Eukaryotic Gene Prediction

Kelli Davies

2009 December 12

Introduction:

The advent of large-scale genome sequencing has revolutionized the field of genetics and biology. Sequencing projects require sophisticated computational analysis to manage vast collections of data. Scientists first sequenced a genome in 1977, that of a small bacteriophage consisting of 11 genes over 5.4kb of DNA. In the bacteriophage, coding genes comprise 95% of the genome.[1] Since then, numerous prokaryotic and eukaryotic genomes have been sequenced, including the mouse (*M. musculus*) genome, the human (*H .sapiens*) genome and the model plant *Arabidopsis thaliana* genome. Gene prediction in eukaryotic genomes can be especially difficult given the large genome size, the low proportion of coding regions, and the frequent splice events due to the presence of introns (non-coding segments between the exons that code for a gene). For example, the human genome contains approximately 25,000 genes over 30 million base pairs. Given that the average protein encoded by these genes is 350 amino acids, this means that only about 1% of the genome actually codes for proteins and these regions must be separated from the remaining 99% of the genome.[1] The actual coding regions are frequently interrupted by introns that are removed from the mRNA transcripts through splicing.

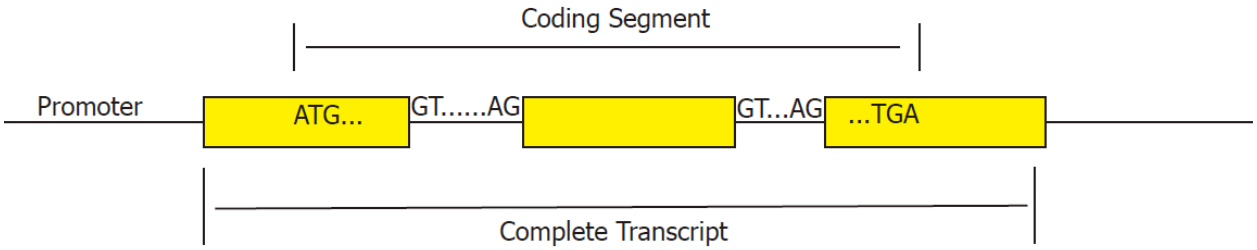


Figure 1: The typical structure of a eukaryotic gene with exons shown in yellow. Splicing signals for the intervening introns include the typical GT donor site and AG acceptor site. Translation signals for the coding segment of the transcript include the start codon (ATG) and a stop codon (such as TGA). This figure is adapted from Figure 3.2 and Figure 3.3 of Majoros 2007.

Gene structure includes a promoter region to allow for temporal and spatial (i.e. cell type) expression by trans-regulatory elements. In addition to the introns that are removed from the pre-

mRNA transcript via splicing, the actual mRNA transcript includes 5' and 3' regulatory regions that are not translated (these UTRs can be important for miRNA regulation). Complete gene structure is therefore very complicated. The primary focus of most gene prediction programs is to identify all genes in a given sequence and to exactly identify the boundaries of regions corresponding to the coding segments of exons.[1]

Traditionally, forward genetic screens involve mapping mutations to specific genetic loci. Now, proper annotation of completely sequenced genomes facilitates accelerated biological discovery. Once genes are identified, further analysis of the coding sequence can be used to identify protein domains to help illuminate function and determine candidates for further study. Homologs in different organisms can be identified via sequence alignment and then further studies can be conducted to determine if a protein in one organism has analogous function to its most closely related protein in another organism. Genome analysis has revealed that the genetic evolutionarily history of some organisms, especially plants, has been characterized by large duplication events. These duplications can complicate functional characterization, as single mutants often display no phenotype due to a closely related gene (or genes) in the organism that exhibits redundant or compensatory function.[2] By identifying all genes in the genome, one can analyze these sequences to guide the generation of multiple mutants of closely related genes that may manifest a mutant phenotype (thereby elucidating gene function). Correct annotation of the genome is therefore critical for a wide range of biological applications and for the accelerated functional characterization of genes.

