Global Analysis of Microbial Translation Initiation Regions

Alan P. Boyle and John A. Boyle[1]
Department of Biochemistry and Molecular Biology,
Mississippi State University, Mississippi State, MS 39762

The availability of genomic sequences from multiple bacteria has allowed global comparisons of patterns. Here we present a graphical comparison of normalized base frequencies in the vicinity of translation starts for both eubacteria and archae. The results show that most eubacterial Open Reading Frames (ORFs) are preceded by a distinctly recognizable Shine-Dalgarno (SD) sequence pattern. However, some eubacteria deviate from this arrangement and have diminished SD patterns or completely lack this sequence. On the other hand, some archae seem to use both SD sequences and leaderless transcripts in their translation initiation process. This is dependent on the position of a gene within an operon. Most archae seem to have other regular sequences located upstream from the typical SD location. Both eubacteria and archae have a surprising repetitive pattern seen within the averaged ORFs. The eubacterial and archaeal averaged patterns are slightly different from each other, and individual organisms within each domain vary from the averages. Nevertheless, the existence of such a periodicity within ORFs may allow the development of new techniques to identify actual genes from ORFs.
Keywords: translation initiation, eubacterial, archaeal, Shine-Dalgarno, alignment

Eubacteria initiate their translational process by binding mRNA to the small ribosomal subunit. This occurs because of a complementarity between a sequence at the 3' end of 16S rRNA and the Shine-Dalgarno (SD) sequence just 5' to the initiation codon (Shine and Dalgarno, 1974; Gualerzi and Pon, 1990). However, rarely, some eubacteria have been shown to lack an untranslated leader of sufficient length to contain an SD sequence (Van Etten and Janssen, 1998). In addition, there is some evidence that *Mycoplasma* species may have a high proportion of leaderless transcripts. On the other hand, archae may more routinely use heterogenous mechanisms for translation initiation (Saito and Tomita, 1999). Work on two different species, *Pyrobaculum aerophilum* (Slupska et al., 2001) and *Sulfolobus solfataricus* (Tolstrup et al., 2000), has shown that while many genes seem to have SD sequences in the proper location, a significant number of others are likely to have leaderless transcripts.

The availability of large numbers of complete genomic sequences provides the opportunity to search for patterns within and among genomes. In particular there were 56 eubacterial and 11 archaeal sequences that were published by the middle of January, 2002 (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html). Newly sequenced genomes are subjected to computational methods that produce a collection of open reading frames (ORFs) that presumably correspond to the genes in the organisms. This type of analysis has already led to insights into the translational process (Saito and Tomita, 1999; Sakai et al., 2001; Ma et al., 2002).

We construct a matrix using a maximum likelihood statistical approach (Hertz and Stormo, 1996) and combine it with a graphical representation of the results to show results averaged over all ORFs for all available sequenced microbial genomes. This approach can reveal common patterns and deviations from these patterns for microorganisms. We discuss the results as they relate to translation initiation mechanisms.

MATERIALS AND METHODS

Sequences for the following genomes were available as of 01-22-02 at the Entrez Genomes section of the National Center for Biotechnology Information (NCBI) Web site (http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html) (hereafter referred to as the NCBI Microbial Genomes website):

[1]Corresponding author. (FAX 662-325-8664; email jab@ra.msstate.edu)

Agrobacterium tumefaciens Cereon circular chromosome (Goodner et al., 2001)
Agrobacterium tumefaciens Dupont circular chromosome (Wood et al., 2001)
Aeropyrum pernix K1 (Kawarabayasi et al., 1999)
Aquifex aeolicus chromosome (Deckert et al., 1998)
Archaeoglobus fulgidus (Klenk et al., 1997)
Bacillus halodurans C-125 (Takami et al., 2000)
Bacillus subtilis (Kunst et al., 1997)
Borrelia burgdorferi chromosome (Fraser et al., 1997)
Brucella melitensis chromosome I, chromosome II (DelVecchio et al., 2002)
Buchnera sp. APS (Shigenobu et al., 2000)
Campylobacter jejuni (Parkhill et al., 2000)
Caulobacter crescentus (Nierman et al., 2001)
Chlamydophila pneumoniae CWL029 (Kalman et al., 1999)
Chlamydophila pneumoniae AR39 (Read et al., 2000)
Chlamydophila pneumoniae J138 (Shirai et al., 2000)
Chlamydia trachomatis (Stephens et al., 1998)
Chlamydia muridarum chromosome (Read et al., 2000)
Clostridium acetobutylicum chromosome (Nolling et al., 2001)
Deinococcus radiodurans R1 chromosome 1, chromosome 2 (White et al., 1999)
Escherichia coli K12 (Blattner et al., 1997)
Escherichia coli O157:H7 EDL933 (Perna et al., 2001)
Escherichia coli O157:H7 (Hayashi et al., 2001)
Halobacterium sp. NRC-1 (Fleischmann et al., 1995)
Haemophilus influenzae (Ng et al., 2000)
Helicobacter pylori 26695 (Tomb et al., 1997)
Helicobacter pylori J99 (Alm et al., 1999)
Lactococcus lactis subsp. *lactis* (Bolotin et al., 2001)
Listeria monocytogenes EGD-e (Glaser et al., 2001)
Listeria innocua (Glaser et al., 2001)
Methanobacterium thermoautotrophicum (Smith et al., 1997)
Methanococcus jannaschii chromosome (Bult et al., 1996)
Mesorhizobium loti chromosome (http://www.kazusa.or.jp/rhizobase/)
Mycobacterium tuberculosis H37Rv (Cole et al., 1998)
Mycobacterium tuberculosis CDC1551 (Fleischmann, R.D., D. Alland, J.A. Eisen, L. Carpenter, O. White, J. Peterson, R. DeBoy, R. Dodson, M. Gwinn, D. Haft, et al., Whole genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. Unpublished (listed in NCBI Microbial Genomes website).
Mycobacterium leprae (Cole et al., 2001)
Mycoplasma genitalium (Fraser et al., 1995)
Mycoplasma pneumoniae (Himmelreich et al., 1996)
Mycoplasma pulmonis (Chambaud et al., 2001)
Neisseria meningitidis MC58 (Tettelin et al., 2000)
Neisseria meningitidis Z2491 (Parkhill et al., 2000)
Nostoc sp. PCC 7120 (Kaneko et al., 2001)
Pasteurella multocida (May et al., 2001)
Pseudomonas aeruginosa (Stover et al., 2000)
Pyrococcus abyssi (Heilig, R. *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution Unpublished (listed in NCBI Microbial Genomes website).
Pyrococus horikoshii (Kawarabayasi et al., 1998)
Rickettsia conorii Malish 7 (Ogata et al., 2001)
Rickettsia prowazekii (Ogata et al., 2001)
Ralstonia solanacearum (Salanoubat, M., S. Genin, F. Artiguenave, J. Gouzy, S. Mangenot, M. Arlat, A. Billault, P. Brottier, J.C. Camus, L. Cattolico, et al., Genome sequence of the plant pathogen *Ralstonia solanacearum* Unpublished (listed in NCBI Microbial Genomes website).
Salmonella typhimurium LT2 (McClelland et al., 2001)
Salmonella typhi (Parkhill et al., 2001)
Sinorhizobium meliloti (Capela et al., 2001)
Staphylococcus aureus N315 (Kuroda et al., 2001)
Staphylococcus aureus Mu50 (Kuroda et al., 2001)
Streptococcus pneumoniae TIGR4 (Tettelin et al., 2001)
Streptococcus pneumoniae R6 (Hoskins et al., 2001)
Streptococcus pyogenes (Ferretti et al., 2001)
Sulfolobus solfataricus (She et al., 2001)
Sulfolobus tokodaii (Kawarabayasi et al., 2001)
Synechocystis PCC6803 (Kaneko et al., 1996)
Thermoplasma acidophilum (Ruepp et al., 2000)
Thermoplasma volcanium (Kawashima et al., 1999)
Treponema pallidum (Fraser et al., 1998)
Thermotoga maritima (Nelson et al., 1999)
Ureaplasma urealyticum (Glass et al., 2000)
Vibrio cholerae chromosome I, chromosome II (Heidelberg et al., 2000)
Xylella fastidiosa chromosome (Simpson et al., 2000)
Yersinia pestis chromosome (Parkhill et al., 2001)

