Atypical Regions in Large Genomic DNA Sequences

Stewart Scherer*[1], Mary Sara McPeek†[2], and Terence P. Speed†

* Human Genome Center

Lawrence Berkeley Laboratory 74–157

Berkely, CA 94720

†Department of Statistics

University of California

Berkeley, CA 94720

Classification: Genetics

Key Words: DNA sequence analysis, Markov chain, chromosome structure

[1] Corresponding author. Current address: Microbiology Department, University of Minnesota School of Medicine, Box 196 UMHC, Minneapolis, MN 55455

[2] Current address: Department of Statistics, University of Chicago, Chicago, IL 60637

ABSTRACT

Large genomic DNA sequences contain regions with distinctive patterns of sequence organization. We describe a method using logarithms of probabilities based on 7th-order Markov chains to rapidly identify genomic sequences that do not resemble models of genome organization built from compilations of octanucleotide usage. Databases have been constructed from *E. coli* and *S. cerevisiae* DNA sequences greater than 1000 nucleotides and human sequences greater than 10000 nucleotides. Octanucleotide counts from a single large genomic sequence can be used in certain analyses. Atypical genes and clusters of genes have been located in bacteriophage, yeast, and primate DNA sequences. We consider criteria for statistical significance of the results, offer possible explanations for the observed variation in genome organization, and give additional applications of these methods in DNA sequence analysis.

INTRODUCTION

Large contiguous genomic DNA sequences have been determined for a number of species. The human genome project and similar efforts for important model systems will be producing such sequences at an increasing rate. The most widely used computational technique for the analysis of new sequence information is the comparison of the sequence with all other known sequences for regions of similarity. With the growth of the public databases, this approach has become increasingly successful with about one-third of all large open reading frames in new genomic sequences or new cDNA sequences being related to known genes (1, 2). The genome projects will have a major impact on this approach because a decreasing fraction of database entries will be sequences of known function with the wealth of biological information associated with them.

Existing computational tools for the analysis of DNA sequences were generally developed for the analysis of individual genes. Most sequence analysis software reports tables of regions of interest based on a scoring system. This tabular presentation becomes unwieldy when faced with sequences of tens or hundreds of kilobases.

Statistical analyses of DNA sequences have been used since the origins of molecular biology. It was first noted that nearest neighbor frequences of different species varied greatly and these differences could not be explained by the known variation in base composition (3). With the advent of efficient DNA sequencing techniques, it became clear that there were wide variations in codon usage (4), tetranucleotide (5), and hexanucleotide (6) frequencies among species. The increase in available sequence data as a result of the genome projects will permit the statistical evaluation of frequencies of longer oligonucleotides.

Building on these observations, we have developed a flexible approach to the analysis of large genomic DNA sequences based on Markov chain models. A database is built from a large collection of sequences and the frequency of occurrence, termed usage, of all possible sequences of length eight is tabulated. This length was chosen for several reasons. First, it is the largest size for which effective statistics could be developed when the complete genomes of the important model microorganisms become available. Second, it is about the smallest size where protein-DNA interactions might be reflected. Finally, this size represents 16 bits of information and offers certain computational economies. Probabilities are calculated for a window of constant length at every possible position along a query sequence and the results are presented graphically. The query can be tested against a variety of models constructed by the selection of different sets of sequences for the underlying database. The display indicates how well different regions of the sequence resemble the model.

Our statistical model assumes stationary behavior of the query sequence but makes no biological assumptions about the nature of that sequence or those used to build the database. For example, coding and non-coding regions are treated equivalently. For long genomic sequences, one would expect substantial local statistical variations. We discuss the validity of these asumptions and criteria for the determination of significance of the extreme values that are observed.

By building a database from all known sequences from a single species we have found regions in large query sequences with atypical organization in both prokaryotes and in eukaryotes. These results agree well with our understanding of DNA organization of these regions and highlight areas for future biological investigations. This approach can also rapidly provide certain information about internal repetition and other features of the query sequence. We discuss possible explanations for the presence of genes and gene clusters with unusual organization and other applications of this approach to DNA sequence analysis.

METHODS

DNA sequence data: all DNA sequences used are derived from those in Genbank release 75. Compilations of sequences from a particular species used the name in the organism field of the entry. Sequences less than 1000 bases in length (10000 for human DNA) were not included and the reverse complement of each sequence was added to the compilations.

