

A rapid method of whole genome visualisation illustrating features in both coding and non-coding regions

R. Hall

L. Stern

Dept. of Computer Science and Software Engineering
The University of Melbourne, Melbourne, Victoria 3010
Australia, e-mail rshall@cs.mu.oz.au

Abstract

The application of Fourier analysis to a genome can be used as an indicator of gene coding regions. We have developed a visualisation of the Fourier spectra that allows convenient whole chromosome scanning for genes and other features. The method's rapid operation suits its application as a first pass analysis. Fourier analysis indicates a strong periodicity of 3 in coding regions of several different organisms and is independent of the orientation of the gene. A bitmap display of the Fourier spectra over a sliding window gives rapid visualisation and localisation of coding regions in the chromosomes of a number of different organisms. Non-coding features such as regions of repetitive DNA, are visualised at the same time. The method works particularly well on organisms with a skewed base composition such as the malaria parasite *Plasmodium falciparum* and the protozoan *Leishmania major*.

Keywords: Fourier analysis; *Plasmodium falciparum*; Repeats; genome visualisation.

1 Introduction

The frequency content of a signal can be determined by the analysis of its Fourier transform (O'Neil 1991). When DNA is viewed as a signal, a discrete Fourier transform provides a method suited to the detection of periodic arrangements of bases in a genome. Before a spectral analysis of a genomic sequence can be performed, it must be converted from a string of four component bases to a numerical array. The numerical representation determines which features of the genome are highlighted by the analysis. This translation can be performed in a number of ways. Silverman and Linsker (1986) represent each base as the vertex of a tetrahedron in three dimensional space and the genome sequence is transformed into an array composed of three dimensional vectors. A Fourier transform is performed on each of the three sequences made up of a directional component from the sequence vectors. The resulting spectrum is the sum of the three Fourier transforms. Tiwari *et al.* (1997) use four binary strings to represent the occurrence of each base in the nucleotide sequence, summing the individual spectra to give an overall sample spectrum for the genome. Using this method, a repeating period of 3 has been found in coding regions (Fickett & Tung 1992, Tiwari *et al.* 1997). Fourier analysis has also been employed as one of a number of weighting factors in the determination of intron splice sites (Huestis & Saul 2001).

Copyright ©2004, Australian Computer Society, Inc. This paper appeared at The Second Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand. Conferences in Research and Practice in Information Technology, Vol. 29. Yi-Ping Phoebe Chen, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

For a clear Fourier spectrum to be generated, a reasonable sized sample is required. Larger samples give relatively poor resolution, making it difficult to locate features exactly using only the Fourier transform. This paper shows that an appropriate visual representation of the Fourier analysis can provide a guide to gene location and other periodic features in the genome.

2 Algorithm

A numerical representation of a sample of DNA is constructed for each of the four nucleotides. The four arrays are implemented using binary strings as described by Tiwari *et al.* (1997), where each occurrence of the relevant base is indicated by a 1 and any other base, by a 0. For example, a short sequence such as ACTGAGCTA is transformed into the four strings: 100010001 referring to the A nucleotide, 010000100 for C, 000101000 for G and 001000010 for T. A Fourier transform was performed on each array, giving a spectrum which indicates if that particular base is appearing periodically in the DNA sample. The sum of the squares of the individual spectrum components is taken to produce an overall Fourier analysis for the particular sample of the genome. The algorithm was implemented in C.

An overall view of a genomic sequence is constructed by taking sequential, overlapping samples and performing a combined Fourier analysis on each. The samples are overlapped to detect periodicities spanning two adjacent samples. The spectrum from each analysis is scaled by a constant factor and converted to a column of grey-scale pixels. The resulting bitmap has a width of the genome length divided by half the sample size and a depth of half the sample size (only the first half of the Fourier transform being significant).

