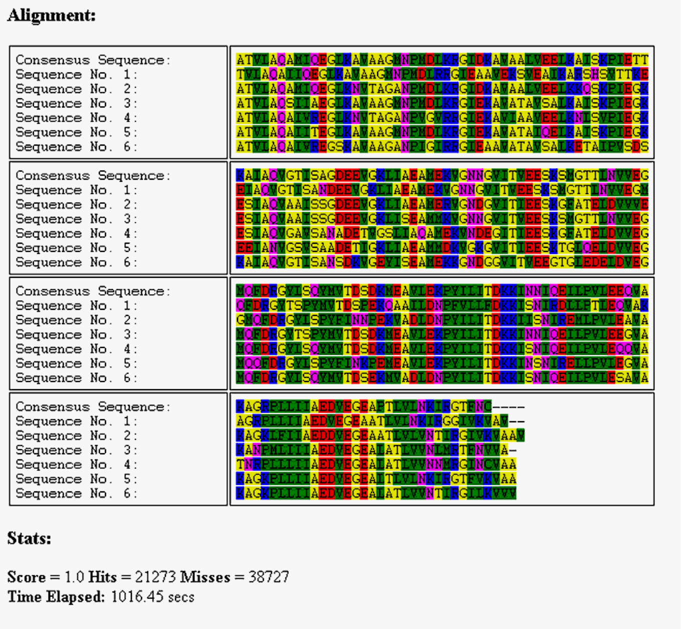


Figure 2. An illustration of a folded protein, created by the swiss-pdb viewer program.

substitution-subsets can each be given a colour in a visualization, as illustrated in figure 3.

Another attempt to visualize such properties is given by Williams et al. [14]. They represent each amino acid position with five vertical spaces, with the spaces filled according to the chemical properties of the residues (e.g. the lowest space is filled in if the residue was hydrophobic). This leads to sequences resembling Morse code, with some structural features claimed to be highlighted by the resulting pattern of dots. Similarly, Ninio & Mizraji [7] employed graphical coding of DNA bases; this was effective at highlighting simple, repetitive patterns. Another related representation is the *sequence logo* developed by Schneider et al. [10]. The properties of a protein's amino acids may also be visualized in the form of a line graph, as illustrated by figure 4 which shows the protein rhodopsin using the Eisenberg hydropathic scale.

However much of this work simply presents small scale details, and there is little attempt to provide a "higher level" view which visualizes structural features spanning a number of positions in the sequence. Trying to pick out mid-scale features from these kinds of visualizations is like looking closely at a picture printed in a newspaper, where you see lots of dots and find it difficult to see the details of the picture. The aim of out current work is to provide visualizations which enable us to pick out such mid-range features.