Software was developed using Perl and run on Sun OS 5.7. Data was read from standard formatted data files found at the Entrez-Genome site hosted by

the NCBI (NCBI Microbial Genomes website) and the information needed to align the open reading frames was extracted. Each eubacterial and archaeal DNA sequence was downloaded in FASTA format (*.fna) and the open reading frame information was selected from the same site in ProTein Table (*.ptt) format. No reformatting of the data was required. The only input required by the software was the location of the genome files and the start/stop locations for the alignment (in this case -70 to +50 where 0 is the first base in the start codon). The program first removed all the end-line characters and calculated the G-C content of the strain. Next, each ORF segment was aligned with the start codon beginning at the zero location (ORFs indicated as being in the opposite direction were computed as reverse compliments and aligned). The bases at each position were then totaled by summing over all ORFs. The sums were 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 by using the G-C content of the organism. Next, a log probability of any given base in the region was calculated by taking the log of these normalized values (Staden, 1984; Hertz and Stormo, 1996; Stormo, 2000). The values were then output into a data file. This was repeated for each genomic file. An average for each bacterial domain was calculated from all the frequencies in that domain. A total of 58 eubacterial sequences were run requiring 1598 seconds and totaling 146,335 ORFs. A total of 11 archaeal sequences were run requiring 235 seconds and totaling 23,361 ORFs.

Subtraction of individual organismal patterns from the domain average was accomplished by direct subtraction of frequencies at each base. The result was reploted to show the differences in frequencies.

Sulfolobus solfataricus and *Halobacterium* sp. NRC-1 genomes were examined to identify ORFs that were likely to be members of operons. In order to qualify, ORFs must have been assigned an identity at the NCBI Microbial Genomes web site and be part of a recognizable cluster of related genes. The selected genes were related by function (as in collections of ribosomal protein genes or subunits of a protein like NADH dehydrogenase) or by being part of a pathway (like proteins in the cobalamin biosynthetic pathway). The gene cluster had to be closely linked with fewer than ten bases separating the genes (many overlapped slightly). The initial gene in the putative operon had to be separated by 50 to 100 bases from the preceding gene and had to have no identifiable functional relationship with it. No hypothetical protein genes were used.

A version of the program used for this work is accessible at the following web site: http://www.msstate.edu/dept/biochemistry/CBIG/. In addition the aligned genomes of all microbial species examined are available at this site.

RESULTS

Examination of individual eubacterial genomic sequence patterns and the eubacterial average over all reported ORFs (Figure 1) reveals distinctive patterns near the start codon. The averaged start codons themselves show the expected overwhelming presence of A, T, and especially G in the third position (base number 2 in the representation used). There is a general enrichment of A's both upstream and downstream from the start codon with the obvious exception of the Shine-Dalgarno region and a decline in A as the last base before the start. There is also a general decrease in G's in this start codon proximal region except for SD. The expected SD sequences upstream from start contribute to a distinctive pattern. The canonical sequence of AGGAGG may vary somewhat in its location with respect to translation start. The net result is a distribution of this sequence over a range of locations in the genomic patterns and in the eubacterial average. The high G content of SD yields a broad G peak centered around -9 to -10 bases upstream from start. There is a rise in A content upstream of this peak and a definite decline within the peak. T and C frequencies drop off dramatically in the SD region.

Individual eubacterial genomes vary from the average pattern. A few eubacteria maintain a preponderance of A's over G's throughout the SD region. This is true for *Caulobacter crescentus*, *Mycobacterium tuberculosis*, and *Xylella fastidioasa* (Figure 2). These still have the decline in T's and C's as observed for others. Some eubacteria, e.g. *Bacillus subtilis* (Figure 3) and *Staphylococcus aureus* (see Figure 1), have SD regions whose center is shifted slightly upstream from the average and have a higher frequency of G's and lower frequency of T's and C's in this region as compared to the average. This is highlighted by subtracting the organismal pattern from the eubacterial average (Figure 3).

Synechocystis PCC6803 has a clear deviation from the average pattern (Figure 4). While there is a rise in G's at SD, it is not nearly so pronounced as in the average. The decline in C's is also not conspicu-

ous. There is no obvious explanation for this anomaly. *Deinococcus radiodurans* also presents an anomalous pattern (Figure 4). Instead of a G peak in the SD region, it displays a prominent A peak and only small decreases in T and C. There is a T peak at position -7. The patterns are consistent for both chromosomes.

Other notable exceptions from the eubacterial pattern are seen in two *Mycoplasma* species, *Mycoplasma genitalium* and *Mycoplasma pneumoniae* (Figure 4). While both evidence the general rise in A's and decline in G's near the start codon, each lacks evidence of SD sequences. In contrast *Mycoplasma pulmonis* fits the standard eubacterial pattern except that a larger proportion of its ORFs start with something other than ATG. Removal of the *M. genitalium* and *M. pneumoniae* frequencies from the eubacterial average has little affect on the pattern since together they only contribute 1173 ORFs out of the almost 150,000 found in the overall average.

The average over all archaeal sequences shows a much less uniform pattern than for eubacteria (Figure 5). While this could be due to the smaller sample size of archaeal sequences, the number of ORFs involved is still over 23,000. The variations are more likely due to the diverse nature of the archaeal kingdom. It is also likely due to diverse translation initiation mechanisms for different classes of genes within individual archae (Tolstrup et al., 2000; Slupska et al., 2001).

There are some similarities between the archaeal and eubacterial averages. There is an enrichment in A's immediately upstream and downstream from the start codon. The last base before this codon is depleted in A. There is a decrease in G before the start, but, in contrast to eubacteria, there is no general depletion in G after the start codon. While there is a clear G peak and decline in A's and C's in the SD region, there is no drop in T's here. In addition the SD pattern is not so pronounced as in the eubacterial case. Individual archae lack distinct SD patterns. *Thermoplasma acidophilum, Aeropyrum pernix*, and *Halobacterium* sp. NRC-1 have none of the standard base distributions here. On the other hand, with the exception of *Aeropyrum pernix*, they and the other archae have apparent consistent structure in the region upstream from the SD location that is seen in no eubacteria (Figure 5). From -27 to -30 there is a preponderance of T's and -19 to -26 and -31 to -35 shows a preference for A in all the archae. These regions seem to show a diminished amount of C and G although this is not pronounced.