Several methods can be employed to identify genes within a genome. These approaches can be broadly grouped into sequence alignment and ab initio prediction. In sequence alignment, genetic sequences are compared to known genes in other organisms (via BLAST) or to EST/cDNA data. The use of BLAST searches to explore a new genome is limited because many genes will not have a recognizable homolog[3] With ab initio prediction, intrinsic signals within the sequence (such as start and stop codons and splice sites) are combined with probabilistic models to predict whether or not a given sequence has gene structure. This critique will focus on gene prediction through use of ab initio programs and also consider alternative approaches such as EST alignment.

Methods in commonly used Gene Prediction Programs:

One major class of gene prediction programs are ab initio programs that use complex algorithims to identify gene signals within sequences. These programs generally have several

assumptions, including no overlapping genes, no nested genes (such as a gene located in the intron of another gene), no partial genes (due to missing sequence in genome assembly), no frameshifts, no noncanonical signals (such as splice sites other than GT and AG), and no alternative splicing. [1] In order to predict gene structure, the programs search a sequence for signal sensors. These sensors often include start codons (ATG), stop codons (TAG, TGA, TAA), splice donor sites (typically GT), splice acceptor sites (typically AG), and promoter sites (TATA box and CAT box upstream of coding sequence).[3] Notably, most of these nucleotide sequences are short and nonspecific. Simply identifying these sites in a long DNA sequence therefore is not very useful. Identification of these sequences is combined with coding statistics, in which probability is used to determine the likelihood of a gene being present in a sequence and its exon boundaries. Hidden Markov Models are frequently used in ab initio programs because they allow for the modeling of various hidden states and the probability of moving from one state to another.

Ab initio programs also require a training set so that the program parameters (such as HMMs and weight matrices) can be optimized such that the program is able to predict genes for the genome composition of a particular organism. Notably, some organisms have significant codon bias that can be extremely predictive of exons.[1] Different organisms have different nucleotide composition in the signal sensor regions, exon, and introns.[1] Weight matrices can model these differences; for instance, in a particular organism the AG splice sites may be flanked by a T at some high frequency, and this probability could be trained into the program. Exons and introns can also have a greater probability to be of a certain length in specific organisms.[4] The G+C content of organisms can vary[5] and in many organisms, coding portions of exons have elevated G+C levels.[1] For example, *Arabidopsis* has an overall G+C content of 36%, with coding G+C 44% and noncoding 32%.[6] (Please also see Supplemental Figure 7). One must also consider overall GC content more generally in order to account for codon bias. Within open reading frames, GC rich genomes tend to have a G or C at the third position whereas AT rich genomes tend to have an A or a T.[7] Thus, the performance of a given ab initio program depends not only on the algorithms used in the program but also on adequate training which requires a representative test set of genes.

<u>GENSCAN:</u>

The GENSCAN (Burge and Karlin 1997) program at MIT provides versions of the program trained for *Arabidopsis*, Maize, and Vertebrates. This commonly used program explicitly scores for transcription and translation signals[8] and uses an explicit-duration HMM. [3] For signal sensors,

GENSCAN identifies donors and acceptor splice sites, start codons, stop codons, promoters, and polyadenylation signals and these signal sensors are not specific to G+C density.[1]

