DNA Sequence Analysis: The Development of a Custom Software Tool

Tim Hunt, David R. Musgrave

The Waikato Institute of Technology
Hamilton, New Zealand
[1]tim.hunt@wintec.ac.nz

[2]The University of Waikato

Proceedings of the 15th Annual NACCQ, Hamilton New Zealand July, 2002 www.naccq.ac.nz

ABSTRACT

The development and application of a new software tool, for the analysis of DNA sequence data, is described. The software was developed in a modular fashion using reusable Borland Delphi components. An algorithm to calculate the G+C% of the imported sequence was implemented with user options for choosing either the window and shift variables within the genome sequence as a whole or identified genes in the input file. To assist users to determine the significance of the data, an option is also provided to allow the comparison of the sequence data with a random data set and to allow further analysis, data can be exported to a text file. Analysis of the *Aeropyrum pernix* K1 genome is presented and compared with results obtained with other published tools. The significance of choosing an appropriate window size is described. A number of published algorithms that present the opportunity for further development are also reviewed.

1. INTRODUCTION

The quantity of genome data that is currently being made freely available (via the WWW) is increasing at an exponential rate. The volume of data is such that information extraction from this data is lagging the production of new data. Biologists have turned to Information Technology for tools to analyse the data for useful information. This area of research is known as BioInformatics.

A principle focus of genome analysis is the identification of genes for medical, pharmaceutical and agricultural purposes. DNA is composed of building blocks called deoxyribonucleotides consisting of a deoxyribose sugar, a phosphate group and one of four nitrogen bases: adenine (A), thymine (T), guanine (G) and cytosine (C). In its most simple form a gene is a linear sequence of nucleotides that provides the cell with the information for the synthesis of a protein that catalyses a particular cellular reaction. Once the nucleotide sequence of a DNA molecule has been obtained, there is a range of tools available to help with the identification and analysis of genes. The nucleotide sequence of DNA is an information system that is slowly being understood and although we have learned a great deal about the DNA sequence of genomes and individual genes, we are still a long way

from completely understanding how DNA programmes the complex events that describe life. It is of particular importance to understand how the structure of genomic DNA, and the way in which the DNA is packaged into chromosomes, control how the set of genes in the genome are utilised. For example the methylation of CG (when G immediately follows C in the nucleotide sequence) in genomic DNA is an epigenetic mark involved in such processes as X chromosome inactivation, imprinting and gene silencing. CG methylation has recently been shown to be influenced by the methylation of the histone proteins that package DNA into chromatin and subsequently control its use Tamaru and Selker (2001). A lack of understanding of the influence that the state of chromatin has on developmental processes is thought to result in the developmental aberrations and death of cloned animals because even though the cells used to generate animal clones have a full complement of DNA the way in which it is packaged into chromosomes is not always appropriate to set up the correct developmental programme Vastag (2001).

2. SEQUENCE ANALYSIS

There are numerous tools available for gene discovery, the most popular being the set of tools called BLAST Altschul et al. (1990). This set of tools provides a method of searching databases containing both fully sequenced genomes and partial sequences in order to match a DNA or protein sequence to sequences located in the databases.

2.1 G+C% ANALYSIS

One of the defining characteristics of a genome is its G+C% defined as (NG + NC) / (NG + NC +NT + NA) x 100 % where NG/C/T/A are the number of G, C, T, or A containing nucleotides respectively in the sequence; for example the overall G+C% of the *Aeropyrum pernix* K1 genome is 56.3%, Kawarabayasi *et al.* (1999). The G+C% signature of a genome is significant because it can vary considerably both within a genome and between genomes of different species Bernardi *et al.* (1997). The G+C% has a significant affect on the structure and function of a genome. Firstly, the G+C% of a genome determines the relationship that exists between the nucleotide sequence of the messenger RNA encoded by DNA and the amino acid sequence in the resulting protein. This relationship is known as the codon bias and it significantly affects the genetic relationship between organisms Karlin et al. (1998). Secondly, eukaryotic genomes contain regions known as satellite DNA that contain repeating sequences that do not code for genes but play an important structural role in chromosomes (for review see Miklos 1985). Satellite DNA has a low G+C% compared with the majority of the DNA and so has been used to help identify regions of genomes that are likely to code for genes. Thirdly the variation in G+C% of a DNA sequence with respect to the genome as a whole may indicate that the DNA has been acquired by a recent horizontal transfer from another species rather than having been acquired vertically by a normal reproductive event Koonin *et al.* (2001). For example a review by Karlin (1998) shows a plot of the G+C% of the Helicobacter *H.pylori* highlighting a region of low G+C% corresponding to the known pathogenicity island (Pa-I), a mobile DNA element involved in bacterial virulence that has presumably been transferred to *H.pylori* from an organism whose genome is significantly different in G+C%. Also whole genome comparisons have indicated significant differences in gene content among bacteria that are phylogenetically related indicating that genome evolution could not be reasonably described in terms of vertical descent alone Tatusov *et al.* (1996). Horizontal gene transfer (HGT) is therefore particularly important in the evolution of prokaryotic genomes by the incorporation of foreign DNA. Because of this, HGT is also an important factor in the debate surrounding the possibility of the transfer and subsequent spread of genes from genetically modified organisms to other organisms in the environment.