Alignment:

```
Consensus Sequence:   ATVLAQAMIQEGLKAVAAGMNPMDLKRGIDKAVAALVEELKAISKPIETT
Sequence No. 1:       TVLAQAIIQEGLKAVAAGMNPMDLRRGIEAAVEKSVEAIKARSHSVTTKE
Sequence No. 2:       ATVLAQAMIQEGLKNVTAGANPMDLKRGIDKAVAALVEELKKQSKPIEGK
Sequence No. 3:       ATVLAQSIIAEGLKAVAAGMNPMDLKRGIEKAVATAVSALKAISKPIEGK
Sequence No. 4:       ATVLAQAIVREGLKNVTAGANPVGVRRGIEKAVIAAVEELKNISVPIEGK
Sequence No. 5:       ATVLAQAIITEGLKAVAAGMNPMDLKRGIEKAVATAIQELKAISKPIEGK
Sequence No. 6:       ATVLAQAIVREGSKAVAAGANPIGIRRGIEAAVATAVSALKETAIPVSDS

Consensus Sequence:   KAIAQVGTISAGDEEVGKLIAEAMEKVGNNGVITVEESKSMGTTLNVVEG
Sequence No. 1:       EIAQVGTISANDEEVGKLIAEAMEKVGNNGVITVEESKSMGTTLNVVEGM
Sequence No. 2:       ESIAQVAAISSGDEEVGKLIAEAMERVGNDGVITIEESKGFATELDVVVE
Sequence No. 3:       ESIAQVAAISSGDEEVGKLISEAMMKVGNNGVITVEESKSMGTTLNVVEG
Sequence No. 4:       ESIAQVGAVSANADETVGSLIAQAMEKVNDEGITIEESKGFATELDVVEG
Sequence No. 5:       EEIANVGSVSAADETIGKLIAEAMMDKVGKGVITIEESKTGLQELDVVEG
Sequence No. 6:       KAIAQVGTISANSDKVGEVISEAMEKKGNDGGVITVEEGTGLEDELDVEG

Consensus Sequence:   MQFDRGYISQYMVTDSDKMEAVLEKPYILITDKKINNIQEILPVLEEQVA
Sequence No. 1:       QFDRGYTSPYMVTDSPEKQAAILDNPFVLLFDKKISNIRDLLPTLEQVAK
Sequence No. 2:       GMQFDRGYISPYFINNPEKVADLDNPYILITDKKIISNIREMLPVLEAVA
Sequence No. 3:       MQFDRGYTSPYMVTDSDKMEAVLEKPYILITDKKINNIQEILPVLEEGVA
Sequence No. 4:       MQFDRGYISQYMVTDSDKMEAVLEKPYILITDKKISNIQEILPVLEQQVA
Sequence No. 5:       MQQFDRGYISPYFINKPEMEAVLEKPYILITDKKINSNIRELLPVLEGVA
Sequence No. 6:       MQFDRGYISQYMVTDSEKMVADLDNPYILITDKKISNIQEILPVLESAVA

Consensus Sequence:   KAGRPLLIIAEDVEGEAPTLVINKIRGTFNC----
Sequence No. 1:       AGRPLLIIAEDVEGEAATLVINKIRGGIVKVAV--
Sequence No. 2:       KAGKLFIIAEDDVEGEAATLVLVNTIRGIVKVAAV
Sequence No. 3:       KANPMLIIIAEDVEGEALATLVVNLMRTFNVVA-
Sequence No. 4:       TNRPLLLIIAEDVEGEALATLVVNNMRGINCVAA
Sequence No. 5:       KAGKPLLIIAEDVEGEALTLVINKIRGTFVKVAA
Sequence No. 6:       KAGKPLLIIAEDVEGEALATLVVNTIRGILKVVV
```

Stats:

Score = 1.0 **Hits** = 21273 **Misses** = 38727
Time Elapsed: 1016.45 secs

Figure 3. Using colour to overlay substitution-subsets onto a number of related sequences.

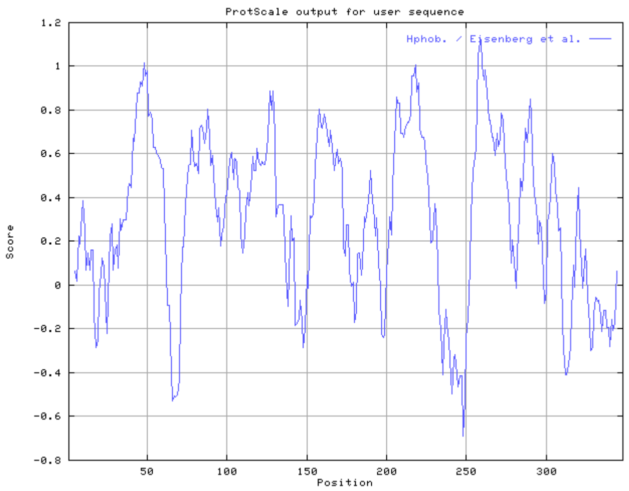


Figure 4. A line-graph representation of a chemical property of a protein sequence.