Examination of the results for *Sulfolobus solfataricus* reveals some of the complexity of archaeal genes (Figure 6). Its genomic average is very similar to the overall archaeal average. However, a sampling of 63 genes that are likely to be transcribed as internal members of polycistronic mRNAs shows a pattern that presents a clear SD signal with little upstream structure. Sampling 40 genes that are likely to be the initial sequences in polycistrons presents a noisy average but reveals no SD pattern and suggests the possible presence of the A-T rich and C-G poor region from -35 to -19. Combination of the patterns from the two gene classes would yield the sequence structures found upstream from the translation start signal in the overall genomic pattern for *Sulfolobus*.

Sampling 72 genes that appear to be internal and 29 that appear to be the first in an operon in *Halobacterium* sp. NRC-1 gives a similar result (data not shown). An A-T rich region is seen from -38 to -25 with the typical T peak surrounded by A peaks when the operon initiating genes are averaged. A G rich region from -12 to -8 and correspondingly C poor region is suggestive of an SD region for genes internal to operons. This region is obscured in the overall average.

The gene sequences downstream from the start codon show a surprising regularity in all bacteria examined. This is not an artifact of the alignment; combination of random genomic sequences (data not shown) presents only noise. A similar regularity has been observed by others for individual genes and organisms (Fickett, 1984; Tsonis et al., 1991; Suckow et al., 1998).

DISCUSSION

The statistical and graphical approach used here shows averages of the sequenced eubacterial and archaeal genomes. Patterns that relate to the translational processes are readily visualized. Comparisons with individual organisms reveal possible deviations from the standard processes. As expected, most eubacteria have a distinct indication of a Shine-Dalgarno region located just upstream from the initiation codon. There is also a general enrichment of A's near the start with the exception of a clear decrease at position -1. There is no clear evidence of a downstream box that has been suggested to be a translation enhancer in the region of +7 to +12 (Sprengart et al., 1996).

Deviations from the average results for eubacteria

could be caused by several factors. Misidentification of ORFs by various gene finding programs could produce noise in the data. An ORF may be incorrectly extended in the 5' direction beyond the actual start codon simply because there is an open reading frame available beyond that codon. This would place the real SD site within the gene when the alignment is done. Some sequences identified as ORFs may not be real genes. This would also introduce noise in the alignment. Conceivably, some genes may contain translational elements that are quite different from the average. It might be an interesting application of this technique to identify ORFs that deviate in some statistical way from the average. Do they cluster in some fashion? Do they have unusual characteristics as genes or do they look like misassigned sequences?

Some organisms deviate in definite but not critical ways from the average. A few have a G peak in the SD region that is less than the A peak but still have clear decreases in T and C. *B. subtilis* and *S. aureus* have intense but shifted SD signals. This is most clearly seen by subtracting the average results from the organismal results. These variations from the average are not likely to represent large scale differences in translation initiation.

Some organismal averages are distinct from the eubacterial average. *Synechocystis* shows an exceptionally weak SD pattern and also has reasonably prominent C frequencies on either side of the start codon. Other work has suggested that this bacterium is somewhat different in its translation initiation and may have different classes of genes (Osada et al., 1999; Sakai et al., 2001). However, examination of only the highly expressed class of genes (Mrazek et al., 2001) (data not shown) does not show much variation from the overall average. *Deinococcus radiodurans* shows the typical lower values for C and T frequencies in the SD region, but there is an A peak here for both chromosomes. G tends to increase around -8 to -9. This is unlike any other eubacterial genome. Examination of the 16S ribosomal RNA genes in both *Synechocystis* and *Deinococcus* shows the expected AGGAGG sequence (NCBI Microbial Genomes website) and so altered complementarity is not to be expected. The regularity seen within the coding region is also clearly different for *Deinococcus*. These two eubacteria are not related in any fashion with *Synechocystis* being a cyanobacterium and *Deinococcus* being one of the most unusual eubacterial extremophiles (White et al., 1999). It will be interesting to compare these results with any newly sequenced cyanobacteria to see if a unique translation initiation pattern emerges.

The results seen with *Mycoplasma genitalium* and *Mycoplasma pneumoniae* highlight their high frequency of leaderless transcripts (Weiner et al., 2000). This phenomena, whereby an mRNA is produced with few bases in front of the start codon, has been observed in every type of organism but is usually rare (Van Etten and Janssen, 1998). Translation initiation seems to be accomplished through the actions of either bacterial or eukaryotic equivalents of IF-2 (Kyrpides and Woese, 1998; Grill et al., 2000; Grill et al., 2001). These two *Mycoplasma* species show no indication at all of an SD region. While *Synechocystis* and *Deinococcus* are unusual, they still evidence a diminished G peak; *M. genitalium* and *M. pneumoniae* have no distinguishable G peak. The other species of mycoplasma in this study, *Mycoplasma pulmonis* and the related organism *Ureaplasma urealyticum*, have a readily distinguishable SD pattern and therefore are not likely to have leaderless transcripts.

Archeal translation initiation is not as well understood as eubacterial initiation. Genomic analysis has suggested that at least some archae use a combination of eubacterial and eukaryotic mechanisms (Salin et al., 1991; Saito and Tomita, 1999). Some archaeal initiation factors are homologs of the eukaryotic equivalents (Bult et al., 1996; Keeling et al., 1998; Kyrpides and Woese, 1998) yet many archaeal genes have clear SD regions, appropriately complimentary regions at the 3' end of their 16S rRNA molecules, and no 5' CAP structure. Some archaeal transcripts are leaderless (May and Dennis, 1990; Condo et al., 1999; Slupska et al., 2001). Recent work has suggested that in at least some archea, the specific translation initiation mechanism depends on whether the gene in question is located internal to an operon or is an isolated gene or the first gene in an operon (Tolstrup et al., 2000; Slupska et al. 2001).

The results here reflect both the diversity of the archaeal domain and the diversity of mechanisms within individual organisms. Not all of the archae, especially *Halobacterium* and to a lesser extent *Thermoplasma acidophilum* and *Aeropyrum pernix*, have a clearly defined SD region. However, with the exception of *Aeropyrum pernix*, all the archaeal genomes have at least some indication of common sequence elements further upstream from an expected SD site. There are A rich regions located around -34 and -24 and a T rich region centered at -28. The T region and the -24 A region could readily

correspond to an A box structure (consensus TTTA(A or T)A) (Wich et al., 1986; Reiter et al., 1988; Thomm and Wich, 1988; Reiter et al., 1990) slightly diffuse out due to heterogeneity in location. This element represents a transcription start signal yielding leaderless transcripts having no SD region and is found upstream of isolated genes and first genes in operons in *Pyrobaculum aerophilum* (Slupska et al., 2001) and *Sulfolobus solfataricus* (Tolstrup et al., 2000; She et al., 2001). It was not found upstream of internal genes in operons. The results presented here for *Sulfolobus* and *Halobacterium* support this conclusion and lead to the suggestion that each genome result and the overall archaeal average represents a superposition of at least the two classes of genes. The SD region centered at -9 is from internal members of operons; the putative A box (the combination of the T peak at -28 and the A peak at -24) is the transcription signal for leaderless products of isolated genes and first genes in operons. The A region at -34 may be another type of transcriptional signal since it is likely to be too far upstream to be involved in translation initiation. It seems that most if not all archae studied here may use at least two mechanisms for translation initiation based on these results.

