BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION
Vol. 31, No. 1, pp. 64–68, 2003

Multimedia in Biochemistry and Molecular Biology Education

Visualization of Aligned Genomic Open Reading Frame Data*

Received for publication, April 26, 2002, and in revised form, July 3, 2002

Alan P. Boyle and John A. Boyle‡

From the Department of Biochemistry and Molecular Biology Mississippi State University Mississippi State, Mississippi 39762

Students can better appreciate the value of genomic data if they are asked to use the data themselves. However, in general the enormous volume of data involved makes detailed examination difficult. Here we present a web site that allows students to study one particular aspect of sequenced genomes. They are able to align the open reading frames (ORFs) of any available genome that is of reasonable size. The ORFs may be aligned using either the start codon or the stop codon as the starting points. Results will readily show the presence of common ribosome binding sites as well as reveal interesting order within the ORFs that is nonexistent outside of them. Students will be able to ask various questions involving comparisons of genomes and see the results presented in both a tabular and graphic format. An example problem is presented under "Results."

Keywords: Alignment, visualization, genomes, ORF.

The past few years have seen an explosion of DNA sequencing results. In particular, entire genomes, both prokaryotic and eukaryotic, are now available for study due to rapid advances in sequencing methodologies. The first viral sequence, phi-X174, was published in 1977 and consisted of 5386 bases in a single-stranded genome [1]. The first eubacterial sequence, *Haemophilus influenzae*, was completed in 1995 and was 1,830,137 base pairs long [2]. Yeast and human draft sequence lengths testify to the power of the available methods and the quantity of information produced. However, the enormous size of even the simplest genomes sometimes hinders easy use of the available data.

Students are presented with information relating to cis elements affecting transcription and translation by several different methods. The most common is a simple presentation of "consensus" data. For a more detailed appreciation, they are given tables of aligned sequences of these elements. Recent statistical approaches have provided a more accurate way to present summaries of the data. Two related methods are based on a maximum likelihood statistic. One involves a weight matrix [3], and the other involves an information theory approach [4]. Both methods have their proponents and virtues [5, 6]. A typical output shows cis sites in terms of varying sizes of base logos (see Ref. 7 for example).

Here we present an alternate way to display and examine alignable sequence data. We take advantage of the availability of genomic data to show that open reading frames (ORFs)[1] from entire genomes can be aligned to reveal global patterns in DNA data. We present a web site where students can ask their own questions about sequences and produce the results quickly.

MATERIALS AND METHODS

A Perl program was written to analyze the data. Connection to www.cs.msstate.edu/~apb22/CBIG/AlignORFs.cgi provides a web site that allows a user to do various types of analyses. In addition, this site provides files showing alignments of all eubacteria and archae that were available on February 20, 2002 at www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html [8]. Frequency data is provided from −100 to 200 where the zero base is the first base in the start codon. Graphs of the data are shown from −70 to 100.

The method of Staden is used to determine base frequencies [3]. Each ORF segment is aligned with the first base in the start codon or the last base in the stop codon being the zero base (ORFs indicated as being in the opposite direction are computed as reverse complements and aligned). As in a number line, bases before the zero base are negative; bases after it are positive. The bases at each position are then totaled by summing over all ORFs. The sums are divided by the total number of ORFs to find the real frequency and then normalized by dividing by the expected base content at each position using the G-C content of the organism. The G-C content is previously calculated using all G-C base pairs in the organism. Next a log probability of any given base in the region is calculated by taking the log of these normalized values [3, 5, 9]. The values are then output into a data file. The graphic representation of the data is presented along with the tabulated frequencies.

Students may generate their own frequency data by selecting the appropriate organism from the drop down menus (Fig. 1). Currently the National Center for Biotechnology Information (NCBI) lists the RefSeq accession number and the name of the

* This work was supported by National Science Foundation Grant 302473. This is Publication Number J10137 of the Mississippi Agricultural and Forestry Experiment Station.

‡ To whom correspondence should be addressed. Fax: 662-325-8664; E-mail: jab@ra.msstate.edu.

[1] The abbreviations used are: ORF, open reading frame; S-D, Shine-Dalgarno.

organism. Unfortunately the names do not currently distinguish between bacteria and their plasmids. For instance, *Bacillus subtilis* has five separate listings because of the plasmid sequences available as well as the chromosomal sequence. Knowledge of the appropriate RefSeq number will direct the student to the appropriate selection. There are currently no eukaryotic chromosomes available for alignment. Their size would overwhelm the capabilities of the server being used. The eukaryotic listing represents mitochondrial and chloroplastic sequences.