2.2 DINUCLEOTIDE ABUNDANCES

Karlin *et al.* (1998) summarised the range of dinucleotide relative abundances as defined by the genome signature: ?* = f*XY /f*Xf*Y for all XY, where f*X denotes the frequency of the nucleotide X and f*XY denotes the frequency of the dinucleotide XY. They concluded that dinucleotide relative abundance is a measure of genetic distance and that the dinucleotide signature agrees well with the phylogenetic distance determined by comparison of 16S rRNA genes Woese *et al.* (1990). This data suggests that dinucleotide sequences might reflect the influence of genome sequence on such factors as the functions of the DNA replication and repair

machinery and DNA mutations. In the thermophilic Archaea surveyed, the GC dinucleotide is usually highly underrepresented whereas the CC and GG dinucleotides are usually highly overrepresented. This is consistent with the under representation of CG dinucleotides in vertebrate sequences but not with its overrepresentation in many bacterial genomes.

2.3 THE SLIDING WINDOW SIZE

G+C% values are calculated for a certain number of base pairs and then the window is shifted by a certain number of base pairs before the G+C% for the next window is calculated. A range of sequence window sizes has been used in the published data. Karlin *et al.* (1997) used approximately 50 kbp windows and in a further study Karlin (1999) used sliding windows of 10, 20 and 50 kbp in length. Lio and Vannucci (2000) discuss the problems with choosing the window size and have presented results using 'wavelets'.

3. TOOLS FOR G+C% ANALYSIS

There are a number of tools available for G+C% analysis, each tool having different options and limitations and they are discussed here. The European Molecular Biology Open Software Suite (http://www.emboss.org/) offers the DEA: analyse_seq tool (http://www.isrec.isb-sib.ch/~mstadler/exercise2_2/analyse_seq.html). The tool does not allow the importation of the sequence files, but instead you 'copy and paste' a sequence into the window provided. You are able to choose the window and shift sizes. The G+C% results can be either displayed as a graph or the raw G+C% values can be saved in a text file for further analysis. The Institute for Genome Research (TIGR) provides the 'Comprehensive Microbial Resource' that includes a tool for G+C% analysis, http://www.tigr.org/tigr-scripts/CMR2/GCDisplay.spl?asmbl_id=49. To use this you choose the genome that you wish to analyse from the list of genomes already available on the web site. This eliminates the requirement of uploading the genome that you are interested in, however this can be a restriction if you have a sequence that is not available on the server. You can change the window size, but not the shift size, which is set to be equal to the window size. The results are displayed as a graph and the data can also be saved to a text file for further analysis. The Genome Information Broker (http://gib.genes.nig.ac.jp/) provides a tool called 'GC Plot' which allows the window size to be chosen, but the window shift is set automatically to an unspecified value. You do not need to upload data as the published genomes are already on the server. The results are only available as a graphic, with no option for looking at the raw data. The Sanger Institute (http://www.sanger.ac.uk/) provides a tool called Artemis that can be downloaded to run on your local computer. It is a JAVA application and one of its options is G+C% analysis. Data is loaded from a standard file format and the G+C% graph is displayed. The window size can be selected and the shift is automatically fixed to 10% of the window size. Results are graphically displayed but there is no option for saving the raw data.

4. WCBI DNA ANALYSER

For this study, the Waikato Centre for BioInformatics (WCBI) DNA Analyser (WDA) has been developed as a freely available tool for G+C% sequence analysis. An object oriented design approach using Borland Delphi (http://www.borland.com/) was used and the WDA runs on the MSWindows platform. The user can enter the window size and shift size as separate parameters. The G+C% is graphically displayed as a function of sequence position. The software can import the two most common genome sequence file types-GenBank and FASTA. When importing GenBank formats, the WDA can use the information of identified genes in the file to create plots of gene size and composition. The data can also be exported to a Comma Separated Value (CSV) file for further analysis. The software is available from the author on request.

4.1 SOFTWARE ARCHITECTURE

Emphasis was placed on creating separate reusable components for each implemented function, thus enabling the easy incorporation of future functions. The two main visual components developed were the CDNASequence and CDNAGPlusCAnalysis components, which handle the storage of raw sequence data and subsequent analysis of that data respectively. In addition the main screen has an associated Delphi unit that coordinates the requests from the user such as importing and exporting of data.

4.1.1 The CDNASequence visual component

The CDNASequence component is derived from the Delphi TComponent class and contains two arrays and seven procedures (methods). The main array holds the raw base pair (bp) data as single characters; the second array holds data on the size and position of gene data read from a GenBank format data file. There are two procedures for reading the different file formats, a procedure (GetBP) that returns the bp of a specified position, a procedure for returning the total length of the genome, a procedure to create a random sequence and two procedures associated with analysing annotated gene information.

4.1.2 The CDNAGPlusCAnalysis visual component

The CDNAGPlusCAnalysis component (derived from the Delphi TComponent class) does not hold any data but instead uses the data held in the CDNASequence component. It has a single procedure that calculates the number of each bp type in a DNA sequence when given the start position and length. It calls the GetBP procedure of the CDNASequence component and increments a counter for each bp type and then returns the totals back to the calling procedure for G+C% calculation.