The repetitive pattern seen with the averaged coding sequences is consistently present in all organisms examined. It is especially interesting that a pattern is seen even when all ORFs over both the eubacteria and archae are averaged (Figures 1 and 5). However, the observed patterns are slightly different in each case. There is a clear periodicity of three bases seen and this is likely due to codon structure. Nevertheless, there is no reason to expect such a regularity in base positions. Both eubacteria and archae have higher than expected G frequencies in the first codon position and lower than expected in the second position. This begins immediately after the start codon in archae but becomes prominent in eubacteria only after several codons. Conversely, T is elevated in the second position but diminished in the first position. The other bases show lower effects in the averages. Individual organisms can show patterns that are distinct from the averages. Figure 2 shows that *Caulobacter crescentus* has regularly higher C and lower A in the third position. Figure 5 shows that *Halobacterium* has higher C and lower T in the third position and both higher A and T in the second position. The set of all 69 organisms examined has much variety but each member always shows some regular pattern.

The pattern observed internal to the ORFs is not seen upstream nor is it seen in random sequences. The pattern is observable even if few ORFs are examined as is seen in Figure 6 when 63 and 40 ORFs are aligned. In particular, the G repeat stands out from the noisy background. Preliminary work shows that alignment of multiple codons within one gene also shows periodicity. This agrees with work done for some genes in *Pyrococcus* (Suckow et al., 1998). There has been some suggestion that examination of periodicity within ORFs could assist in gene identification (Tiwari et al., 1997). It remains to be seen whether eukaryotes will have similar patterns.

Detailed analysis should allow comparison of individual ORFs to organismal averages. This may prove useful in identifying ORFs that represent outliers in the data. Are they atypical because they are members of a particular gene class? The *Sulfolobus* and *Halobacterium* genes that begin operons show this characteristic. On the other hand, further study of some abnormal ORFs might eventually show that they have been incorrectly identified as genes.

The combination of a visual approach with a maximum likelihood statistical treatment of aligned ORFs can be powerful in revealing global patterns around translation initiation sites in both eubacteria and archae. The availability of large numbers of genomes allows formulation of reasonable averages for the two domains considered here. Since a broad range of bacteria were included, it is not likely that the results are skewed in any particular way. No bacterium contributed more than 8% of the total ORFs in question.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Carolyn Boyle for her advice on statistical validity of methods used. This work was supported by NSF Award No. EPS0082979. This is publication number J10086 of the Mississippi Agricultural and Forestry Experiment Station.

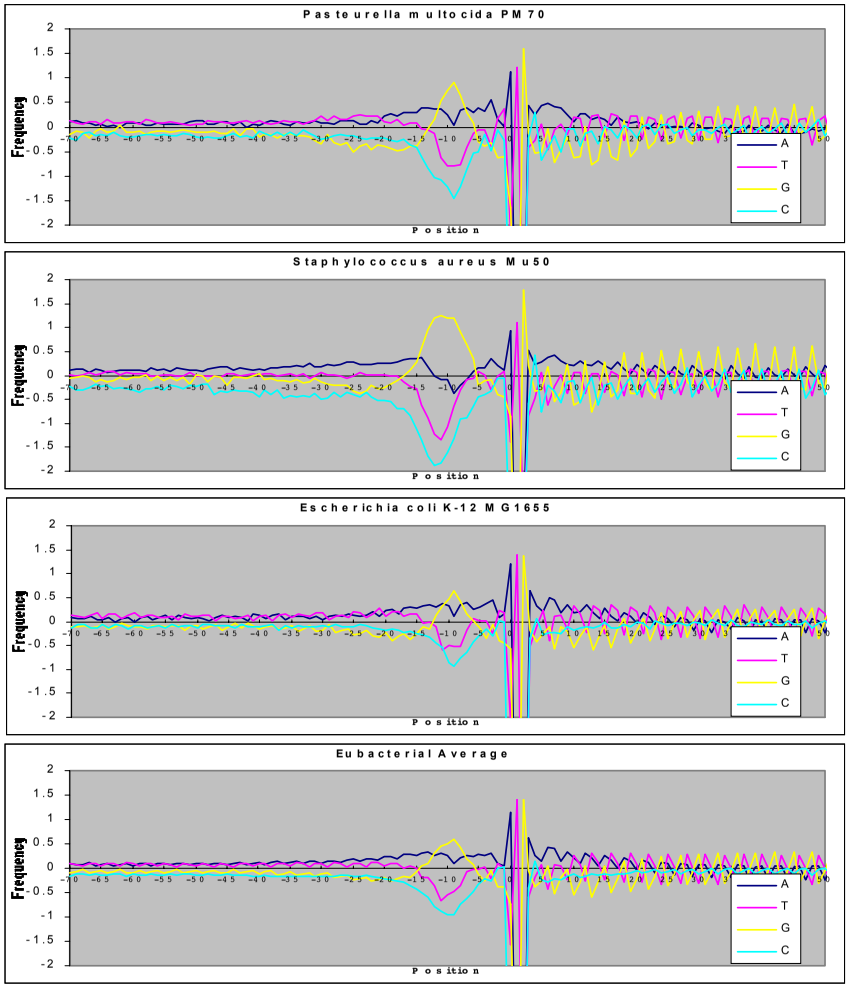

Figure 1. Aligned ORFs for three representative eubacteria, *Pasterella multocida* PM70, *Staphylococcus aureus* Mu50, and *Escherichia coli* K-12 MG1655, and for all ORFs of all eubacterial sequences available as of 01-22-02. Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The base position immediately preceding the start codon is -1. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9.