Octamer usage statistics: the usage of all octamers $f(b_1, b_2 ... b_8)$ (b is any base, f is the tabulated count) and total usage $N = \sum_{all\ 8mers} f$ were determined. ¿From this data, conditional probabilities were calculated $P(b_8|b_1...b_7) = (f(b_1, b_2...b_8)+k)/\sum_{b_8}(f(b_1, b_2...b_8)+k)$ where $k = 0.01N/4^8$. This flattening constant k is approximately unity for the databases used below that were built from large collections of genomic sequences. See Good (7), for a discussion of this topic. Translated log base 2 of these values were determined as $Z_i = log_2 P(b_i|b_{i-7}...b_{i-1}) + 2$. Dinucleotide and trinucleotide frequencies were determined from the octamer usage. These values were similarly zero adjusted and used to determine values of Z.

Confidence limits: DNA sequences of a length equal to that of the query sequence were generated randomly using the usage of 7mers in the database to provide an initial point and using the conditional probabilities described above. 100 such sequences were used to establish 95% confidence limits for the upper limit of the plots.

Computation: custom software to construct the usage tables, evaluate the Markov statistics, and plot the data were written in the C programming language. The programs were run on Sun workstations and output Tektronix 4014 or Postscript graphics.

Graphical presentation: scores for sequence blocks of length $j + 1$ were calculated for all values of i where $i + j$ is less than the sequence length, $Y_{i,j} = -\sum_{k=i}^{k=i+j} Z_k$ was plotted at $X = i + \frac{j}{2}$. The Y axis was divided by the mean $\mu_j = (\frac{j+1}{N})\sum_{all\ 8mers} Z \cdot f$ such that the values presented are multiples of μ. The entire sequence is plotted in a single frame with the X axis being the position on the sequence. Because of variation in the lengths of the query sequences, widths of the peaks for similar sized structures vary between figures. Note that the plots are $-log_2 P$. Therefore, sequences that do not resemble the underlying model will point upwards and those overrepresented in the database will point downwards. Three horizontal lines are shown. The upper line is the position where all values would plot if the database had a uniform distribution of all sequences. The central line is the mean of the distribution, μ. If the sequence resembled the underlying model, one would expect a noise band centered on this position. The third line is twice the mean. In some figures, the position of all open reading frames beginning with a methionine codon that could translate into proteins > 100 amino acids are shown with rightward frames as the top three lines and leftward frames on the lower three lines.

RESULTS

All figures in this work present the comparison of a DNA sequence with an octamer usage database contructed from genomic DNA sequences from a particular species. Generally all known sequences from a given species were used as described in Methods. In some cases, the counts were tabulated from a single DNA sequence. Probabilites were determined for all subsequences of a given length (512 or 1024 bases) in the sequences under consideration and values for each position along the sequence are plotted as described in Methods.

Bacteriophage λ: Among large bacterial DNA sequences, phage λ is one of the most intensively characterized (8). In particular, it exhibits a modular organization with adjacent genes often having related functions (9). Several portions of the genome are known to be dispensible in the lytic phase of its life cycle and function while λ is resident as a prophage in the *E. coli* chromosome. Figure 1 presents an analysis with the entire phage genome as query and *E. coli* sequences used to construct the model. Panel A uses a 7th-order Markov chain analysis while panel B uses 1st-order Markov statistics (i.e. dinucleotide frequencies). While the overall pattern is similar, it is clear that the 1st-order chain, which is built using nearest-neighbor frequencies, does not effectively reveal the block structure present in the phage genome. Several features stand out in the plot. First, the late genes A–J clearly are representative of typical *E. coli* sequences. The first significant peak near the end of the late transcript is the *lom* gene (18695–19582, ref. 10). This gene is transcribed during lytic growth but is not essential for plaque formation. Other regions that produce the downward peaks in the right half of the genome include the recombination, replication, and lysis genes. The regions not essential for lytic growth, most notably the b2 region (leftward from 27731), deviate most from the model.

Saccharomyces chromosome 3: This sequence was the first complete eukaryotic chromosome to be determined (2). Because of the extreme length of the yeast sequences being tested, the size of the window was doubled to assure statistical significance. This also makes the remaining larger features easier to detect given the compression of the X-axis. Figure 2, panel A shows the pattern produced using the entire chromosome as query with the model built from all other yeast sequences greater than 1kb. While the extent of variation from the model is not as extensive at that observed with phage λ, there is a large region centered about coordinate 210000 that does not strongly resemble yeast sequences. Several other statistically significant variations are observed. Notable is the lowest point in the graph which is at the location of a Ty element. This is expected for sequences that are highly represented in the database.

Panel B of figure 2 shows a plot where yeast chomosome 3 is tested against an *E. coli* model. Clearly, no part of the sequence strongly resembles bacterial DNA with little variation along the sequence. None of the low points on this plot align with the peaks in panel A or panel C.