ORIGIN

1	aaataataat	aaaaattaag	tgactcatgc	attatcctac	gaggtaaaaa	tatgttataa
61	attgtcccag	actaccatca	atttagggac	aatagtgttt	aagggatggc	cttcggagct
121	ggcagctcgc	gggttcaaac	tcgcgtaggg	cccgagttct	agttatagtt	gcgtggattt
181	agataaattg	agtatgatct	ctcagtttta	tatcaatact	taccctcttt	attaatcata
241	attaacattg	ttacaacgaa	tagagtggtc	actcccgcca	acaggattat	ccaccacata
301	tggaatcctg	ctaaaatcat	atatacacct	atagctatga	gagataagga	ggttggcatg
361	aaaaattgta	atagctcgat	cgtttcccga	cttagtcctg	gtattacata	tgcttccggc
421	ctttttacag	atagaaaaac	ggtatatcct	gctaatattt	caatagataa	atgtgaagct

Table 1: A sample of a GenBank format data file showing DNA sequence data

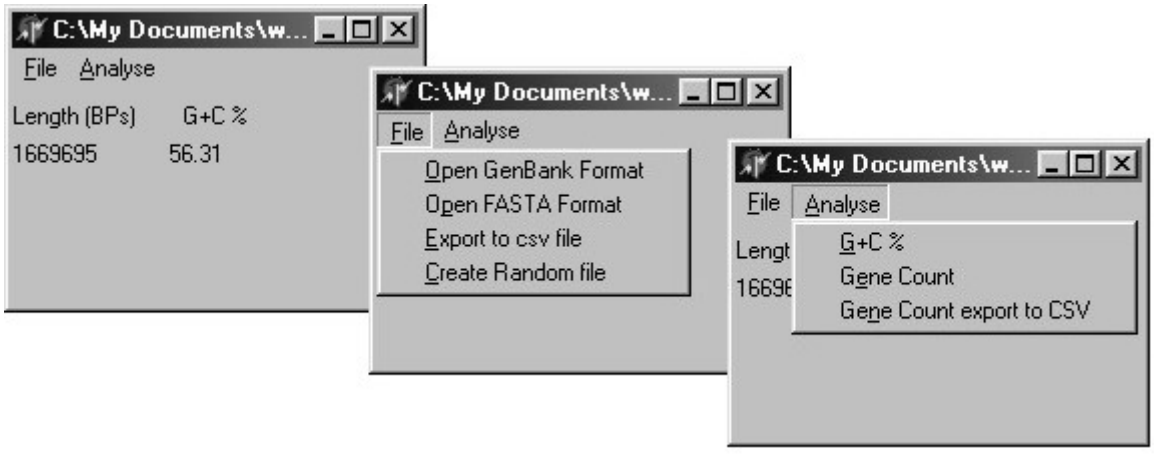

Figure 1: The main user interface of the WDA showing the three different states of user interaction

Gene complement(1665992..1666417)
/gene=”APE2616"
CDS complement(1665992..1666417)
/gene=”APE2616"
/note=”similar to PIR:H69385 percent identity:40.288 in
139aa.”
/codon_start=1
/transl_table=11
/product=”141aa long hypothetical protein”
/db_xref=”PID:d1045420"
/db_xref=”PID:g5106323"
/translation=”MIDTSVFADYYLLYPGDPERHERSRTVLDKLSLRDVIVYEPFLF
EIELRAVLVRRIPPEQALRIVDTTLKHVNVVREEELHDKAAEIALITGCRAVDAYFIA
TAKHVDGILITNDKVMKDNAQKIGVKAYYLLDNQDYTKL”

Table 2: A sample of a GenBank format data file showing annotated gene information in the file header