The line of pixels representing each Fourier sum is scaled and transformed such that a high power in the spectrum is represented by a dark pixel, while low power values are assigned light pixels. The scaling factor provides a means of enhancing lower peaks as all peaks over an arbitrary value can be marked black. This allows spectrum powers which are only marginally above the background noise to be assigned the same visual significance as major peaks. Due to the sequential sampling, small peaks over a number of contiguous samples produce a distinct line in the bitmap. For example using a sample of 256 bases, a continuous horizontal line at a pixel column height of 85 indicates a strong period 3 in these samples ($256/85 = 3$).

Using a sample size of 512 bases, a one million base genome produces a grey-scale bitmap in the PGM format 256 pixels x 3900 pixels in about 20 seconds on a 333 MHz Sun Ultra Sparc. The convenient size means that multiple bitmaps can be viewed and compared using standard graphics software. An X Window program is in development which will allow convenient scrolling of the bitmap image along with scales giving sequence offsets and periodicity values.

3 Experimental

The Fourier analysis and visualisation was used to examine DNA sequences from a number of organisms.

3.1 Structure in Non-Coding Regions

3.1.1 *Plasmodium falciparum*

The genome of the malaria parasite *Plasmodium falciparum* is about 30 million base pairs in size and is made up of 14 chromosomes. The chromosomes range from 650 kb to 3400 kb in length. The genome has a high proportion of the adenine/thymine bases (82% in chromosome 2). An annotation of chromosome 2 was published in 1998 (Gardner *et al.* 1998), and is still in the process of being confirmed (Gardner 1998, Gardner *et al.* 1999, Huestis & Fischer 2001, Huestis & Saul 2001).

Figure 1 shows the Fourier analyses of the first 100 kb of *P. falciparum* chromosome 2 represented as a bitmap, using a sample size of 512 bases. This sample size proved to be the best compromise between resolving short sequences with periodic features and providing a strong spectrum. Annotated exons (Gardner *et al.* 1998) are shown at the top of the figure for comparison.

Numerous areas with strong periodic properties are visible as horizontal lines. Referring to the annotated exons, periodicity is apparent in both coding and non-coding regions. In chromosome 2, a section with a period of 7 is visible in the 800 bases at the end of the chromosome. Also seen in figure 1 is a region from 1 kb to 2 kb with a period of 14. A strong period 12 feature is visible from offset 10.2 kb to 12 kb, followed by a 10 kb section with a prominent period 21 repeat. These features in the first 20 kb of chromosome 2 correspond to known repeating regions in non-coding DNA. For example, the 10 kb section from 12 kb to 22 kb corresponds to the known 21 base pair repeat region *rep20* (Gardner *et al.* 1998). A less pronounced repeat with a period of 128 at 2.8 kb to 4.6 kb is visible at the higher resolution available on the screen. All the repeat regions observed correlate with previously described areas of high compressibility (low complexity) (Stern *et al.* 2001). Repeating regions have a characteristic “laddered” appearance due to the harmonics of the Fourier transform (O’Neil 1991).

3.2 Structure in Coding Regions

3.2.1 *Plasmodium falciparum*

Looking again at *Plasmodium falciparum* chromosome 2 (Figure 1), a well defined broken line is seen at pixel column height 171. This denotes a strong repeating signal with a period of 3 ($512/171 = 3$). The period 3 lines correlate well with exon locations as described in Gardner *et al.* (1998) shown above the bitmap, although the locations of the beginnings and the ends are not always exact.

Two repeat structures with a laddered appearance appear in a coding region of chromosome 2, at 92–98 kb. They lie within the annotated gene PFB0095c which encodes the erythrocyte membrane protein PfEMP3 (Gardner *et al.* 1998). PfEMP3 is known to have a section of repeating amino acids (Pasloske *et al.* 1993).

3.2.2 Other Organisms

Leishmania are protozoans with 36 chromosomes, a genome size of about 34 million base pairs and a base composition of 62% GC. A Fourier bitmap was generated from *Leishmania major* chromosome 1 and the period 3 line correlated with the annotation of Myler *et al.* (1999) (data not shown). Because of *L. major*’s higher coding density, the coding region ends were less clearly defined than with *P. falciparum*.