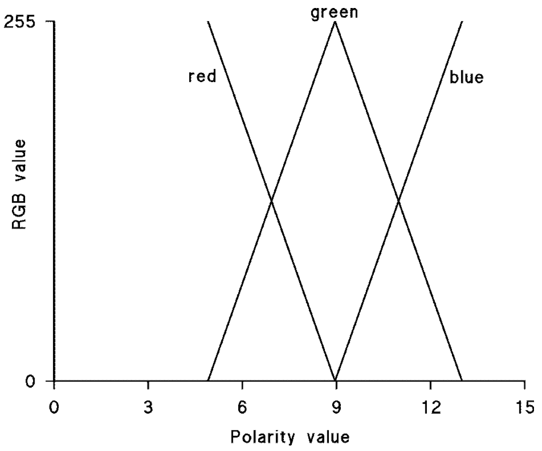


Figure 5. The mapping from polarity to RGB colour values.

4. Colour and visualization

Clearly colour is an important aspect of visualization [3, 13]. Colour can be used to highlight features in a complex data set, and to give an illusion of distance, e.g. text in some colours seems "closer" than text in others. Colour can be used to label discrete entities by assigning different colours to different data items or different types of data, and it can be used to label continuous data by assigning values on some continuum through colour space to values of a continuous variable. People have a natural affinity to colour (though colour-blindness of various sorts is common), and colour can be used to add another dimension to a visualization which can be readily perceived orthogonally to other ways of representing data.

The work below extends this use of colour in a new way. Individual items from a sequence of data are assigned colours, however we make use of merging colours together to get a higher level view of a parts of a sequence.

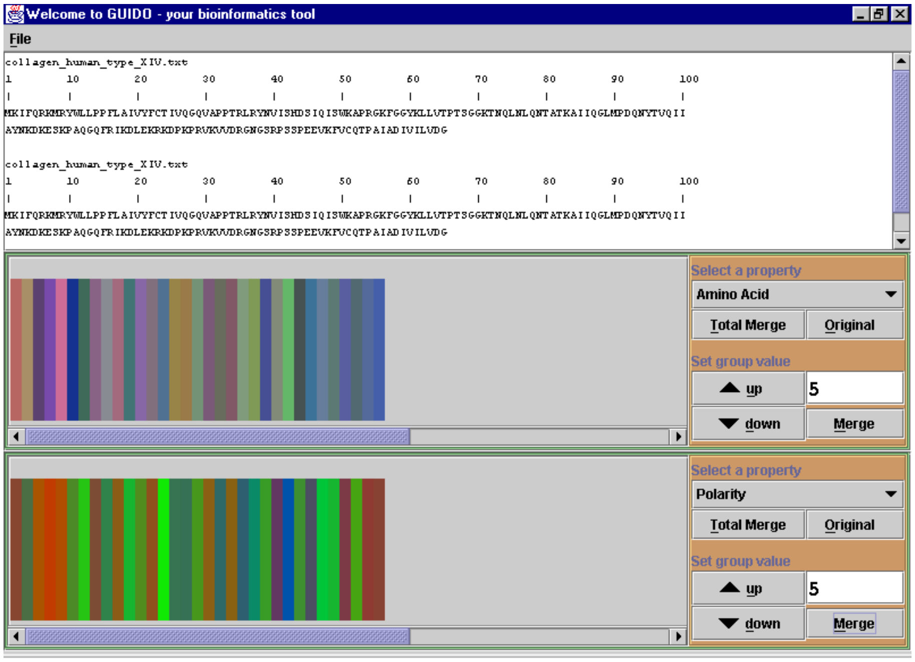
5. Using colour merging to visualize sequences

We have been investigating how the *merging* of colours can be used as a visualization technique in bioinformatics. By merging we mean the combining of colours representing various pieces of information into a single colour, as in paint mixing or the merging which occurs when several coloured lights are shone on the same area.