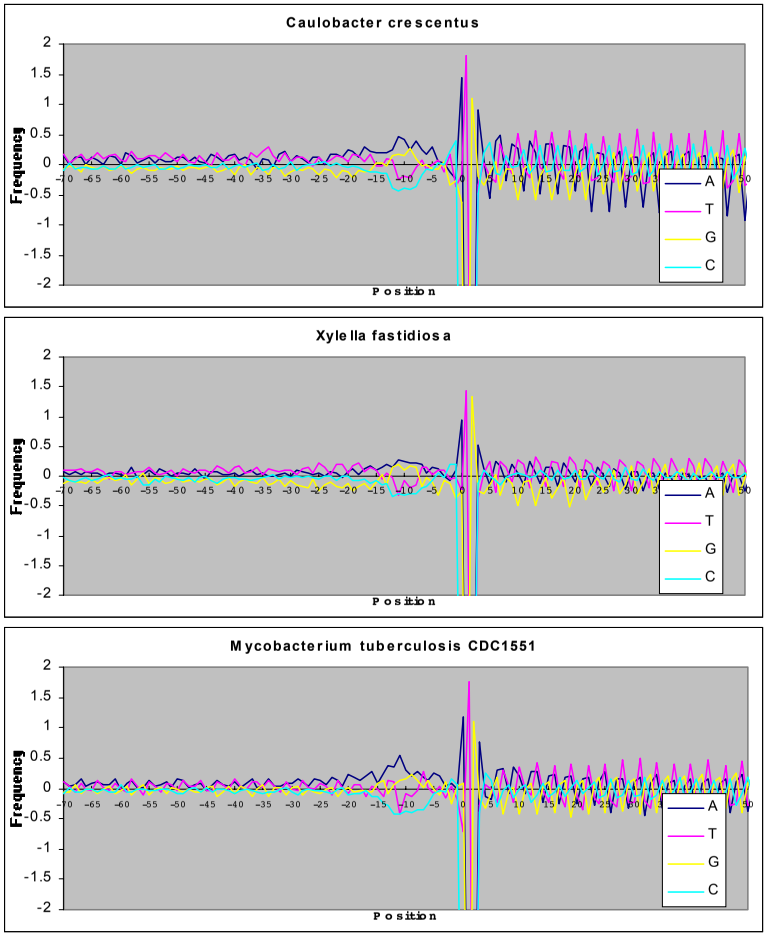

Figure 2. Aligned ORFs for *Caulobacter crescentus*, *Mycobacterium tuberculosis*, and *Xylella fastidioasa*. Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9.

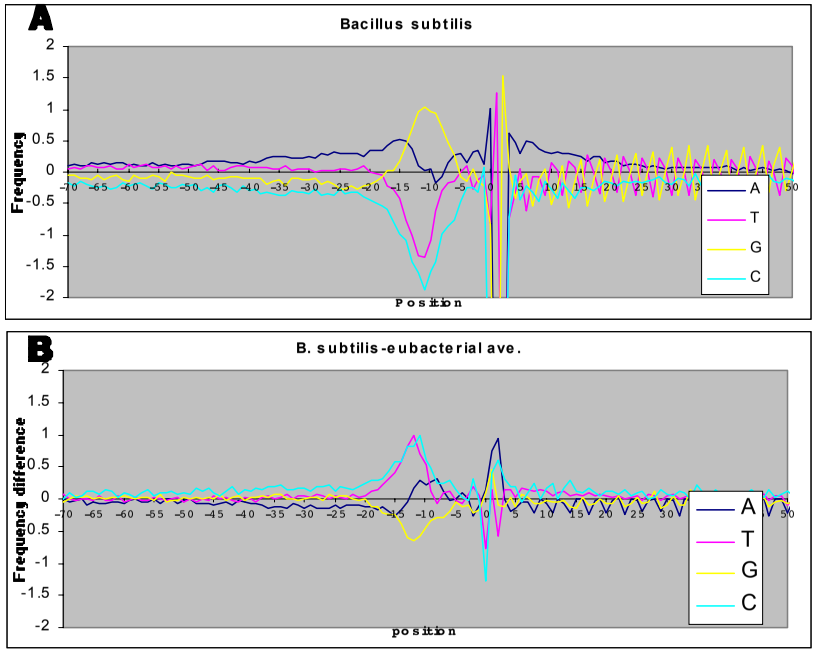

Figure 3. (A) Aligned ORFs for *Bacillus subtilus*. Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9. (B) Differences between the weighted frequencies at each position for *Bacillus subtilus* and those of the eubacterial average.

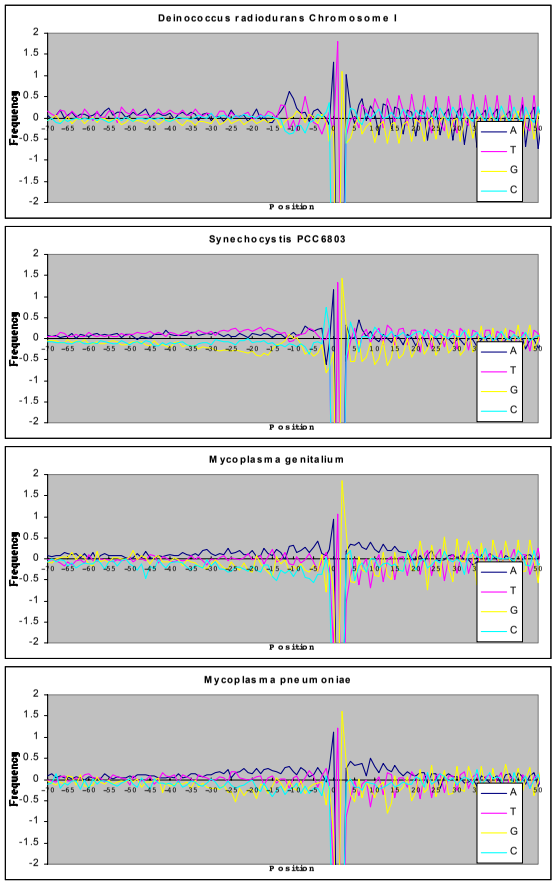

Figure 4. Aligned ORFs for *Deinococcus radiodurans, Synechocystis* PCC6803, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9.

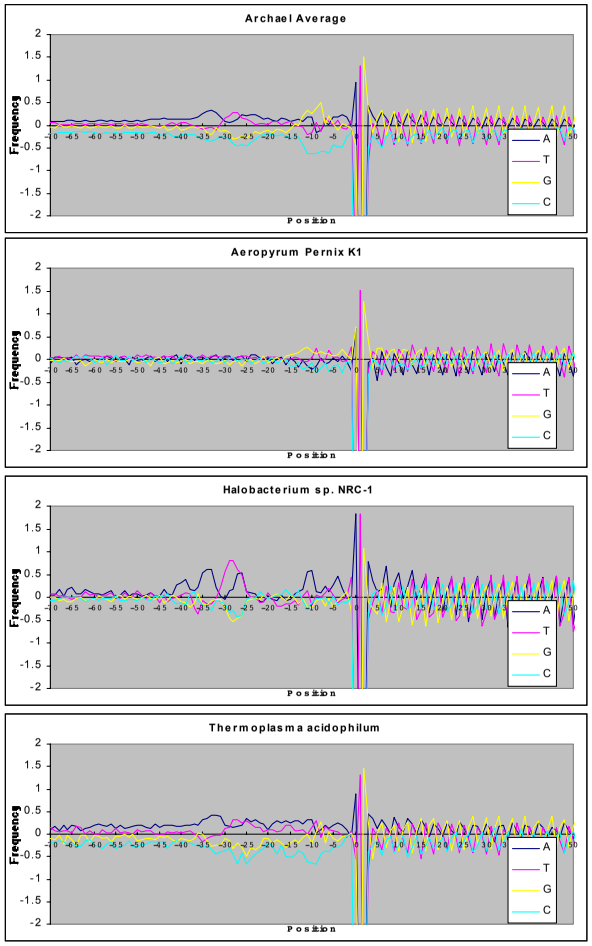

Figure 5. Aligned ORFs for all archaeal sequences available as of 01-22-02 as well as aligned ORFs for *Thermoplasma acidophilum, Aeropyrum pernix*, and *Halobacterium* sp. NRC-1. Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9.