Panel C of figure 2 is a plot where the database is built solely from the actual sequence of yeast chromosome 3. Most of the sequence shows no significant variation from other sequences on the chromosome. Several features are of interest. The lowest point on the plot is again the Ty element. The other three pronounced downward peaks are *HML (12000), MAT (199000), and HMR (292000).* This indicates that any sequence representing about 1% of a database will be readily detected without resorting to a conventional homology search. The two highest points are the region around 210000 which was found not to resemble yeast sequences in general and the glucokinase gene (50000).

Figure 3 shows an enlargement of the region near 210000 on chromosome 3. The window size has been reduced to show smaller features of the plot. Note that the most atypical regions align well with the largest open reading frames of the region; however, many large open reading frames nearby such as *TSM1* (201000–204000) and *THR4* (216000) show more typical organization.

Saccharomyces chromosome 1: Among several large yeast genomic sequences examined by this approach, including a total of about 100kb from chromosome 5, the most significant variation from the general yeast model was noted for this sequence from chromosome 1 (figure 4). The highest peak in the plot again aligns with an open reading frame. While the function of this gene in the region of this peak is unknown (11), it is homologous to the hamster RCC1 locus. Given the role of this gene in the cell cycle, one would expect it to have some type of yeast counterpart; however, it is unclear why such a gene would require atypical organization.

β-globin genes: Figures 5A and 5B show 7th-order and 2nd-order analyses of the human β-globin gene cluster (12) using all other human sequences greater than 10kb to build the database. This higher cutoff is necessary because a high proportion of the database is cDNA sequences. The presence of highly repeated DNA sequences such as ALU clearly influence the 7th-order results but not the 2nd-order plot. All of the strong downward peaks are at the position of ALU elements. If the model is built with a short related sequence such as the 40kB *Galago crassicaudatus* (bush baby) β-globin gene cluster (13) the only significant downward peaks are seen at conserved regions including the coding sequences for the β-globin-like genes and a conserved regulatory region (not shown).

Figure 6A shows the results of a 7th-order plot with the bush baby β-globin cluster as query and a human model built from all sequences greater than 10kb. Note the large region around position 7000 that differs greatly from the human model. A small portion of this region was known to be highly homologous to the *E. coli* sequence IS186. Our statistical anlysis detected a much larger region of unusual organization. The larger region is highly homologous to the *E. coli entD* gene (14) with the left end homologous to IS421. Figure 6B shows the same sequence plotted using an *E. coli* model. While the *entD* homology fits well with the *E. coli* model, the other peak in figure 6A does not. This region (near 32000) contains a short segment homologous to the *E. coli lac* operon.

DISCUSSION

We have developed a simple method, using 7th-order Markov chains, to locate genes and clusters of genes in bacterial and yeast DNA sequences that are quite dissimilar from other sequences in those species. Our approach makes no assumptions about the nature of the sequences being tested and can scan tens of kilobases per second. Similar analyses can also reveal repeated elements and rapidly locate regions of homology between two sequences. This approach should permit the rapid location of interesting regions in large DNA sequences for subsequent computational analyses and biological experimentation.

If a stationary 7th-order Markov model were appropriate for the query sequences, then for most subsequences of length $j+1$ the value $-log_2 P(b_i, b_{i+1}...b_{i+j})$ should be approximately $(j+1)H$ where H is the entropy of the Markov chain. Our analysis differs in that our $-log_2 P$ values are derived from a database of other sequences, zero adjusted, and conditional on $b_1...b_7$. The evidence we have accumulated argues strongly that the 7th-order Markov model does not hold globally and forms the basis for identifying biologically interesting regions in large sequences.

We have established our confidence limits using a simulation procedure (see methods). Other efforts to determine the probability of the plots crossing any value assumed a normal distribution for the values being plotted because each would represent the sum of many identically distributed random variables. Q–Q plots (not shown) of these values for large sequences clearly indicate that this assumption is not correct and that the distributions of the plotted values lack the tails that would expand the range of values that might be expected. The variance determined by fitting a line to these plots was used in efforts to calculate the probability of extreme values arising in Markov chains of the length under study. These latter values gave 95% confidence limits far outside the range of the plots shown. While it is important not to equate biological and statistical significance (15), the confidence limits determined by the simulation procedure are quite effective in identifying interesting regions in the plots.