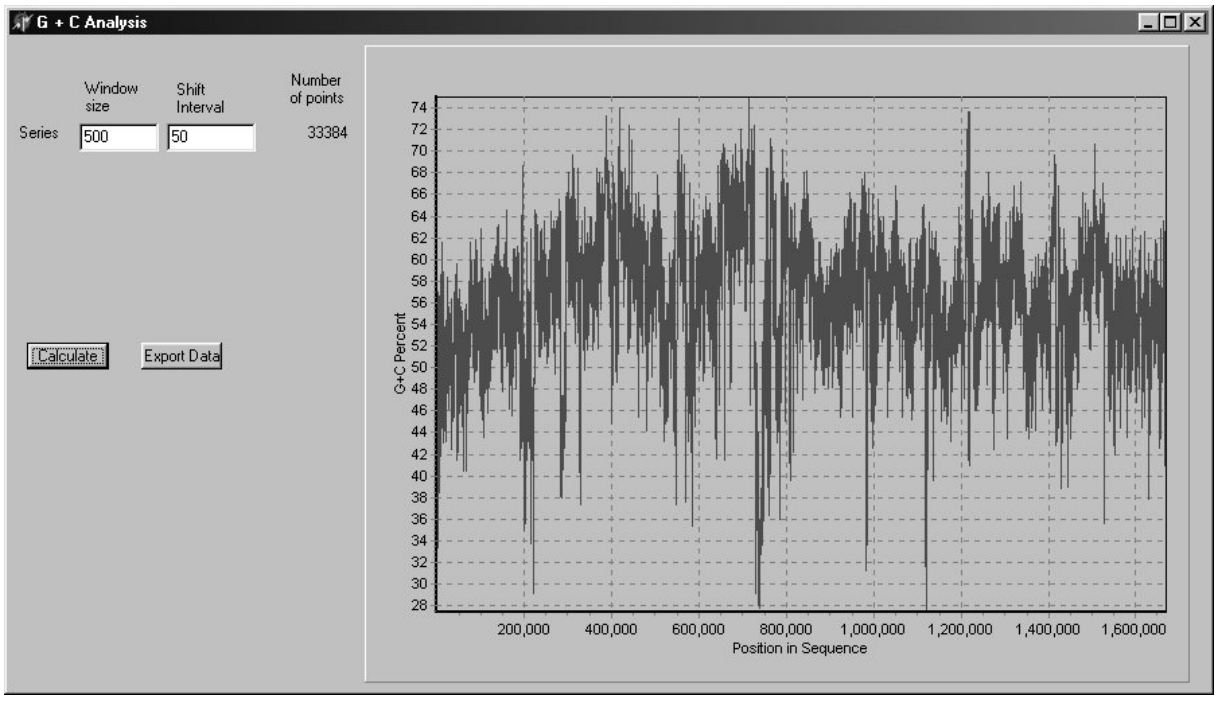

Figure 2. The G+C Analysis user interface of the WDA showing the input boxes for window size and shift on the left and the plot of G+C% versus Position in Sequence

4.2 USER INTERFACE

Figure 1 shows the main user interface in the three different states of user interaction. The first screen gives the total length (number of base pairs) of the imported sequence as well as the average G+C% for the sequence. If the sequence is a complete genome, this number should be the same as that published for the genome. The second view shows the four options available in the File menu. The 'Open GenBank Format' is used to import the GenBank format (see Table 1) of genome data. This is the preferred format as it is annotated with information such as the position and size of identified genes in the header of the file (see Table 2).

An alternative format that is available for importing DNA sequence data is called FASTA. This format has no annotated information and so is not as useful as the GenBank format. However the WDA provides an option for importing this data type in case this is the only format available. The 'Export to csv file' option exports the raw DNA sequence in comma separated value (CSV) format to allow other software tools to import a CSV file if required. The 'Create Random File' option creates a random sequence (length specified by the user) of letters T, A, G and C which can then be analysed by the software.

Figure 2 shows the user interface that is displayed when the G+C% option is chosen from the Analyse menu. The user is able to enter the window size and shift size before pressing the calculate button. The shift size is the number of letters in the sequence that the array pointer is incremented before the next G+C% is calculated. Once the G+C% has been calculated for all points, the results are plotted on screen. The 'Export Data' button provides the facility to export the plotted data in csv format for further analysis and manipulation e.g. using MSExcel. The 'Gene Count' option from the Analyse menu calculates the size of each gene and its G+C% value and the 'Gene Count export to CSV' option creates a CSV file with the start and end position of each gene along with the gene size and G+C% value.

5. RESULTS OF GENOME ANALYSIS

The WDA was used to analyse the genome of the archaeaon *Aeropyrum pernix* K1 (Kawarabayasi

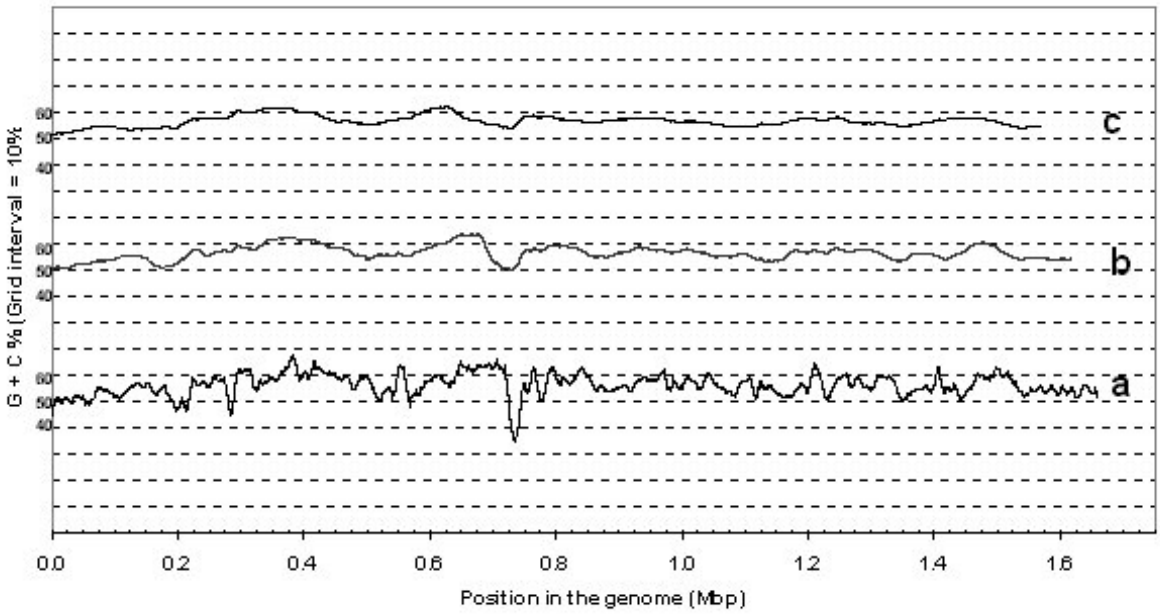

Figure 3. G+ C % as a function of position in *Aeropyrun pernix* K1 genome. a) window size = 10 000 bp, b) window size = 50 000 bp, c) window size = 100 000 bp. Window shift = 100 bp for all plots