Streptomyces coelicolor is a soil-dwelling, filamentous bacterium. The genome is in the order of 8.6 million base pairs in length and has a GC bias of 72.12% (Bentley *et al.* 2002). *Escherichia coli* is a well known laboratory organism with a GC content of 50.4%. Both these bacteria showed an almost continuous period 3 line, consistent with their high coding density. For *E. coli* the line was not overly strong. In both cases the breaks in the period 3 line were poorly defined, because of the high coding density.

Looking at higher organisms, Fourier bitmapping was also applied to the set of sequences ALLSEQ, compiled by Burset and Guigo (1996) for the benchmarking of gene finders. The set of 500 genes (overall GC 49%) were concatenated and then analysed to form a single bitmap using a sample size of 256 bases. Using the Fourier bitmap, lines indicating strong period 3 evidence were far less numerous than the number of known coding regions, with approximately 30 clearly defined areas of strong periodicity. Figure 3 shows three genes with a distinctive period 3 for the entire length of the annotated exon. In addition to the strong period 3 line, these three genes show multiple laddered structures typical of repetitive regions.

Sources of structure in coding regions

It was of interest to investigate the base distributions responsible for the observed periodicity. Analysis of the vectors for individual bases in the first 100 kb of *P. falciparum* (figure 4) shows that the C and G analyses produce the dominant periodicity of three, with minor contributions from A and T.

Two genes in the region were selected for further study, PFB0010w (25.232 kb to 31.168 kb) and PFB0065w (66.529 kb to 67.545 kb). Detailed analysis showed that the period 3 signal was not constant along the length of these genes. Looking at individual bases within a region of high signal strength, the distribution of each base over the three codon positions was determined (Table 1).

For each of these two genes the period 3 line on the combined spectrum appears to originate from a different kind of periodicity within the gene. PFB0010w gives a strong indication on the G and T bitmaps. Sampling a region of high periodicity 3 shows codon position three has a high proportion of T. The concentration of G is only moderately dominant at codon position one, but is enough to register on the bitmap. Although the percentages for C are almost the same as G, the C bitmap shows no period 3 signal. A possible explanation is that the codon position two C bases are evenly distributed throughout the sample while the codon position one G bases are grouped together. PFB0065w is evident on the G Fourier bitmap. Sampling shows a high G in codon position one. The C spectrum is also reasonably strong.

In order to investigate the role of codon bias in periodicities, a section of a *P. falciparum* gene’s amino acid sequence was used in one experiment to generate codons with a high GC bias and in another to generate random codons. The period 3 line was still evident in both cases.

For *Leishmania major* the strongest and most consistent period 3 signal was generated by A, which is extremely sparse in the third codon position. The genes shown from the ALLSEQ data set all had a reduced proportion of T, especially in codon positions one and three, and a high proportion of A in codon position two. The two trichohyalin genes both have a high proportion of C in position one and G in position three.

4 Discussion

Fourier analysis has been used previously for gene finding by Tiwari *et al.* (1997). We have found that although the Fourier method gives a good indication of coding regions, especially with some organisms, its resolution is not fine