To analyze sequences for donor sites, GENSCAN uses a maximal dependence decomposition (MDD) method, which can be thought of as a tree in which internal nodes correspond to specific decisions and leaf nodes correspond to solutions. One can move from one node to another based on the defined predicates of each node, which are associated with probabilities. In GENSCAN, the leaves of the MDD tree are weight matrices (WMMS) that include the GT splice site three bp from the beginning of the 9bp sequence. Acceptor sites are not modeled with an MDD but instead are modeled using a weight array matrix (WAM) and the AG site is offset at 20bp in the 23bp WAM. Start codons, stop codons, and polyadenylation signals (AATAAA consensus) are modeled using simple weight matrices of lengths 12bp, 6bp, and 6bp respectively. The other signal sensor used by GENSCAN, the promoter, is modeled using a 15bp WMM for the TATA-box which must be 14 to 20 bp from 8bp WMM for the CAP site. Since approximately 30% of eukaryotic genes do not have a TATA-box, a split model is used in which the presence of a TATA box is assigned a probability score of 0.7 and no TATA box is assigned a score of 0.3.[9] Lastly, exons, introns, UTRs, and intergenic regions in GENSCAN are all modeled using 5th-order Markov chains (three-periodic for exons). GENSCAN trains two different 5th-order Markov chains, one for low G+C density (0-43%) and one for high G+C density (43-100%). [1] Additionally, GENSCAN is able to predict the presence of partial genes.[9]

FGENES and FGENESH: There are various FGENE (Find Genes) programs, including the pattern based FGENES and the HMM based FGENESH. These programs are found at the SoftBerry website. The FGENESH provided at the SoftBerry website provides numerous version of the program, which have been trained on a wide variety of organisms. These include chicken, frog, human, mouse, drosophila, honey bee, phytophthora, ustilago, algae, dicot plants, monocot plants, and many more.

The basic premise of the initial FGENES is a pattern-based method that uses dynamic programming and discriminant classifiers in order to produce exon candidates. Candidate exons are first identified by searching for all open reading frames bordered by known signals (e.g. ATG…GT, AG-GT, AG… STOP). These candidate exons are then placed in order given their relative position (5' to 3') and a maximum score is calculated for each possible path of compatible exons. Lastly, any identified promoters or poly(A) tails are scored and placed at the appropriate terminal exon.[10]

The FGENESH program version uses Hidden Markov models similar to GENSCAN. The FGENESH is considered different from other programs because it does not place as much weight on content terms (such as codon usage) as it does on signal terms.[4] It also uses a Bayes Theorem to calculate the probability of exons. Similar to GENSCAN, FGENESH uses separate potentials for regions with low GC content (less than 45%) and those with high GC content (45% and greater).[10] There is also a version of FGENESH (FGENESH_GC) on the website which allows for exon prediction using the non-canonical GC splice donor.

An additional version of FGENESH is FGENESH+ which combines the FGENESH HMM based prediction program with information from protein homologues. It does this through an additional calculation that implements a Smith-Waterman for alignment of predicted exons with the protein homolog.

GeneMark.hmm: GeneMark.hmm is another HMM based program that was originally made for gene prediction in bacteria. This program has since been modified for eukaryotic gene prediction and also has a self-training version of the program requiring a sequence of 10MB. Like GENSCAN, GeneMark.hmm uses an explicit-duration HMM, which means that each state of the HMM has an associated length distribution.[3] This HMM is also referred to as a hidden semi-Markov model. The HSMM provides hidden states for the initial and terminal exons, introns, intergenic regions, single exons, splice sites, initiation sites, and termination sites. These hidden sites emit nucleotide sequences of fixed length and are modeled by positional Markov chains. The protein coding state is modeled by a three-period Markov chain and the order of the Markov chain (up to 5^{th} order) is determined by the training sequence.[5]

GeneMark.hmm has also developed a self-training version of the program for eurkaryotic genomes. This is an important algorithm because gene prediction programs need to be trained in order to optimize the parameters for a specific organism. Typically, this training process requires the use of a large, already verified training set which can be hard to acquire. For self-training, the program uses an unsupervised iterative estimation of gene parameters through a process known as Viterbi training. The Viterbi training provides a method to estimate the parameters for the Hidden Markov Model. The algorithm uses the genome sequences labeled by the Viterbi algorithm and then re-estimates its parameters to compute new sequences until a convergence point is reached.[5]

GeneSeqer@PlantGDB : A spliced alignment program specific for plants can be found at GeneSeqer@PlantGDB. The program provides specific splice site models for *Arabidopsis*, Maize,