The start and stop codons of the ORFs of the genomes provide the starting points for the alignments. However, it is possible to align base regions that do not include either the start or stop codons so long as the regions selected use either of these codons as reference points. For example, to align bases −200 to −100 in front of an ORF, simply use these numbers as the start and stop positions relative to the start codon.

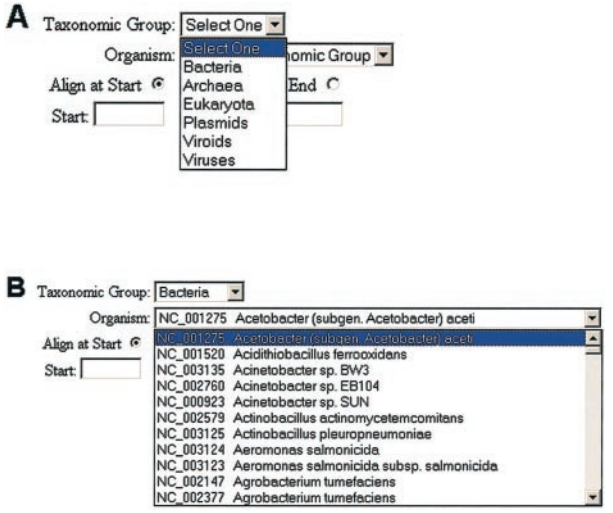

FIG. 1. **Details of the web site www.cs.msstate.edu/~apb22/CBIG/AlignORFs.cgi showing drop-down menus specifying the taxonomic group and the specific organism and RefSeq accession number.**

RESULTS

Alignments of eubacteria are shown at the web site at www.msstate.edu/dept/biochemistry/CBIG/Eubacteria.htm, and alignments of archae are shown at www.msstate.edu/dept/biochemistry/CBIG/Archaea.htm. An example of these alignments is shown in Fig. 2. The start codon is indicated by the prominent frequencies at positions 0, 1, and 2. A, T, and G show high positive frequencies in these locations, while the other bases at each location have low frequencies. The Shine-Dalgarno (S-D) ribosome binding site located just upstream from the start codon [10, 11] with the consensus sequence AGGAGG is also clearly seen. Since the location of the site is not exact, it appears as a G-rich region preceded by an A-rich region. This is uniform for most eubacteria. However, this alignment technique shows that not all eubacteria or archae have the expected ribosome binding site.[2] *Mycoplasma genitalium* and *Mycoplasma pneumoniae* both use leaderless transcripts [12] and so lack the A-rich, G-rich pattern (Fig. 3).

The repeat pattern seen within the ORFs is not seen upstream, nor is it seen in random sequences. Preliminary work shows that alignment of multiple codons within one gene also shows periodicity. This agrees with work done for some genes in *Pyrococcus* [13]. There has been some

[2] A. P. Boyle and J. A. Boyle, submitted.