This colour merging would seem to have potential as a visualization technique. The core idea is that a mapping is established between values which items in a data set can take and colours. A view of the information can then be created where each item in the set is displayed by displaying a piece of that colour. Then to "zoom out" to a coarser scale view of that information, we blend the colours together using some kind of averaging process in colour space.

An application has been created which allows the display of bioinformatic sequence data in this fashion. The sequence data is read in using the standard FASTA file format [6]. This data is displayed as a "bar-code" across the screen. Initially each of the basic units in the sequence (bases for DNA sequences, amino acids for protein sequences) are allocated a thin rectangle. For proteins a number of different features of the sequence can be illustrated. In the default view each of the amino acids is assigned a a discrete colour, chosen by picking twenty colours widely scattered around the RGB colour space. This can be changed to display one of a number of chemical characteristics of the amino acids, for example their polarity and hydrophobicity: an example of the scale used is illustrated in figure 6. These "bar-codes" can be stacked on top of another to compare different properties of the same sequence, or to compare several sequences for similarity (figure 5).

The user can then adjust the resolution of the image by choosing a "merge value" m. Starting from the beginning, the sequence is divided into blocks of length m. These are then replaced by a single merged colour. This new colour is calculated by regarding the colours as points in an RGB colour-space, and taking the centroid of those points to get a new point which is then interpreted as a colour.



Welcome to GUIDO - your bioinformatics tool

File

```
collagen_human_type_XIV.txt
1       10        20        30        40        50        60        70        80        90        100
|        |         |         |         |         |         |         |         |         |         |
MKIFQRKMRYWLLPPFLAIVYFCTIVQGQVAPPTRLRYNVISHDSIQISWKAPRGKFGGYKLLVTPTSGGKTNQLNLQNTATKAIIQGLMPDQNYTVQII
AYNKDKESKPAQGQFRIKDLEKRKDPKPRVKVVDRGNGSRPSSPEEVKFVCQTPAIADIVILVDG

collagen_human_type_XIV.txt
1       10        20        30        40        50        60        70        80        90        100
|        |         |         |         |         |         |         |         |         |         |
MKIFQRKMRYWLLPPFLAIVYFCTIVQGQVAPPTRLRYNVISHDSIQISWKAPRGKFGGYKLLVTPTSGGKTNQLNLQNTATKAIIQGLMPDQNYTVQII
AYNKDKESKPAQGQFRIKDLEKRKDPKPRVKVVDRGNGSRPSSPEEVKFVCQTPAIADIVILVDG
```

Select a property: Amino Acid

Set group value: 5

Select a property: Polarity

Set group value: 5

Figure 6. The interface to the application.

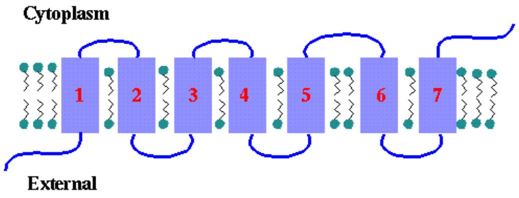


Figure 7. The structure of the *rhodopsin* transmembrane protein.