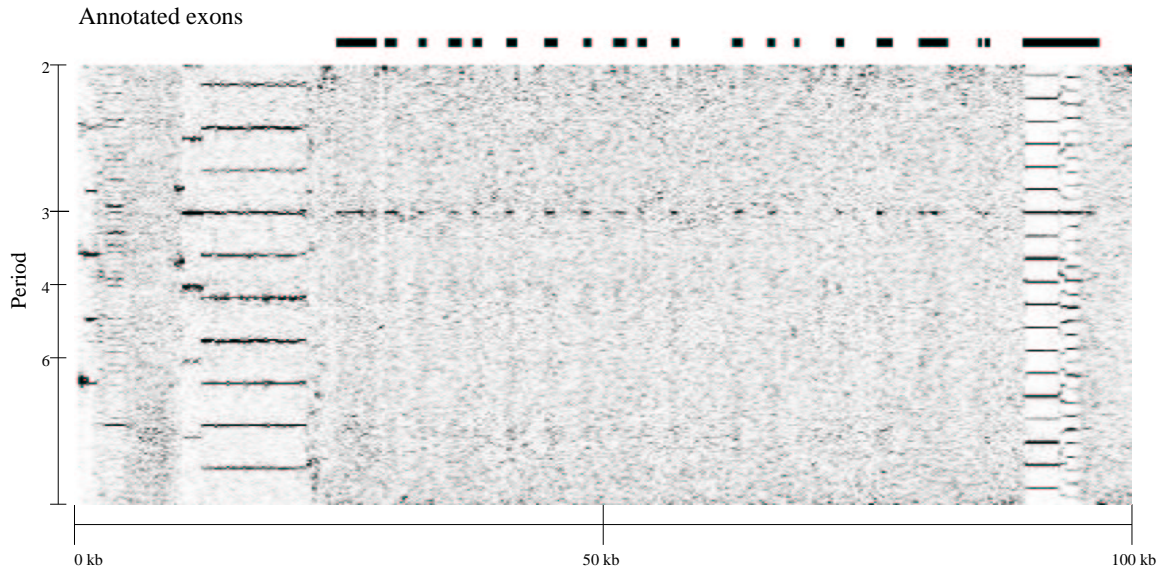


Figure 1: *P. falciparum* chromosome 2: Fourier analysis bitmap of the first 100 kb. The heavy dotted line above the bitmap indicates the location of exons as annotated in Gardner *et al.* (1998). Period is given as the sample size divided by the frequency of a repeating pattern.

Codon Position	PFB0010w				PFB0065w			
	%A	%C	%G	%T	%A	%C	%G	%T
1	39	32	57	18	39	19	64	24
2	41	55	31	22	18	63	25	29
3	20	13	12	60	43	18	11	47

Table 1: Percentage distribution of each base over the three codon positions in a sample of 255 bases. Samples are taken from high period 3 regions of *P. falciparum* chromosome 2, positions 30.3 kb (PFB0010w) and 67.2 kb (PFB0065w).

enough to show exon boundaries clearly even when an incremental sliding window is used. Since Fourier analysis is limited by the amount of periodic data which can be extracted from a genome, we have chosen to use a limited sliding window. This results in a bitmap of large genomic regions that is small enough to provide an overall view, with no further loss of resolution.

The strength of the periodicity 3 signal varies over coding regions. For visual ease we adjusted the threshold assignment of pixel values, so that a consistent line is seen over the length of the coding region. Because of the visual nature of the bitmap, a series of low Fourier signals just above the background noise is discernible as a line to the observer when thresholding is used.

Anastassiou (2000) has used a color bitmap representation of a Fourier analysis on short sections of the *S. cerevisiae* genome, employing color to show phase information indicating reading frames. In our experiments with *P. falciparum* the phase angle was not consistent enough to give an indication of the reading frame.

Our analysis of a number of *P. falciparum* coding regions showed a high G prevalence in codon position one and this concurs with the coding distribution described by Saul and Battistutta (1988). The concentration of G and C in the coding regions and the tendency for their proportion to peak at a single codon position enhances the period 3 signal. Highly structured areas in noncoding regions were also observed in *P. falciparum*, characterised by a ladder appearance.

Our experiments with codon substitution suggest that the period 3 signal is a function of the amino acid se-

quence, rather than of the codon bases. This is consistent with the findings of Tiwari *et al.* (1997). We have found that although a biased base composition is not a necessary condition for detecting periodicity 3, it enhances the signal and allows a clearer delineation of the coding and non-coding regions.

5 Conclusion

Fourier bitmapping provides a convenient and rapid overview of a genome. Within a few seconds, the genome of an organism can be assessed for coding density, approximate gene locations and repeat regions. In most of the organisms studied, Fourier bitmapping produces a period 3 line which correlates with annotated coding regions and is particularly prominent in genomes with a biased base composition.