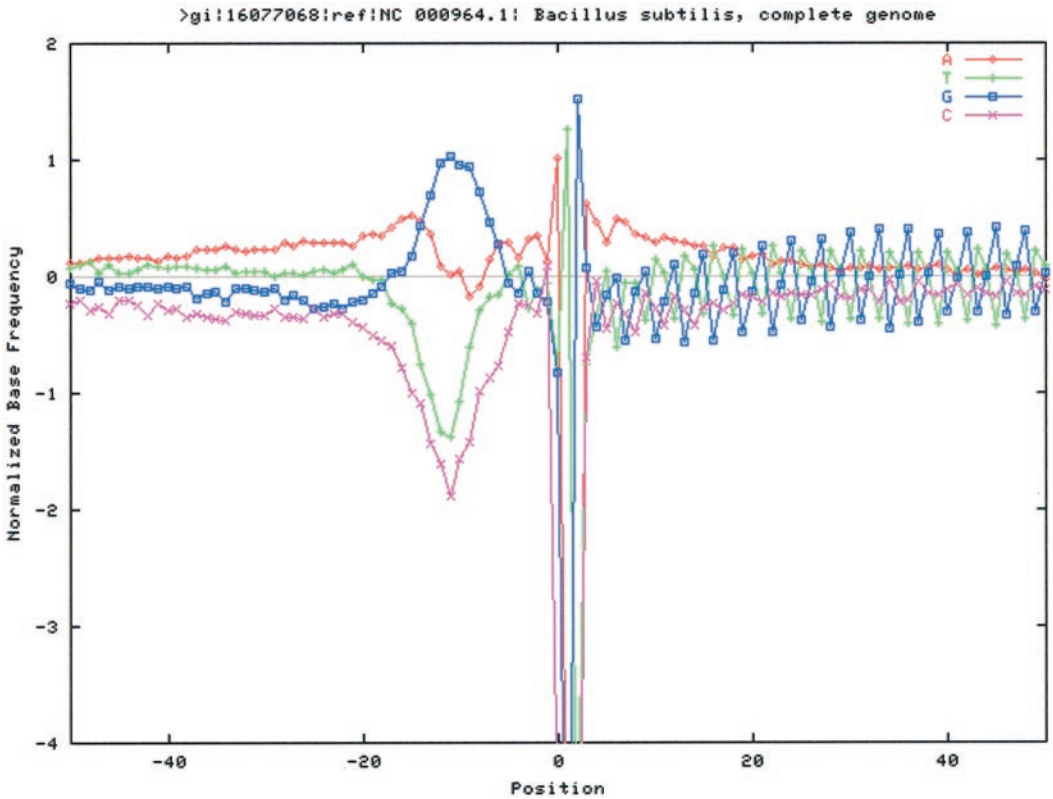

FIG. 2. **Alignment of all the open reading frames found in the genome of *B. subtilis*.** *Position* shows the base position relative to the start codon where the first base in this codon is designated as zero. Normalized base frequency is computed as indicated under "Materials and Methods."

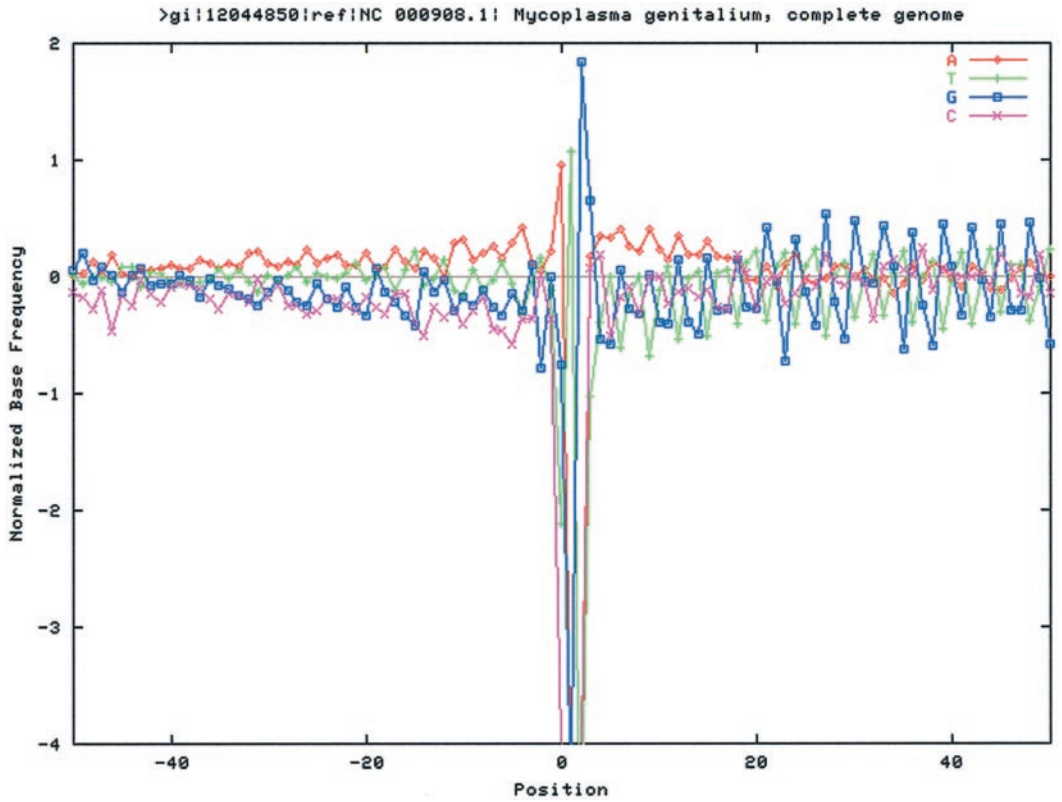

FIG. 3. **Alignment of all the open reading frames found in the genome of *M. genitalium*.** *Position* shows the base position relative to the start codon where the first base in this codon is designated as zero. Normalized base frequency is computed as indicated under "Materials and Methods."

suggestion that examination of periodicity within ORFs could assist in gene identification [14]. Visualization of the pattern makes clear the differences between coding and non-coding sequences in bacteria. Students can use this web site simply to view aligned sequences or as a research and learning tool to explore for similar or new patterns among the other available genomes.