Rice, and Medicago. Geneseqer uses spliced alignment with ESTs and full-length cDNAs.[11] For the gene prediction, the ESTs and cDNAs can be source native or nonnative (i.e. from putative homologs in other species). The EST and cDNA can be either supplied by user input or retrieved from a database. The spliced alignment uses ESTs alongside a genomic sequence and the resulting aligned regions are assigned as exons and gaps as introns. The gaps are generally flanked by the stereotypic GT and AG donor and acceptor splice sites. Prediction is based on both the sequence similarity and splice site strength. [12]

GeneSeqer allows for gene prediction based on the most current EST and cDNA sequences available. These sequences can help overcome what is known as "annotation lag" which refers to the fact that major annotation projects are often outdated in that they do not reflect the most recently available EST and cDNAs. GeneSeqer also allows one to improve gene prediction accuracy by including homologous ESTs from non-native origin (other organisms). Potential for improved gene prediction has been demonstrated using two separately annotated loci on *Arabidopsis* chromosome 5: At5g62600 and At5g62590. Researchers used spliced alignment of both native and nonnative resources to predict that these two loci actually correspond to a single gene of 27 exons[11], which would have been missed using just ab initio gene prediction programs.

In 2005, a GeneSeqer with the ability to align sequences with non-canonical splice sites (GC donor site) was developed. Approximately 1-2% of introns are non-canonical in *Arabidopsis* and rice and modeling of GC donor splice sites thus enables detection of these introns, significantly improving gene structure prediction.[13]

There are also various other ab initio programs and splice alignment programs. A common method to enhance gene prediction power is to combine various programs or to combine various methods into a single program. For example, the FGENESH+ program discussed uses both the HMM along with a BLAST to known proteins. A new program known as HaMStR also combines HMMs with a BLAST in order to make use of EST transcripts for ortholog identification.[14] There are also programs such as SLAM and TWINSCAN that use comparative genomics to analyze two genomes. The basic premise behind these programs is that highly conserved regions likely correspond to genes. SLAM uses an HMM combined with sequence alignment while in TWINSCAN the sequence alignment is performed first.[15]

Evaluation of Gene Prediction Programs

General Criteria

To evaluate program performance, the specificity and sensitivity of results are often analyzed at the nucleotide, exon, splice acceptor, and splice donor level.

Sensitivity (or recall) refers to the objects in a class that are correctly identified as a member of that class. The general equation is:

True Positives / (True Positives + False Negatives) where true positives are the positives correctly identified by the program and the false negatives are the ones that the program missed.

In considering exons, for example, the specificity would be the amount of correctly identified exons divided by the total actual exons in the sequence (it is thus important to have a reliable test set).

Specificity (or precision) how many objects identified as a member of a class actually belong to that class. The general equation is:

True Positives/ (True Positives + False Positives) where the true positives are the positives correctly identified by the program and the false positives are incorrectly identified as being part of the class. In considering exons, for example, the specificity would be the amount of correctly identified exons divided by the total exons the program identified.[1] Exon sensitivity (SN) and specificity (SP) tends to be much lower than nucleotide SN and SP because both exon boundaries must be correctly identified in order for the exon to be considered a true positive. Similarly, SN and SP values for a gene tend to be even lower because every exon must be correctly specified in order for the gene to be considered a True Positive.[3] Generally, SN and SP values are calculated at the splice site, nucleotide, and exon level rather than the gene level.

Results from 3 gene prediction programs on a 20kb sequence from *Arabidopsis*

To look at the output generated by various gene prediction programs, I selected a 20kb region from the *Arabidopsis* genome and put it into three different programs. I used the forward strand of chromosome 2 from bases 11207501 to 11230701. I ran the following programs (versions of which have been trained on *Arabidopsis*): FGENESH, GENSCAN, and GeneSeqer@PlantGDP. I compiled the results (exon boundaries) for the first three genes in the region as shown in Supplemental Table 1. All these programs analyze the forward and reverse strands. Graphical output for all three tables is shown in supplemental figures 1-3. The location numbers correspond to the actual sequence (see final pages of this paper) inputted into the program and not the actual location on the chromosome.