5.1 Acknowledgments

Sequences for *P. falciparum*, *L. major*, *S. coelicolor* and *E. coli* were obtained from the National Center for Biotechnology Information (www.ncbi.nlm.gov), National Library of Medicine, National Institute of Health (USA). An updated *P. falciparum* annotation was obtained from The Walter and Eliza Hall Institute (<http://www.wehi.edu.au/MalDB-www/ae1362ra.htm>), Melbourne, Australia. We would like to thank Robert Huestis for many useful suggestions and Lloyd Allison for critical reading of an early draft of this manuscript.

References

- Anastassiou, D. (2000), 'Frequency-domain analysis of biomolecular sequences', *Bioinformatics* **16**, 1073–1081.
- Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., Hornsby, T., Howarth, S., Huang, C.-H., Kieser, T., Larke, L., Murphy, L., Oliver, K., O'Neil, S., Rabbinowitsch, E., Rajandream, M.-A., Rutherford, K., Rutter, S., Seeger, K., Saunders, D., Sharp, S., Squares, R., Squares, S., Taylor, K., Warren, T., Wietzorrek, A., Woodward, J., Barrell, B. G., Parkhill, J. & Hopwood, D. A. (2002), 'Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)', *Nature* **417**, 141–147.
- Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. A., Devlin, K., Feltwell, T., Gwilliam, S. G. R., Hamlin, N., Harris, D., Holroyd, S., Hornsby, T., Horrocks, P., Jagels, K., Jassal, B., Kyes, S., McLean, J., s. Moule, Mungall, K., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutter, S., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Whitehead, S., Woodward, J. R., Newbold, C. & Barrell, B. G. (1999), 'The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*', *Nature* **400**, 532–538.
- Burset, M. & Guigo, R. (1996), 'Evaluation of gene structure prediction programs', *Genomics* **34**, 353–367.
- Gardner, M. J. (1998), 'Direct submission'. Submitted (02-NOV-1998) Institute of Genomic Research, Naval Medical Research Center, 9712 Medical Center Drive, Rockville, MD 20814, USA. Email: gardner@tigr.org.
- Gardner, M. J., Tattelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shun, K., Jing, J., Aston, C., Lai, Z., Schwartz, D. C., Partea, M., Salzberg, S., Zhou, U., Sutton, G. G., Clayton, R., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. (1998), 'Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*', *Science* **282**, 1126–1132.
- Gardner, M. J., Tattelin, H., Carucci, D. J., Cummings, L. M., Smith, H. O., Fraser, C. M., Venter, J. C. & Hoffman, S. L. (1999), 'The malaria genome sequencing project: Complete sequence of *Plasmodium falciparum* chromosome 2', *Parasitologia* **41**, 69–75.
- Huestis, R. & Fischer, K. (2001), 'Prediction of many new exons and introns in *Plasmodium falciparum* chromosome 2', *Mol Biochem Parasitol* **118**, 187–199.
- Huestis, R. L. & Saul, A. (2001), 'An algorithm to predict 3' intron splice sites in *Plasmodium falciparum* genomic sequences', *Mol Biochem Parasitol* **112**, 71–77.
- O'Neil, P. V. (1991), *Advanced Engineering Mathematics*, third edn, Wadsworth Publishing Company, Belmont, California, chapter 17.13, pp. 1116–1126.
- Pasloske, B. L., Baruch, D. I., van Schravendijk, M. R., Handunnetti, S. M., Aikawa, M., Fujioka, H., Taraschi, T. F., Gormley, J. A. & Howard, R. J. (1993), 'Cloning and characterization of a *Plasmodium falciparum* gene encoding a novel high-molecular weight host membrane-associated protein, PfEMP3', *Mol Biochem Parasitol* **59**, 59–72.
- Stern, L., Allison, L., Coppel, R. L. & Dix, T. I. (2001), 'Discovering patterns in *Plasmodium falciparum* genomic DNA', *Molecular and Biochemical Parasitology* **112**, 71–77.