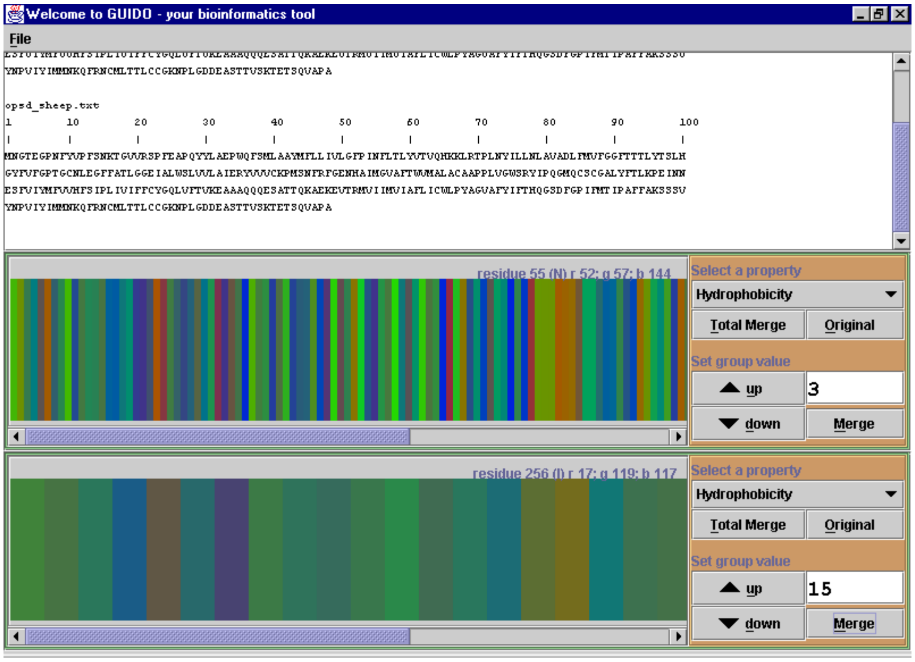
6. Applications 1: Features of sequences

We can see two main areas of applications for this work. The first is as an exploratory tool to allow bioinformatics researchers to pick out areas of sequences which have distinctive features contrasted with neighbouring regions.

As a case study consider the protein *rhodopsin*, which is a molecule which plays a part in processing light which falls on the retina of the eye. This is a *transmembrane* protein, which winds in and out of a cellular membrane, as illustrated in figure 7. An interesting question is to work out which parts of the protein sequence form the (seven) transmembrane regions.

The protein sequence for rhodopsin was loaded into the program and a number of characteristics examined. the characteristic which proved to have most distinctive patterns for this protein was *hydrophobicity*, which is represented in the program using a scale which represents highly hydrophobic regions as blue colours through to red colours for contrasting hydrophilic regions. The size of the merge group was varied from 3 to 15, and regions of high contrast noted; in particular a number of red-blue boundaries were identified (at various scales) which indicate a large change in the character of the protein at that point (figure 8). It is notable that some regions are more visible at certain merge values, which demonstrates the usefulness of being able to adjust the merge resolution.