One major difference is that the ab initio programs (GENSCAN and FGENESH) predicted additional genes than did the spliced alignment. GENSCAN and FGENESH both predict a single exon gene from approximately 19 to 20.5 kb of the input sequence. There appears to be no corresponding EST data for this, as GeneSeqer does not show any EST/cDNA alignment for this region (ESTs shown in red in this program). With this data alone, it is difficult to say for sure if this region corresponds to a gene. It could be a gene that GeneSeqer missed it because it is lowly expressed or expressed under only certain conditions of specific tissues, and thus there is no corresponding EST or cDNA. In this case, it would be considered a false negative and would lower the GeneSeqer sensitivity score. On the other hand, it could be a pseudogene that does not actually encode a gene in *Arabidopsis*. In this case, it would be a false positive and would lower the specificity score of GENSCAN and FGENESH.

Additionally, the ab initio programs both predict a two exon gene from 21.5 to 23kb of the input sequence. GeneSeqer does not shown a prediction for a protein (proteins are displayed in orange) but does show a very small EST alignment that would correspond to one of the two exons, without showing any sequence or predicted gene sequences for the other exon region. Thus, there is very little EST and cDNA data for this gene and GeneSeqer cannot make a prediction. Alternatively, it could be another pseudogene that encodes a short transcript but is not translated.

All three programs predicted the first gene in this region, AT2g26330. This is a characterized gene consisting of 27 exons with a known cDNA sequence. The GeneSeqer identified all 27 exons, which is expected given that the cDNA for AT2g26330 was used in the spliced alignment (can be seen in red). Notably, GENSCAN identified substantially fewer exons (17 exons) in this region than did FGENESH (26 exons). In this region, the exon sensitivity is lowest for GENSCAN. Many of the exons predicted by GENSCAN are also not properly defined in terms of their boundaries. For several genes (including 1-3), GeneSeqer was able to make strong alignments because there was a lot of EST and cDNA data corresponding to the region. I compiled a table with a list of the exons and their boundaries for these genes (see supplemental table 1). Based on the sequences observed, it appears that overall FGENESH did better than GENSCAN. Based on the cDNAs and the output for GeneSeqer, most of the exons in the gene are 72 bp long. GENSCAN missed several of these and it also missed some smaller exons (31 and 48 bp) in other genes as well. These results appear consistent with a paper showing that GENSCAN is relatively poor (compared to other programs such as FGENES) in its ability to identify smaller exons ranging

from 0 to 74bp.[3] The differences were most pronounced in the paper, however, with very small exons (24 bp or less). Notably, in my sequence, GENSCAN missed a couple medium sized exons (159 and 133) so a strong trend is not apparent with this small data set.

GENSCAN did not accurately predict the final stop codon containing exon for the AT2g26330 locus. It ended this exon prematurely (1223-875 rather than 1223-867 on the reverse strand). This improper boundary means that this exon is not classified as a true positive, although the correct nucleotides in this sequence can still contribute to the nucleotide sensitivity score. After ending the exon prematurely (likely due to a false positive identification of a splice donor site), it then added a final exon for which there is no EST evidence (a false positive for both the exon and nucleotides in the sequence). For the most part, FGENESH exons correspond very well with EST-based predictions, although occasionally an exon boundary is misspecified.