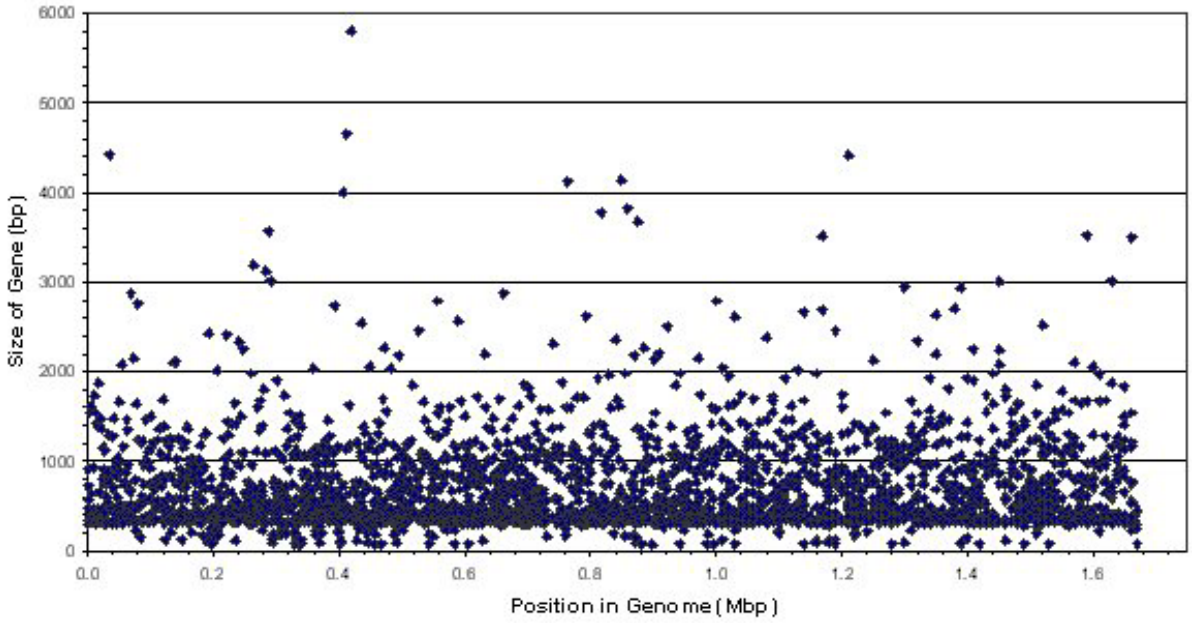

Figure 4. Size of each gene versus position in the in *Aeropyrum pernix* K1 genome