As an example of the type of question that students might pose or be asked to solve, we present the following. Does the aligned ORF pattern of the *Arabidopsis thaliana* chloroplast resemble bacterial ORFs? The student should proceed to the web site and identify the chloroplast sequence for *Arabidopsis* (RefSeq accession number NC_000932). They could then compare the aligned results to those of a standard bacterium like *Escherichia coli* (RefSeq accession number NC_000913). Results could be presented as in Fig. 4. They would see that the chloroplast sequence has the expected ATG at the start of the ORF, but more interestingly, they would also see that it has a region upstream from the start codon that strongly resembles the bacterial Shine-Dalgarno region. It is not as prominent as the bacterial site since the chloroplast has far fewer genes to align and thus has a noisier average. However, the expected AGGAGG pattern is represented by a preponderance of A residues in the region of −11 to −12 and elevated G residues from −10 to −5. There is also a characteristic decline in C residues and especially T residues over this whole site. The student could be prompted to do more analysis and discover that plant chloroplasts have S-D regions that are less uniform in their location as compared with bacteria [15]. This could help explain the elevated G content in the region extending to −17.

DISCUSSION

The alignment software has been used as a research tool to ask questions about translation properties of eubacteria and archae.[2] Because of ease of use and the visual nature of the results, it can also been an excellent way for students to ask their own interesting questions about genomic data.

The inherent nature of the alignment procedure used means that some identifiable point is required as an anchor for the alignment. This procedure uses the designated start or stop codons from open reading frames in sequenced genomes. If a common transcription start were readily available, we could align on it. On the other hand, some archae and eubacteria with leaderless transcripts do show some suggestion of aligned transcription sites at a common distance from the start codon.[2]

The frequencies obtained and the resulting graphs can sometimes be "noisy" if the number of ORFs present is low. Almost no plasmids or viroids will present useful results, and only large viruses will be of any interest. Nevertheless, interesting results are obtainable from the small genomes of plant mitochondria and chloroplasts. The periodicity seen within ORFs is only definitive for organismal genomes.

Future enhancements to this web site will include the ability to align specific genes from one or multiple organ-