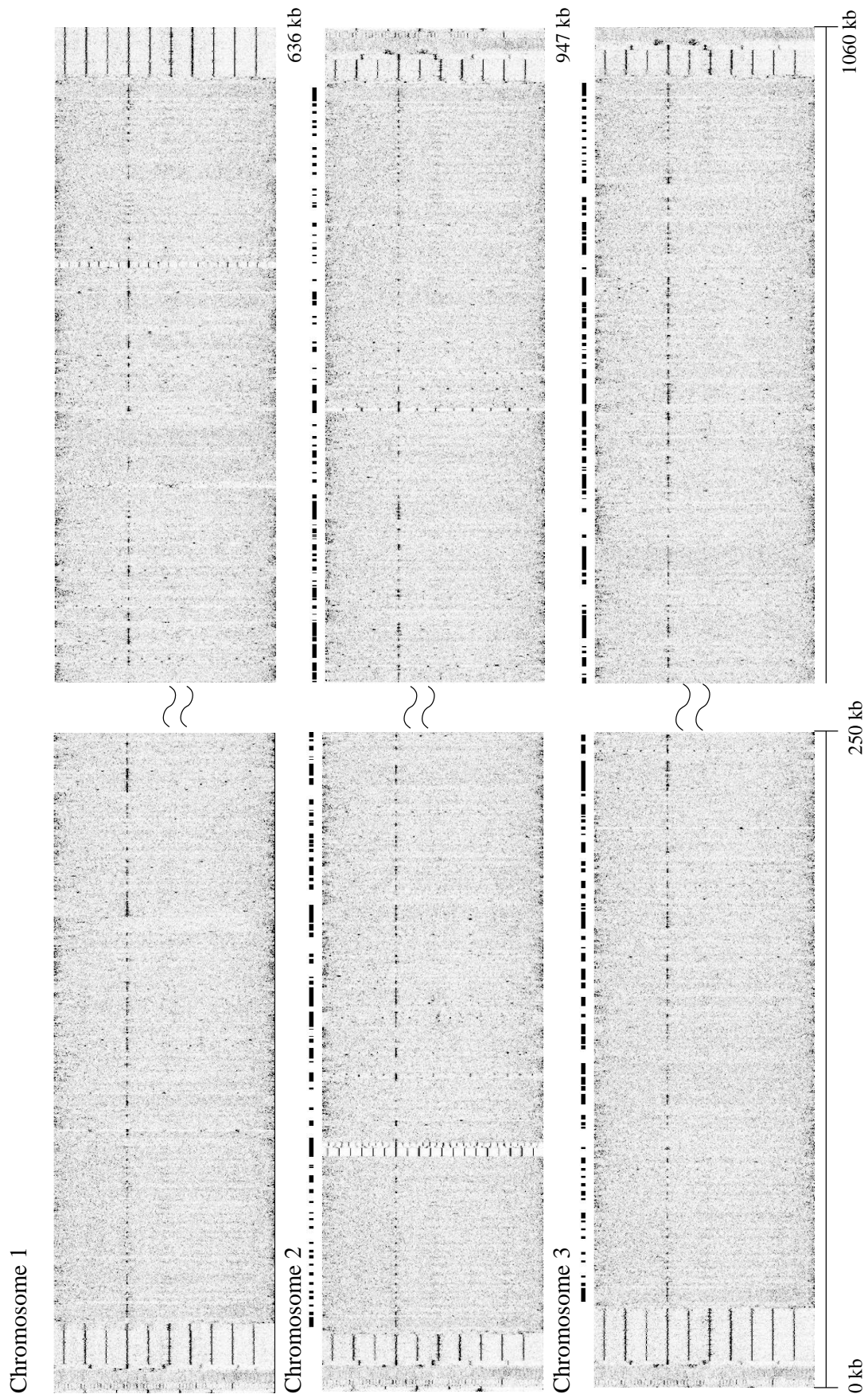


Figure 2: *P. falciparum*: First and last 250 kb of chromosomes 1,2 and 3. Black dotted line above chromosomes 2 and 3 indicate annotated exons (Gardner et al. 1998, Bowman et al. 1999).