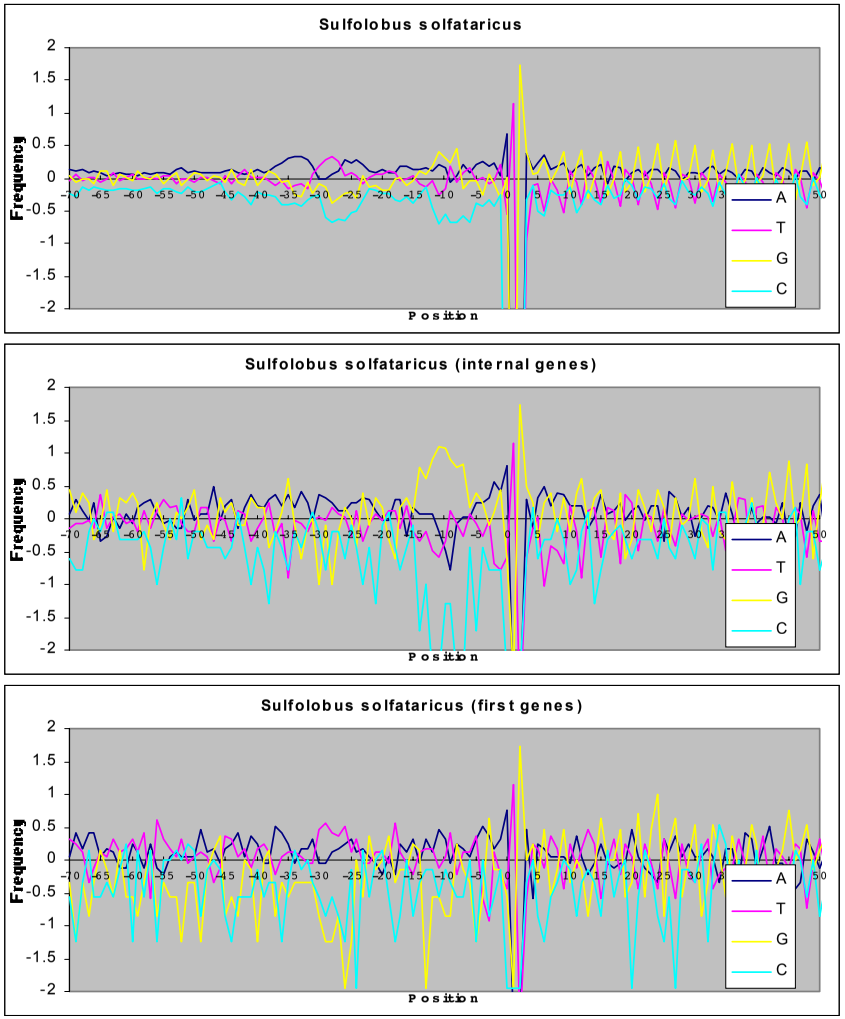

Figure 6. Aligned ORFs for *Sulfolobus solfataricus* as well as ORFs from 63 genes likely to be transcribed as internal members of operons (internal genes) or ORFs from 40 genes likely to be the first genes in operons (first genes). Position represents the base position for each ORF aligned so that the first base in the start codon is aligned as zero on a number line. The frequency represents a weighted, maximum likelihood statistical value calculated for each position as indicated in Methods and in reference 9.

LITERATURE CITED

Alm, R.A., L.-S.L. Ling, D.T. Moir, B.L. King, E.D. Brown, P.C. Doig, D.R. Smith, B. Noonan, B.C. Guild, B.L. deJonge, et al. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176–180.

Blattner, F.R., G. Plunkett, III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474.

Bolotin, A., P. Wincker, S. Mauger, O. Jaillon, K. Malarme, J. Weissenbach, S.D. Ehrlich, and A. Sorokin. 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. Genome Res. 11:731–753.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073.

Capela, D., F. Barloy-Hubler, J. Gouzy, G. Bothe, F. Ampe, J. Batut, P. Boistard, A. Becker, M. Boutry, E. Cadieu, et al. 2001. Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. Proc. Natl. Acad. Sci. USA 98:9877–9882.

Chambaud, I., R. Heilig, S. Ferris, V. Barbe, D. Samson, F. Galisson, I. Moszer, K. Dybvig, H. Wroblewski, A. Viari, et al. 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res. 29:2145–2153.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, III, et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544.

Cole, S.T., K. Eiglmeier, J. Parkhill, K.D. James, N.R. Thomson, P.R. Wheeler, N. Honore, T. Ganier, C. Churcher, D. Harris, et al. 2001. Massive gene decay in the leprosy bacillus. Nature 409:1007–1011.

Condo, I., A. Ciammaruconi, D. Benelli, D. Ruggero, and P. Londei. 1999. Cis-acting signals controlling translational initiation in the thermophilic archaeon *Sulfolobus solfataricus*. Mol. Microbiol. 34:77–84.

Deckert, G., P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay, et al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392:353–358.

DelVecchio, V.G., V. Kapatral, R.J. Redkar, G. Patra, C. Mujer, T. Los, N. Ivanova, I. Anderson, A. Bhattacharyya, A. Lykidis, et al. 2002. The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. Proc. Natl. Acad. Sci. USA 99:443–448.

Ferretti, J.J., W.M. McShan, D. Adjic, D. Savic, G. Savic, K. Lyon, C. Primeaux, S.S. Sezate, A.N. Surorov, S. Kenton, et al. 2001. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc. Natl. Acad. Sci. USA 98:4658–4663.

Fickett, J.W. 1984. Fast optimal alignment. Nucleic Acids Res. 12:175–179.

Fleischmann, R.D., 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.

Fraser, C.M., S. Casjens, W.M. Huang, G.G. Sutton, R.A. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey, M. Gwinn, B. Dougherty, et al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–586.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G.G. Sutton, J.M. Kelley, et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403.

Fraser, C.M., S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E. K. Hickey, R. Clayton, K.A. Ketchum, et al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388.

Glaser, P., L. Frangeul, C. Buchrieser, A. Amend, F. Baquero, P. Berche, H. Bloecker, P. Brandt, T. Chakraborty, A. Charbit, et al. 2001. Comparative genomics of Listeria species. Science 294:849–852.

Glass, J.I., E.J. Lefkowitz, J.S. Glass, C.R. Heiner, E.Y. Chen, and G.H. Cassell. 2000. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407:757–762.

Goodner, B., G. Hinkle, S. Gattung, N. Miller, M. Blanchard, B. Qurollo, B.S. Goldman, Y. Cao, M. Askenazi, C. Halling, et al. 2001. Genome Sequence of the Plant Pathogen and Biotechnology Agent *Agrobacterium tumefaciens* C58. Science 294:2323–2328.