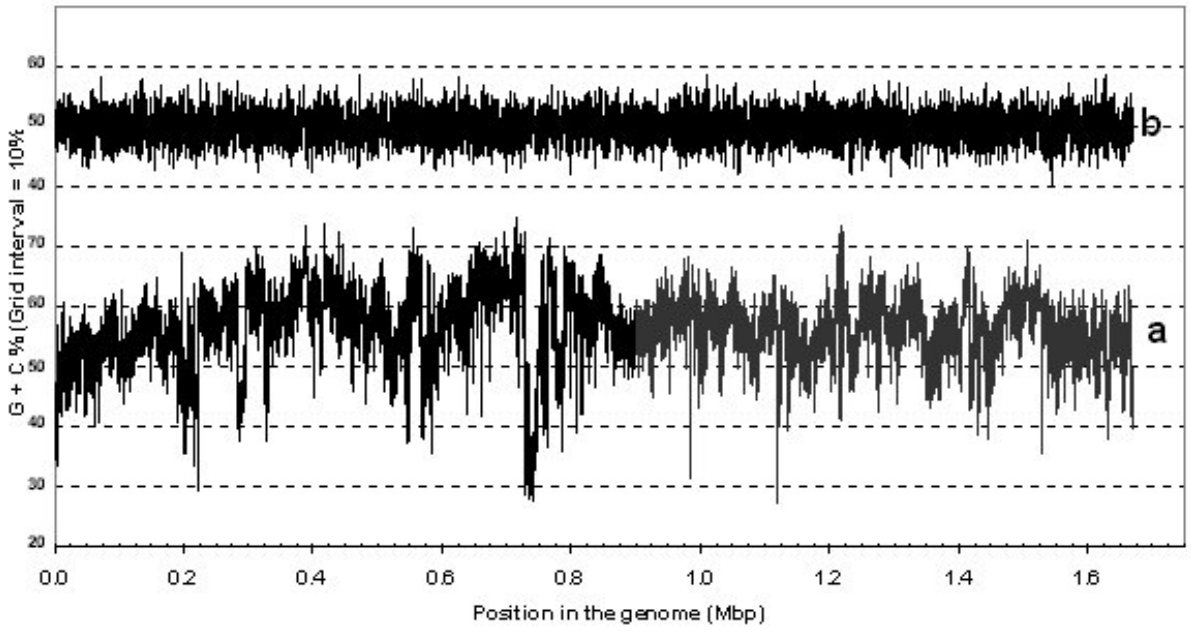

Figure 5. G+C % as a function of position for a window size = 500 bp and a window shift = 30 bp. a) *Aeropyrum pernix*K1 genome, b) Random generated sequence

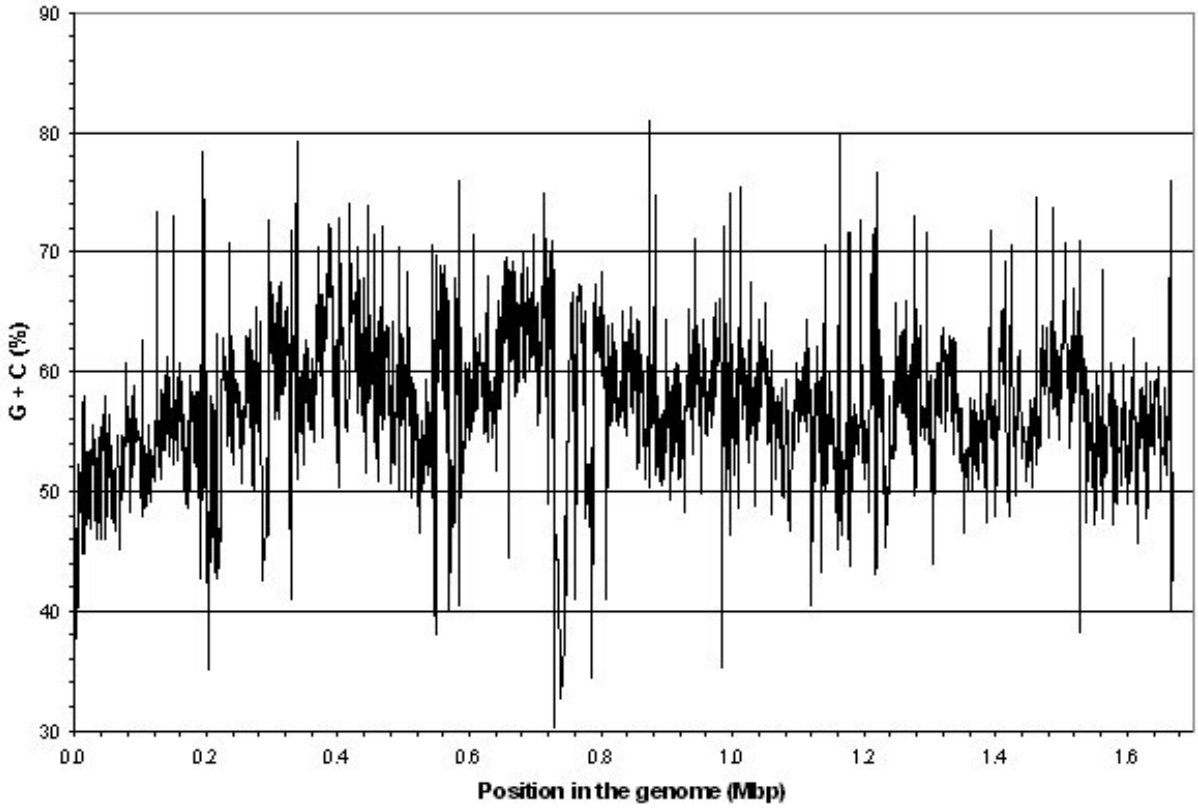

Figure 6. G+C % of Genes as a function of position in genome

1999). This genome was chosen as it has only recently been published and it is relatively small (1.6 Mbp). Figure 3 shows the G+C% plots for window sizes of 10 000 bp, 50 000 bp and 100 000 bp with a shift of 100 bp. The figure shows a smoothing of variation in data as the window size is increased. The choice of window size is therefore important to the information that can be gained for the data. The WDA provides the facility to calculate the size of the genes that have already been isolated in the genome. Figure 4 shows a display of the size of the genes as a function of position in the genome. It is clear from the graph that the bulk of genes so far discovered in this genome are in the range of 300 to 2000 bp in length. There is a very strong 'cut off' in gene size of around 300 bp below which the number of genes is markedly reduced. This may be a product of the method of 'discovering' the genes that Kawarabayasi *et al.* (1999) used. However it is obvious that the use of window sizes in the 10 000 bp to 100 000 bp range could result in missing the majority of genes in this genome: the largest gene so far discovered is only 5800 bp in length. The window size was therefore reduced to 100 bp with a shift of 30 bp and the results are compared with a random data set, Figure 5. The use of a random sequence as a comparison demonstrates that the plot of the actual genome is displaying 'real' structure.