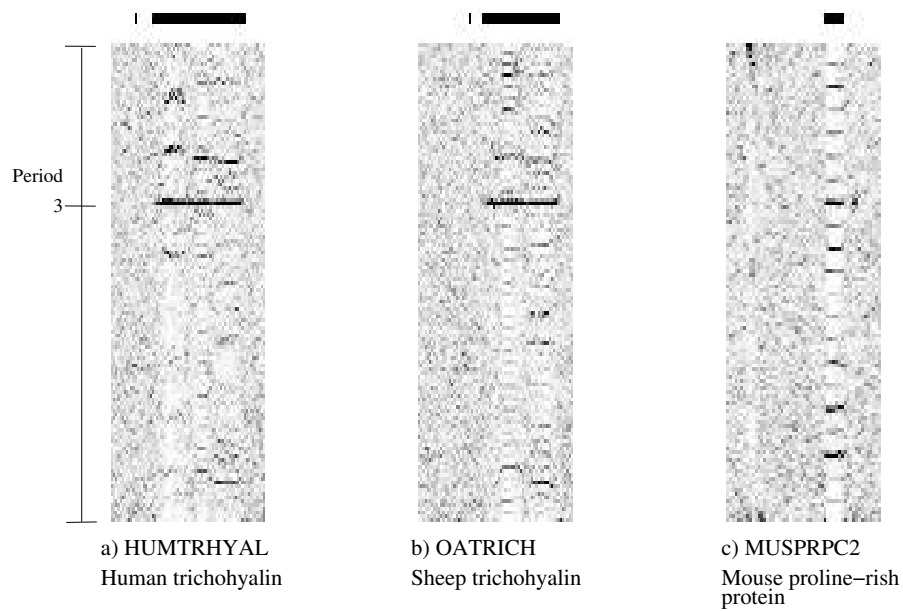


Figure 3: Genes from the ALLSEQ data set. Dark lines above bitmaps indicate annotated exons (Burset & Guigo 1996).

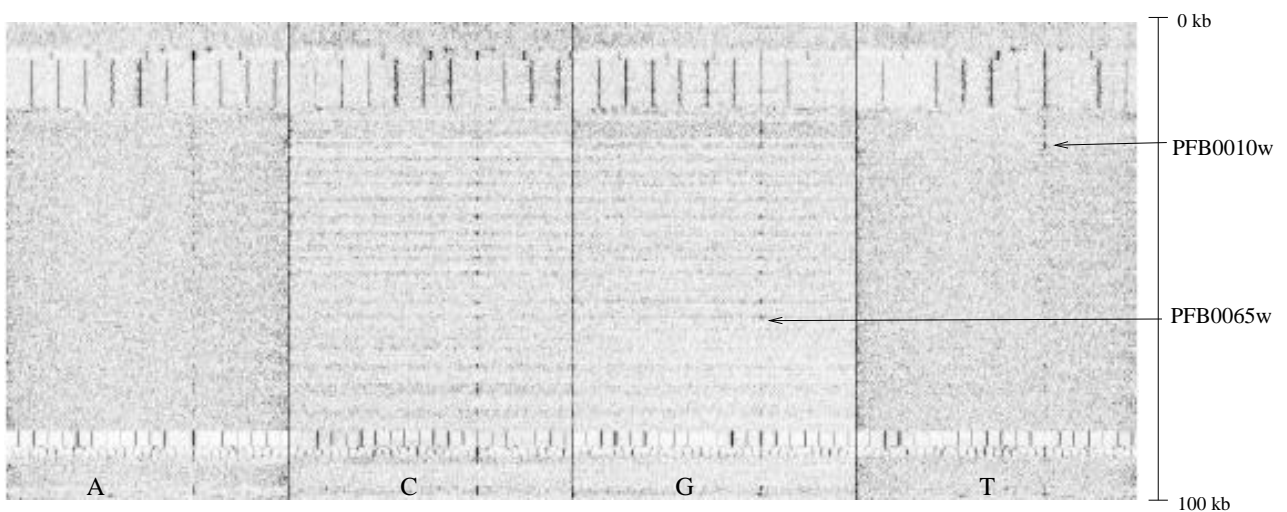


Figure 4: Bitmap representation of the Fourier analysis for each base. The first 100 kb of *P. falciparum* shown.