Grill, S., C.O. Gualerzi, P. Londei, and U. Blasi. 2000. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. EMBO J. 19(15):4101–4110.

Grill, S., I. Moll, D. Hasenohrl, C.O. Gualerziand, and U. Blasi. 2001. Modulation of ribosomal recruitment to 5'-terminal start codons by translation initiation factors IF2 and IF3. FEBS Lett. 495:167–171.

Gualerzi, C.O., and C.L. Pon. 1990. Initiation of mRNA translation in prokaryotes. Biochemistry 29:5881–5889.

Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C.-G. Han, E. Ohtsubo, K. Nakayama, T. Murata, et al. 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8:11–22.

Heidelberg, J.F., J.A. Eisen, W.C. Nelson, R.A. Clayton, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, L.A. Umayam, et al. 2000. DNA Sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483.

Hertz, G.Z., and G.D. Stormo. 1996. *Escherichia coli* promoter sequences: analysis and prediction. Methods Enzymol. 273:30–42.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24:4420–4449.

Hoskins, J.A., W. Alborn, Jr., J. Arnold, L. Blaszczak, S. Burgett, B.S. DeHoff, S. Estrem, L. Fritz, D.-J Fu, W. Fuller, et al. 2001. The Genome of the Bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. 183:5709–5717.

Kalman, S., W. Mitchell, R. Marathe, C. Lammel, J. Fan, R.W. Hyman, L. Olinger, J. Grimwood, R.W. Davis, and R.S. Stephens. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat. Genet. 21:385–389.

Kaneko, T., Y. Nakamura, C.P. Wolk, T. Kuritz, S. Sasamoto, A. Watanabe, M. Iriguchi, A. Ishikawa, K. Kawashima, T. Kimura, et al. 2001. Complete Genomic Sequence of the Filamentous Nitrogen-fixing Cyanobacterium Anabaena sp. strain PCC 7120. DNA Res. 8:205–213.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3:109–136.

Kawarabayasi, Y., Y. Hino, H. Horikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, Ankai, A., H. Kosugi, A. Hosoyama, et al. 2001. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res. 8:123–140.

Kawarabayasi, Y., Y. Hino, H. Horikawa, S. Yamazaki, Y. Haikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, et al. 1999. Complete Genome Sequence of an Aerobic Hyper-thermophilic Crenarchaeon, Aeropyrum pernix K1. DNA Res. 6:83–101.

Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, et al. 1998. Complete Sequence and Gene Organization of the Genome of a Hyper-thermophilic Archaebacterium, *Pyrococcus horikoshii* OT3. DNA Research 5:55–76.

Kawashima, T., Y. Yamamoto, H. Aramaki, T. Nunoshiba, T. Kawamoto, K.Watanabe, M. Yamazaki, K. Kanehori, N. Amano, Y. Ohya, K. Makino, and M. Suzuki. 1999. Determination of the complete genomic DNA sequence of *Thermoplasma volvanium* GSS1. Proc. Jpn. Acad. 75:213–218.

Keeling, P.J., N.M. Fast, and G.I. McFadden. 1998. Evolutionary relationship between translation initiation factor eIF-2gamma and selenocysteine-specific elongation factor SELB: change of function in translation factors. J. Mol. Evol. 47:649–655.

Klenk, H.P., R.A. Clayton, J.-F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson, et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370.

Kunst, F., N. Ogasawara, I. Moszer, A.M. Albertini, G. Alloni, V. Azevedo, M.G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390:249–256.

Kuroda, M., T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa,, I. Kobayashi, L. Cui, A. Oguchi, K. Aoki, Y. Nagai, et al. 2001. Whole genome sequencing of meticillin-resistant *Stapylococcus aureus*. The Lancet 357:1225–1240.

Kyrpides, N.C., and C.R. Woese. 1998. Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families. Proc. Natl. Acad. Sci. USA 95:3726–3730.

Ma, J., A. Campbell, and S. Karlin. 2002. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J. Bacteriol. 184:5733–5745.

May, B.P., and P.P. Dennis. 1990. Unusual evolution of a superoxide dismutase-like gene from the extremely halophilic archaebacterium *Halobacterium cutirubrum*. J. Bacteriol. 172:3725–3729.

May, B.J., Q. Zhang, L. Li, M.L. Paustian, T.S. Whittam, and V.S. Kapur. 2001. Complete nucleotide sequence of an avian isolate of *Pasteurella multocida*. Proc. Natl. Acad. Sci. USA 98:3460–3465.

McClelland, M., K.E. Sanderson, J. Spieth, S.W. Clifton, P. Latreille, L. Courtney, S. Powollik, J. Ali, M. Dante, F. Du, et al. 2001. The complete genome sequence of *Salmonella enterica serovar Typhimurium* LT2: features revealed by comparison to related genomes. Nature 413:852–856.

Mrazek, J., D. Bhaya, A.R. Grossman, S. Karlin. 2001. Highly expressed and alien genes of the Synechocystis genome. Nucleic Acids Res. 29:1590–1601.

Nelson, K.E., R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, et al. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323–329.

Ng, W.V., S.P. Kennedy, G.G. Mahairas, B. Berquist, M. Pan, H.D. Shukla, S.R. Lasky, N.S. Baliga, V. Thorsson, J. Sbrogna, et al. 2000. Genome sequence of Halobacterium species NRC-1. Proc. Natl. Acad. Sci. USA 97:12176–12181.

Nierman, W.C., T.V. Feldblyum, I.T. Paulsen, K.E. Nelson, J. Eisen, J.F. Heidelberg, M. Alley, N. Ohta, J.R. Maddock, I. Potocka, et al. 2001. Complete Genome Sequence of *Caulobacter crescentus*. Proc. Natl. Acad. Sci. USA 98:4136–4141.

Nolling, J., G. Breton, M.V. Omelchenko, K.S. Markarova, Q. Zeng, R. Gibson, H.M. Lee, J. Dubois, D. Qiu, J. Hitti, et al. 2001. Genome Sequence and Comparative Analysis of the Solvent-Producing Bacterium Clostridium acetobutylicum. J. Bacteriol. 183:4823–4838.

Ogata, H., S. Audic, P. Renesto-Audiffren, P.-E. Fournier, V. Barbe, D. Samson, V. Roux, P. Cossart, J. Weissenbach, J.-M. Claverie, and D. Raoult. 2001. Mechanisms of Evolution in *Rickettsia conorii* and *Rickettsia prowazekii*. Science 293:2093–2098.

Osada, Y., R. Saito, and M. Tomita. 1999. Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. Bioinformatics 15:578–581.

Parkhill, J., M. Achtman, K.D. James, S.D. Bentley, C. Churcher, S.R. Klee, G. Morelli, D. Basham, D. Brown, T. Chillingworth, et al. 2000. Complete DNA sequence of a serogroup A strain of *Neisseria menigitidis* Z2491. Nature 404:502–506.

Parkhill, J., G. Dougan, K.D. James, N.R. Thomson, D. Pickard, J. Wain, C. Churcher, K.L. Mungall, S.D. Bentley, M.T.G. Holden, et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica serovar Typhi* CT18. Nature 413:848–852.