From a biologists view, the choice of window and shift size is purely artificial. A more significant start and stop point for calculating each G+C% is to calculate the G+C% for each identified gene. This has been done for the data plotted in Figure 6. [Figure 6. G+C % of Genes as a function of position in genome. TimHuntFig6.jpg] The need for choosing the window and shift sizes has now been eliminated as 'window' size is now exactly the same size as each gene.

5.1 COMPARISON OF THE WDA WITH OTHER TOOLS

Figure 7 shows a comparison of the G+C% of *Aeropyrum pernix* K1 using the WDA, the TIGR tool

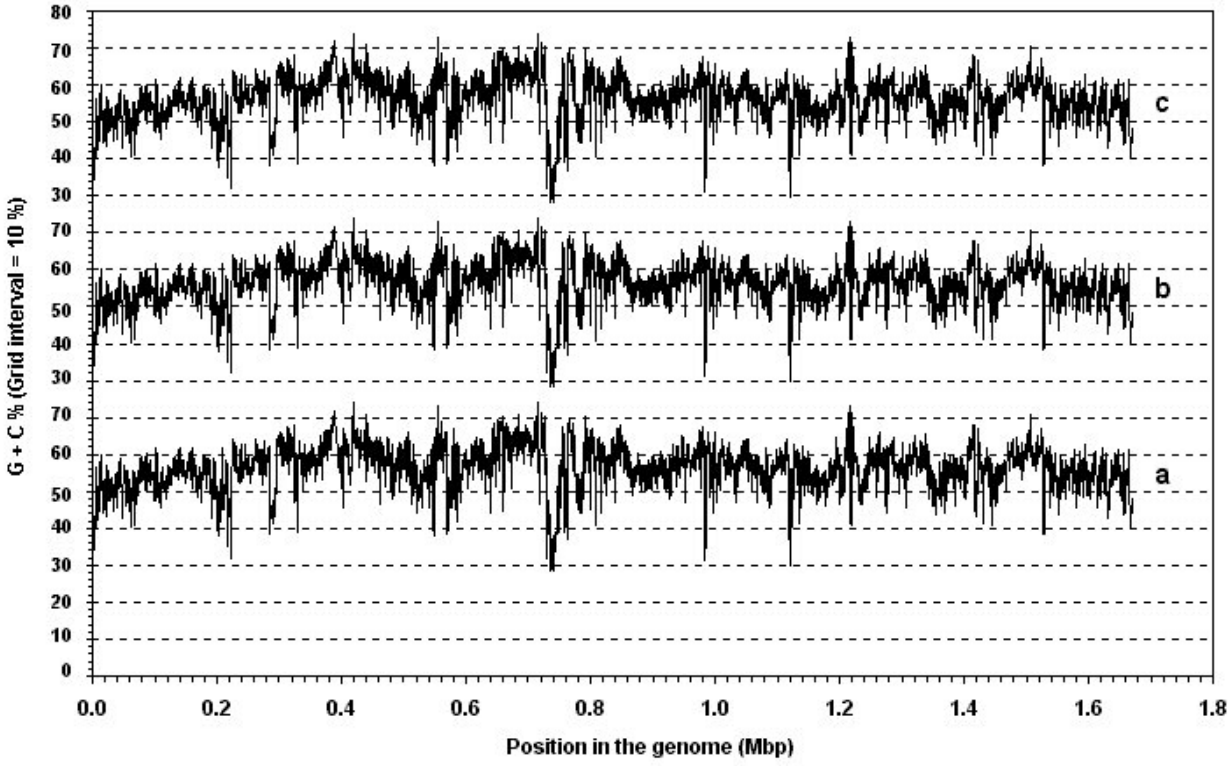

Figure 7. G+C % as a function of position in the *Aeropyrum pernix* K1 genome. a) the WDA, b) the TIGR tool and c) the DEA analyse tool. The window size and shift size for each graph is 500 bp

and the DEA analyse tool as discussed in section 3. The window size and shift were both set at the same value to allow comparison with the TIGR tool that does not allow separate window and shift values. All three tools give exactly the same values of G+C%, which is to be expected when they are analysing the same genome.

5.2 G+C% FEATURES OF THE OF *AEROPYRUM PERNIX* K1 GENOME.

The paper by Kawarabayasi et al. (1999) reporting on the results of sequencing of *Aeropyrum pernix* K1 does not contain G+C% analysis, except for giving the overall value of 56.3%. Faguy and Doolittle (1999) comment that the paper 'leaves the field wide open for speculators such as ourselves'. However, they also do not report on any G+C% analysis. Although G+C% of *Aeropyrum pernix* is given at: http://www.tigr.org/tigr-scripts/CMR2/GCDisplay.spl?asmbl_id=49 the results do not appear to have been published in the normal literature, and no discussion can be found regarding areas of the genome which have a relatively low G+C%. The predominant feature is in the region centered at 739 750 (window of 500 bp) which has a G+C% of 28.4%. This corresponds to the gene APE1186 identified at 739270..740634. This gene is within the region from nucleotide 725000 to nucleotide 750000 which has a significantly lower G+C% than the average with the G+C% dropping abruptly at 725000 from a G+C% of 70% to a G+C% 0f 28.4%. Use of the *Aeropyrum pernix* K1 genome data at the NCBI site (www.ncbi.nlm.nih.gov) allowed us to analyse the DNA strand bias and function of the open reading frames (ORFs) in this low G+C region. ORFs are the protein coding sequence of each gene and in this region 20/25 (80%) are found on the minus strand whereas of the next 25 ORFs only 12/25 (48%) were found on the minus strand as is found for the genome as a whole. In addition, of the 11/25 genes in the low G+C region whose function

is known the majority encode proteins that have a putative role in carbohydrate metabolism.