Markov chains have been used extensively to study DNA sequences (16, 17), particularly with shorter sequences. While some features mentioned above would have been revealed through analysis of base composition and nearest-neighbor frequencies, much additional useful information is present in the higher-order Markov chains. With bacterial sequences, the general outline of the plot emerges in a 2nd-order chain, reflecting the high fraction of coding sequence; however, many parts of the 2nd-order plot show strong downward peaks indicating that reliance on codon bias will excessively weight certain features (not shown). This is not true with human genomic sequences as coding sequences do not produce the predominant features of the plot. 5th-order plots begin to indicate the presence of highly repeated elements such as ALU sequences.

The sequence under study is not included when the database is constructed; however, multiple entries or similar sequences may be present for subsets of the sequence. At present, there is no systematic way to resolve these two possibilities for particular regions of a large DNA sequence. This type of overrepresentation does not significantly affect our conclusions. Using databases the size of known yeast and *E. coli* sequences, an extra copy of a sequence will lower its position in the plot by approximately $0.1 \cdot \mu_j$ (not shown).

There are many possible explanations for why a portion of a chromosome might have drastically different sequence composition from the remainder of the chromosome. Some variation would be expected by chance alone in the stationary model. Unknown selective pressures or structural features of chromosomes may be operating. Additional possibilities include horizontal transmission from an unrelated organism, viral infection, and incorporation of sequences from an organelle into the nuclear genome. The demonstration of transfer of plasmids from bacteria into yeast is one example of the first possibility. Many viruses are known to integrate their DNA into human chromosomes and fractions of the yeast mitochondrial genome have been detected in nuclear DNA. The methods presented here will readily detect hepatitis B virus in the background of human genomic DNA sequences (not shown).

The modular structure of bacterial genomes is built on the clustering of genes with related functions and the genetic exchanges that can take place between species. The central (b2) region of phage λ has long been known to have a markedly different base composition from other parts of the phage. This region is the part substituted in many λ transducing phages with *E. coli* DNA making it surprising that this is the part of the genome least like *E. coli* DNA. Similarly, the *lom* gene and the sequences at the extreme right end of λ are expressed in lysogens (10). All of the regions we find different from typical *E. coli* organization are known to be dispensible to the phage. Such regions of the phage appear be under different selective pressures from the rest of the genome and might be optimized for function in another host.

The sequence of yeast chromosome 3 has been extensively analyzed by a variety of statistical tests. The regions near the glucokinase gene and to the right of MAT have been noted to have unusual base compositions in the third position of their coding regions (18). This agrees well with our findings. While there is much evidence for movement of blocks of genes at the telomeres, much less is known about internal insertions or substitutions in yeast chromosomes. Given a sequence of the size of chromosome 3, the smallest region that can be shown to be statistically distinct with the observed level of variation is about 1kb. Detection of smaller regions will require additional biological or statistical criteria.

The results in figure 6 indicate that this approach will permit the rapid scanning of large sequences for the presence of vector or other bacterial contaminants that might have been included as a result of some step of the cloning or sequencing. Other reports have indicated the presence of *E. coli* IS elements in eukaryotic database entries as the result of homology searches (19, 20). The approach taken here does not rely on homology to known sequences to detect suspect regions.

The usage tables need not be constructed from all sequences of an organism. Any subset of adequate size could be used such as transcribed or translated regions. In fact, the fraction of coding sequences in human DNA is sufficiently low that some large exons in the neurofibromatosis region are identified even when a database including exons at their approximate frequency in the genome is used (not shown).

Our results indicate that caution must be exercised in the construction of data sets for use in gene-finding algorithms. It is clear that several authentic *E. coli* and yeast genes differ significantly from the norm using our statistical measure. Many of the large open reading frames of unusual sequence are in regions known to be dispensible or not related to any sequence in the databases. Using known genes to build the training set will significantly bias the set of reading frames selected by computer programs towards finding more genes like those we already know.

The highly non-random character of DNA sequences greatly complicates efforts at statistically modeling large genomic DNA sequences (21). Our studies provide a simple method to visually display this behavior. Graphical presentation of the results of sequence analysis offers several additional advantages over familiar tabular presentations. For example, the growth in size in public databases and query sequences causes an increase in the score required for statistical significance in a variety of tests. Even with an incomplete understanding of the statistical behavior of DNA sequences, the presence of scores all along a large query rapidly gives the investigator a feel for background scores. Graphs of this type have long been used to indicate features such as protein hydrophobicity. Simple modifications to conventional homology searches can be made to yield scores at each point along the sequence. We expect that this approach to the presentation of DNA sequence analysis results will have wide application.