Parkhill, J., B.W. Wren, K. Mungall, J.M. Ketley, C. Churcher, D. Basham, T. Chillingworth,
R.M. Davies, T. Feltwell, S. Holroyd, et al. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–668.

Parkhill, J., B.W. Wren, N.R. Thomson, R.W. Titball, M.T.G. Holden, M.B. Prentice, M. Sebaihia, K.D. James, C. Churcher, K.L. Mungall, et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527.

Perna, N.T., G. Plunkett, III, V. Burland, B. Mau, J.D. Glasner, D.J. Rose, G.F. Mayhew, P.S. Evans, J. Gregor, H.A. Kirkpatrick, et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533.

Read, T.D., R. Brunham, C. Shen, S.R. Gill, J.F. Heidelberg, O. White, E.K. Hickey, J. Peterson, L.A. Umayam, T. Utterback, et al. 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. 28:1397–1406.

Read, T.D., R.C. Brunham, C. Shen, S.R. Gill, J.F. Heidelberg, O. White, E.K. Hickey, J. Peterson, L.A. Umayam, T. Utterback, et al. 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. 28:1397–1406.

Reiter, W.D, U. Hudepohl, and W. Zillig. 1990. Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection *in vitro*. Proc. Natl. Acad. Sci. USA 87:9509–9513.

Reiter, W.D., P. Palm, and W. Zillig. 1988. Transcription termination in the archaebacterium Sulfolobus: signal structures and linkage to transcription initiation. Nucleic Acids Res. 16:2445–2459.

Ruepp, A., W. Graml, M.L. Santos-Martinez, K.K. Koretke, C. Volker, H.W. Mewes, D. Frishman, S. Stocker, A.N. Lupas, and W. Baumeister. 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature 407:508–513.

Saito, R., and M. Tomita. 1999. Computer analyses of complete genomes suggest that some archaebacteria employ both eukaryotic and eubacterial mechanisms in translation initiation. Gene 238:79–83.

Sakai, H., C. Imamura, Y. Osada, R. Saito, T. Washio, and M. Tomita. 2001. Correlation between Shine-Dalgarno sequence conservation and codon usage of bacterial genes. J. Mol. Evol. 52:164–170.

Salin, M.L., M.V. Duke, D.P. Ma, and J.A. Boyle. 1991. *Halobacterium halobium* Mn-SOD gene: archaebacterial and eubacterial features. Free Rad. Res. Commun.12–13 Pt 1:443–449.

She, Q., R.K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M.J. Awayez, C.C. Chan-Weiher, I.G. Clausen, B.A. Curtis, A. De Moors, et al. 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA 98:7835–7840.

Shigenobu, S., H. Watanabe, M. Hattori, Y. Sakaki, and H. Ishikawa. 2000. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature 407:81–86.

Shine, J., and 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. USA 71:1342–1346.

Shirai, M., H. Hirakawa, M. Kimoto, M. Tabuchi, F. Kishi, K. Ouchi, T. Shiba, K. Ishii, M. Hattori, S. Kuhara, and T. Nakazawa. 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res. 28:2311–2314.

Simpson, A.J.G., F.C. Reinach, P. Arruda, F.A. Abreu, M. Acencio, R. Alvarenga, L.M.C. Alves, J.E. Araya, G.S. Baia, C.S. Baptista, et al. 2000. The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406:151–157.

Slupska, M.M., A.G. King, S. Fitz-Gibbon, J. Besemer, M. Borodovsky, and J.H. Miller. 2001. Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. J. Mol. Biol. 309:347–360.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H.-M. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J. Bacteriol. 179:7135–7155.

Sprengart, M.L., E. Fuchs, and A.G. Porter. 1996. The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*. EMBO J. 15:665–674.

Staden, R. 1984. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res.12:505–519.

Stephens, R.S., S. Kalman, C. Lammel, J. Fan, R. Marathe, L. Aravind, W. Mitchell, L. Olinger, R.L. Tatusov, Q. Zhao, et al. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. Bioinformatics 16:16–23.

Stover, C.K., X.-Q.T. Pham, A.L. Erwin, S.D. Mizoguchi, P. Warrener, M.J. Hickey, F.S.L. Brinkman, W.O. Hufnagle, D.J. Kowalik, M. Lagrou, et al. 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406:959–964.

Suckow, J.M., N. Amano, Y. Ohfuku, J. Kakinuma, H. Koike, and 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.

Takami, H., K. Nakasone, Y. Takaki, G. Maeno, Y. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, et al. 2000. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res. 28:4317–4331.

Tettelin, H., K.E. Nelson, I.T. Paulsen, J.A. Eisen, T.D. Read, S. Peterson, J. Heidelberg, R.T.
DeBoy, D.H. Haft, R.J. Dodson, et al. 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science 293:498–506.

Tettelin, H., N.J. Saunders, J. Heidelberg, A.C. Jeffries, K.E. Nelson, J.A. Eisen, K.A. Ketchum, D.W. Hood, J.F. Peden, R.J. Dodson, et al. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287:1809–1815.

Thomm, M., and G. Wich. 1988. An archaebacterial promoter element for stable RNA genes with homology to the TATA box of higher eukaryotes. Nucleic Acids Res. 16:151–163.

Tiwari, S., S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Comput. Appl. Biosci. 13:263–270.

Tolstrup, N., C.W. Sensen, R.A. Garrett, and I.G. Clausen. 2000. Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*. Extremophiles 4:175–179.

Tomb, J.-F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty, et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–547.

Tsonis, A.A., J.B. Elsner, and P.A. Tsonis. 1991. Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol.151:323–331.

Van Etten, W.J., and G.R. Janssen. 1998. An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*. Mol. Microbiol. 27:987–1001.

Van Etten, W.J., and G.R. Janssen. 1998. An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli*. Mol. Microbiol. 27:987–1001.

Weiner, J., 3rd, R. Herrmann, and G.F. Browning. 2000. Transcription in *Mycoplasma pneumoniae*. Nucleic Acids Res. 28:4488–4496.

White, O., J.A. Eisen, J.F. Heidelberg, E.K. Hickey, J.D. Peterson, R.J. Dodson, D.H. Haft, M.L. Gwinn, W.C. Nelson, D.L. Richardson, et al. 1999. Genome Sequence of the Radioresistant Bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577.

Wich, G., H. Hummel, M. Jarsch, U. Bar, and A. Bock. 1986. Transcription signals for stable RNA genes in *Methanococcus*. Nucleic Acids Res.14:2459–79.

Wood, D.W., J.C. Setubal, R. Kaul, D. Monks, L. Chen, G.E. Wood, Y. Chen, L. Woo, J.P. Kitajima, V.K. Okura, et al. 2001. The Genome of the Natural Genetic Engineer *Agrobacterium tumefaciens* C58. Science 294:2317–2323.

Web Site References

http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html NCBI Entrez-Genome Microbial Genomes

http://www.kazusa.or.jp/rhizobase/Rhizobase

http://www.msstate.edu/dept/biochemistry/CBIG/Mississippi State University Computational Biology and Informatics Group