6. FUTURE WORK

6.1 TOOL DEVELOPMENT

The modular approach to building the WDA lends it to further development by the addition of additional components. A number of alternative algorithms for the analysis of DNA data have been published which provide opportunity for future work. A common method of sequence analysis is to plot the dinucleotide or trinucleotide relative abundance values (e.g. Karlin *et al.* 1998). The GC skew i.e. (G+C)/(G-C) has been used by Lio and Vannucci (2000) to describe the G+C analysis of genome data using wavelets to smooth the profiles and subsequent 'scalogram' application.

6.2 SIGNIFICANCE OF LOW G+C% AREAS IN THE *AEROPYRUM PERNIX* K1 GENOME

There are a number of regions in the *Aeropyrum pernix* K1 genome that show significant deviation in the G+C% value. In particular the region centered on 739 750 bp remains a region of low G+C% when analysed using a wide range of window sizes. The asymmetric strand distribution for ORFs APE 1169 to APE 1196 is unusual in this genome and might be a function of the lower than average G+C% of this genomic region. Of these 25 ORFs 14/25 (56%) were described only as a hypothetical protein because BLAST matches with proteins of known function could not be made. Of the eleven whose function had been predicted the majority had putative roles in carbohydrate metabolism. It is acknowledged that metabolic genes are more susceptible to horizontal transfer because of their presence in operons and because the proteins that they encode do not usually form part of a complex Koonin *et al*. (2001). This data supports our contention, based on aberrant strand distribution and low G+C%, that these genes have been horizontally transferred to the *Aeropyrum pernix genome* K1. This could be tested further by an analysis of codon usage for the ORFs in this region and an analysis of the regulatory regions and transcriptional profile of these genes.

7. DISCUSSION AND CONCLUSION

The WDA has shown to be a useful tool for the investigation of recently published genome data. The main feature of the G+C% analysis of the Aeropyrum pernix K1 genome is the dip in G+C% centred around the 0.74 Mbp region. This feature remains observable no matter what combination of window and shift sizes are used. There are many other regions that display similar 'resilience' to choice of window and shift size. However, this work confirms the observation made by Lio and Vannucci (2000) that the choice of window size is an important variable for G+C% analysis. This work has shown that although compositional changes spanning large regions of the genome are robust to window size choice, there are many small regions that have significant G+C% variation that are not observed if the window size is too large. A window size similar to the size of gene should be chosen to avoid missing important variations of significance.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol Oct 5;215(3), 403-10.

Bernardi, G., Hughes, S. and Mouchiroud, D. (1997). The Major Compositional Transitions in the Vertebrate Genome' Journal of Molecular Evolution, 44(Suppl 1), S44-S51.

Cruveiller, S., D'Onofrio, G. and Bernardi. G., (2000). The compositional transition between the genomes of cold- and warm-blooded vertebrates: codon frequencies in orthologous genes. Gene 261, 71-83.

Faguy, D. and Doolittle, W., (1999). Genomics: Lessons from the *Aeropyrum pernix* K1 genome. Current Biology 9, 23, Dispatch R883-R886.

Karlin, S., Mrazek, J., Campbell, A.M., (1997). Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol, Jun;179(12), 3899-913.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. Current opinion in microbiology 1(5), 598-610.

Karlin, S., Campbel, A. and Mrazek, J., (1998). Annu. Rev. Genet. 32, 185-225.

Karlin, S. (1999). Bacterial DNA strand compositional asymmetry. Trends Microbiol, Aug;7(8), 305-8.

Kawarabayasi, Y. *et al.*, (1999). Complete Genome Sequence of an Aerobic Hyper-thermophilic Crenarchaeon, Aeropyrum pernix K1. DNA Research 6, 83-101.

Koonin, E.V. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55, 709-742

Lio, P. and Vannucci, M. (2000). Finding pathogenicity islands and gene transfer events in genome data. BioInformatics Vol. 16 no 102000, 932-940.

Miklos, G.L.G (1985) in Molecular Evolutionary Genetics, 241-321.

Venter, C. *et al*, (2001). The Sequence of the Human Genome. Science Vol 291, 1304-1351 2001.

Tamaru, H. and Selker E.U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. Nature 414, 277-283

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P. and Hayes W.S. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coil*. Curr. Biol. 6, 279-291

Vaslag, B. (2001). At the cloning circus sideshows abound, while scientists seek a wider audience. J. Amer. Med Assoc. 286,1437-1442

Woese, C.R., Kandler, O. and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eukarya. Proc. Natl. Acad Sci. USA. 87, 4576-4579