ACKNOWLEDGEMENTS

We thank John Mulligan for providing yeast genomic DNA sequences prior to publication. This work was supported by by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Genome Program, of the US Department of Energy under Contract No. DE-AC03-76SF00098 and by NSF grant DMS 9113527 to TPS.

REFERENCES

(1) Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J.M. (1993) *Science* **259**, 1711–1716.

(2) Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, JP. and Benit, P. (1992) *Nature* **357**, 38–46.

(3) Josse, J., Kaiser, A.D. and Kornberg, A. (1961) *J. Biol. Chem.* **236**, 864–875.

(4) Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) *Nucleic Acids Res.* **20**, S2111–2118.

(5) Burge, C., Campbell, A.M. and Karlin, S., (1992) *Proc. Nat. Acad. Sci. USA* **89**, 1358–1362.

(6) Cuticchia A.J., Ivarie, R. and Arnold J. (1992) *Nucleic Acids Res.* **20**, 3651–3657.

(7) Good, I.J. (1965) *The Estimation of Probabilities: An Essay in Modern Bayesian Methods*, Cambridge, MA, MIT Press.

(8) Daniels, D., Schroeder, J., Szybalski, W., Sanger, F. and Blattner, F. (1983) In *Lambda II*, Hendrix *et al.* eds., 467–518.

(9) Campbell, A. and Botstein, D. (1983) In *Lambda II*, Hendrix *et al.* eds., 365–380.

(10) Barondess, J.J. and Beckwith, J.M. (1990) *Nature* **346**, 871–874.

(11) Ouellette, B.F., Clark, M.W., Keng, T., Storms, R.K., Zhong, W., Zeng, B., Fortin, N., Delaney, S., Barton, A. and Kaback, D.B. (1993) *Genome* **36**, 32–42.

(12) Hardison, R. and Miller, W. (1993) *Molec. Biol. Evol.* **10**, 73–102.

(13) Tagle, D.A., Stanhope, M.J., Siemieniak, D.R., Benson, P., Goodman, M. and Slightom, J.L. (1992) *Genomics* **13**, 741–760.

(14) Coderre, P.E and Earhart C.F. (1989) *J. Gen. Microbiol.* **135**, 3043–3055.

(15) Karlin, S. and Brendel, V. (1992) *Science* **257**, 39–49.

(16) Pesole, G., Prunella, N., Liuni, S., Attimonelli, M. and Saccone C. (1992) *Nucleic Acids Res.* **20**, 2871–2875.

(17) Kleffe, J. and Borodovsky, M. (1992) *Comp. Appl. Biosci.* **8**, 433–441.

(18) Sharp, P.M. and Lloyd, A.T. (1993) *Nucleic Acids Res.* **21**, 179–183.

(19) Lamperti, E.D., Kittelberger, J.M., Smith, T.F. and Villa-Komaroff, L. (1992) *Nucleic Acids Res.* **20**, 2741–2747.

(20) Binns, M. (1993) *Nucleic Acids Res.* **21**, 779.

(21) Karlin, S. and Brendel, V. (1993) *Science* **259**, 677–680.

FIGURE LEGENDS

Figure 1. Analysis of the bacteriophage λ sequence. Panel A presents a 7th-order Markov chain analysis; panel B is 1st-order. The block size was 512 nucleotides. In panel A, the 95% confidence limit for significant deviation from the model is at -0.28.

Figure 2. Analysis of yeast chromosome 3. Panel A uses the yeast model; panel B an *E. coli* model; panel C a yeast chromosome 3 model. The block size was 1024 nucleotides. The 95% confidence limit for the upper bound in panel A is -0.15.

Figure 3. An enlargement of part of yeast chromosome 3 with the block size reduced to 512 nucleotides. Open reading frames are indicated at the top of the figure. Note the alignment of the peaks with the largest open reading frame at any position in the region.

Figure 4. Analysis of a region of yeast chromosome 1 using the yeast model. The query sequence is Genbank entry YSCLTESPO. The block size and confidence limits are as described in figure 2. Open reading frames are indicated at the top of the figure. Both of the significant peaks align with open reading frames.

Figure 5. The human β-globin gene cluster. The query sequence used is HUMHBB with a block size of 512 nucleotides. Panel A presents a 7th-order plot; panel B presents a 2nd-order plot. The strong downward peaks in panel A are all at the location of ALU elements in the sequence.

Figure 6. The *Galago cressicaudatus* β-globin gene cluster. The query sequence used is GCRHBEGEB with a block size of 512 nucleotides. A 7th-order Markov chain analysis is shown using the human (panel A) or *E. coli* (panel B) model. Both of the upward peaks in panel A contain sequences highly homologous to *E. coli* DNA sequences.