Table 1 indicates the location of these regions, whilst table 2 presents a comparison of the consensus prediction from the visualization compared with actual known values from experimental data. The visualization was broadly successful at identifying the transmembrane regions of rhodopsin, and would have enabled an expert (who would know that sharp hydrophilic/hydrophobic regions are indicative of transmembrane domains) to identify the protein family to which this belongs. Some of the transmembrane regions (1,5 and 6) were highlighted across all four merge levels, and it was these which compared most strongly with the real data. Regions 2,3 and 7 showed a fair correlation between predicted and actual values, highlighted across three merging levels. Region 4 was poorly highlighted, being weakly identified at one merge level. It is unlikely that this would have been easily identified without some knowl-



```
ESFVIYMFVVHFSIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKSSSV
YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTVSKTETSQVAPA

opsd_sheep.txt
1        10        20        30        40        50        60        70        80        90        100
|        |         |         |         |         |         |         |         |         |         |
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLH
GYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINN
ESFVIYMFVVHFSIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKSSSV
YNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Figure 8. Merged sequences at different merge values.

edge of the number of transmembrane regions in advance. Also, at merge levels 3 and 7, a spurious region was identified between residue numbers 190–196. Nonetheless the overall results are promising.

In current work on these kinds of proteins, transmembrane regions are identified by looking at data on a number of hydrophobic scales: each brings a little more confidence to the interpretation of where the transmembrane regions are located. A similar situation exists with the current application; higher confidence can be put in data which shows similar highlights across a *range* of merge values.

7. Applications 2: An intuitive understanding of protein translation via colour merging

A second area of application for these ideas is in public understanding of science. The ideas of genetic, protein biology and bioinformatics are difficult to explain due to their abstract nature. The idea of colour merging provides an analogy between a process with which many of the general public are familiar (e.g. through mixing paints) and one with which they are not familiar (molecular biology).

In particular the process of taking triplets of DNA bases and converting them into amino acids (as in figure 1) has a nice analogy with merging colours. A small number of base colours can be mixed in different combinations to produce different colours. Providing the proportions are different, then the order of the three chosen base colours will be important too.

An interesting challenge would be to calculate a set of four base colours in an appropriate colour space so that the results of this merging process produce a set of colours so that each amino acid is represented by a set of similar colours.

8. Future work

We have presented a snapshot of current work in progress on developing a novel visualization algorithm and an application in bioinformatics. A number of important challenges remain, such as making a rational choice for the colours rather than just using arbitrary information→colour maps, and testing the system with bioinformatics experts who are not familiar with the details of how the system works.

A full colour version of the paper is available from the author's web site `http://www.cs.kent.ac.uk/people/staff/cgj/`

References

[1] T. Attwood and D. Parry-Smith. *Introduction to Bioinformatics*. Addison Wesley Longman, 1999.

[2] A. Brazma, H. Parkinson, T. Schlitt, and M. Shojatalab. A quick introduction to elements of biology—cells, molecules,

Merge Value	Location of red/blue boundary regions (residue number)						
3	55–60	94–97	136–139		190–192 226–228	271–276	292–294 305–324
7	8–14 57–63	99–105	134–140?		190–196 218–224	260–266	316–322
12	49–60			156–168?	217–228	265–276	
15	46–60	92–105	127–140		211–225	256–270	

Table 1. The predicted values for the location of the rhodopsin transmembrane regions, determined by red/blue boundaries at various scales in the hydrophobicity bar-code. A ? next to a number indicates a weak match.

Transmembrane region number	1	2	3	4	5	6	7
Actual location	37–61	74–98	114–133	153–176	203–230	253–276	285–309
Consensus from visualization	46–63	92–105	127–140	157–168?	211–228	256–276	292–324

Table 2. Comparison of predicted and actual positions of the transmembrane regions in rhodopsin.

genes, functional genomics, microarrays. `http://www.ebi.ac.uk/microarray/biology_intro.html`.

[3] R. Jackson, L. MacDonald, and K. Freeman. *Computer Generated Color*. Wiley, 1994.

[4] J. Klingner. Visualizing sets of evolutionary trees. Technical Report CS-TR-01-26, University of Texas at Austin, Department of Computer Sciences, 2001.

[5] A. Lesk. *Introduction to Protein Architecture*. Oxford University Press, 2001.

[6] A. M. Lesk. *Introduction to Bioinformatics*. Oxford University Press, 2002.

[7] J. Ninio and E. Mizraji. Perceptible features in graphical representations of nucleic acid sequences. In Pickover [8], pages 33–42.

[8] C. A. Pickover, editor. *Visualizing Biological Information*. World Scientific, 1995.

[9] A. J. Robinson and T. P. Flores. Novel techniques for visualizing biological information. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, 1997.

[10] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, 18:6097–6100, 1990.

[11] T. S. Shimizu. *The Spatial Organization of Cell Signalling Pathways: A Computer Based Study*. PhD thesis, University of Cambridge, 2002.

[12] T. S. Shimizu, N. Le Novère, M. D. Levin, A. J. Beavil, B. J. Sutton, and D. Bray. Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. *Nature Cell Biology*, 2:792–796, 2000.

[13] C. Ware. *Information Visualization*. Morgan Kauffman, 2000.

[14] A. Williams, K. Chenault, and U. Melcher. Graphic representations of amino acid sequences. In Pickover [8], pages 6–14.