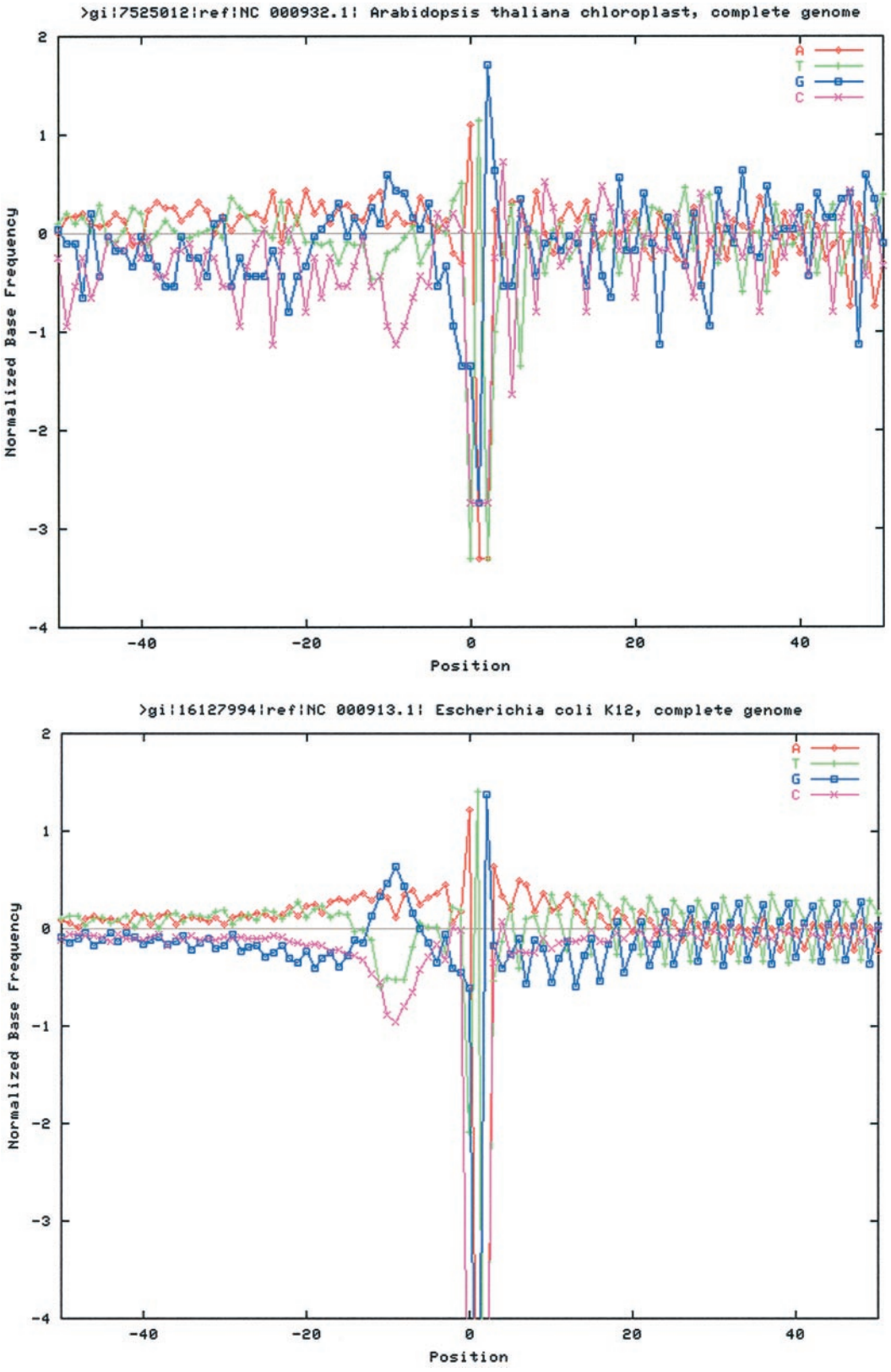

FIG. 4. **Alignment of all the open reading frames found in the genomes of *A. thaliana* chloroplasts (*top*) and *E. coli* (*bottom*).** *Position* shows the base position relative to the start codon where the first base in this codon is designated as zero. Normalized base frequency is computed as indicated under "Materials and Methods."

isms. In this way, if the student can identify an interesting subset of ORFs, they may be readily aligned. However, as indicated above, unless a substantial number of ORFs are used, the results will be quite variable.

REFERENCES

[1] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith (1977) Nucleotide sequence of bacteriophage phi-X174 DNA, *Nature* **265**, 687–695.

[2] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, *Science* **269**, 496–512.
[3] R. Staden (1984) Computer methods to aid the determination and analysis of DNA sequences, *Biochem. Soc. Trans.* **12**, 1005–1008.
[4] T. D. Schneider, G. D. Stormo, L. Gold, A. Ehrenfeucht (1986) Information content of binding sites on nucleotide sequences, *J. Mol. Biol.* **188**, 415–431.
[5] G. D. Stormo (2000) DNA binding sites: representation and discovery, *Bioinformatics* **16**, 16–23.
[6] T. D. Schneider (1997) Information content of individual genetic sequences, *J. Theor. Biol.* **189**, 427–441.
[7] T. D. Schneider, R. M. Stephens (1990) Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res.* **18**, 6097–6100.
[8] NCBI Entrez Genomes, Microbial Genomes: www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html.
[9] G. Z. Hertz, G. D. Stormo (1996) Escherichia coli promoter sequences: analysis and prediction, *Methods Enzymol.* **273**, 30–42.
[10] J. Shine, L. Dalgarno (1974) The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites, *Proc. Natl. Acad. Sci. U. S. A.* **71,** 1342–1346.
[11] C. O. Gualerzi, C. L. Pon (1990) Initiation of mRNA translation in prokaryotes, *Biochemistry* **29**, 5881–5889.
[12] J. Weiner III, R. Herrmann, and G. F. Browning (2000) Transcription in Mycoplasma pneumoniae, *Nucleic Acids Res.* **28**, 4488–4496.
[13] J. M. Suckow, N. Amano, Y. Ohfuku, J. Kakinuma, H. Koike, M. Suzuki (1998) A transcription frame-based analysis of the genomic DNA sequence of a hyper-thermophilic archaeon for the identification of genes, pseudo-genes and operon structures, *FEBS Lett.* **426**, 86–92.
[14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy (1997) Prediction of probable genes by Fourier analysis of genomic sequences, *Comput. Appl. Biosci.* **13**, 263–270.
[15] N. W. Gillham, J. E. Boynton, C. R. Hauser (1994) Translational regulation of gene expression in chloroplasts and mitochondria, *Annu. Rev. Genet.* **28**, 71–93.