The spliced alignment method relies heavily on the available EST and cDNA data. For the results shown above, GeneSeqer was able to identify 27 exons for locus AT2g26330, which corresponds to the known number of exons based on cDNA. It is a given that GeneSeqer should predict the AT2g26330 exons with great accuracy because the entire cDNA is used in the spliced alignment. Because of this, I reran the GeneSeqer program using only a subset of the spliced alignment data (removing cDNA sequences and leaving ESTs only). In this second run, which still included several dozen ESTs corresponding the AT2g26330 locus, GeneSeqer predicted a total of 6 gene regions (rather than the previous 5) and split the AT2g26330 gene into two (please see supplemental figure 4). This makes sense because there is a large region in the middle of the gene without any EST sequences. In this region, GeneSeqer identified one gene with 11 different exons and another gene with 12, and therefore no longer performed better than the HMM-based ab initio gene prediction software programs (GENSCAN identified a single gene of 17 exons in the region and FGENESH identified a single gene of 26 exons). Thus, the spliced alignment program relies heavily on large amounts of EST data (for larger genes at least) and these ESTs are not likely to be available for new genomes. GeneSeqer enables the user to select ESTs from other plant species, which may help to get around this problem. While GeneSeqer functions well for annotation of cDNAs, the use of GeneSeqer for identification of new genes would likely be enhanced by combining the EST splice alignment with an ab initio HMM based algorithm.

In order to investigate the importance of organism specific training, I ran my sequence in programs trained for different species. Ab initio gene prediction programs are trained on specific

organisms because the genome composition can vary significantly from one organism to another, especially for those more evolutionarily distant. (Here, the genome composition refers to the overall and regional GC content, the nucleotide composition in signal regions, and codon bias in ORFs, as previously discussed). There are three versions of GENSCAN on the MIT GENSCAN website: one for *Arabidopsis*, one for Maize, and one for Vertebrates. To look at the importance of training, I ran my 20kb *Arabidopsis* sequence through all three versions of GENSCAN. As previously discussed, the *Arabidopsis* trained GENSCAN identified a total of 7 genes, and for the verified 27-exon AT2g26330 locus, it identified 17 exons, the majority of which were true positives. The vertebrate version on GENSCAN was able to identify 3 exons in the AT2g26330 locus and some exons from other genes as well (15 exons predicted in total for the entire 20kb region, see supplemental figure 5). Interestingly, the Maize version was only able to identify 1 exon for the AT2g26330 locus and a total of 5 exons overall (see supplemental figure 6). Three of these exons do not have any existing EST data (see GeneSeqer output) and do not correspond to any genes predicted by the other programs. They are likely false positives. Thus, training is very important and the *Arabidopsis* gene prediction (for this specific 20kb sequence) is even worse on a Maize trained GENSCAN than on a vertebrate trained GENSCAN. Notably, Maize is a monocot and thus very far diverged evolutionarily from the dicot *Arabidopsis*. Maize also has a distinct GC composition and gradient from the 5' to 3' end of individual genes that is not present in most other organisms that can be difficult to model and integrate into prediction programs during training.[4,16] Maize also has a large amount of transposable elements which might also have profound effects on genome structure that could impact gene prediction. In closely related organisms, the genome will be more similar in composition and structure and there is thus the potential to use a program trained on one organism for another. For example, Rogic *et al* 2001 used programs trained with human sequences on both human and murine test sets and observed only marginal differences in specificity and sensitivity.[3]

The conclusions from my data set are generally limited because it is a small sample size and my test set (those genes in the sequence with complete cDNAs, as splice aligned by GeneSeqer) may have been part of the training set for the development of these programs. In order to assess the accuracy of a program, it is necessary to use a test set that does not overlap with the training set. Oftentimes, the program makers do not disclose both the exact training and test sets. If the training set and test set overlap, then this distorts the data in such a way that the program appears to have a greater accuracy rate than actual. For example, an overlap between the training set and test set data

increased the exon accuracy of an ab initio gene prediction program from 81% to 86%.[1] There have been several previously published studies that have been careful to avoid these pitfalls by selecting genes that were not part of the training set.

Published Studies on the Accuracy on Gene Prediction Programs

Studies to evaluate specificity and sensitivity of gene prediction programs have been conducted in both animals and plants. Rogic *et al* 2001 analyzed mammalian gene prediction (human, mouse and rat) in the FGENES, GeneMark.hmm, Genie, GENSCAN, HMMgene, Morgan, and MZE using a 195 gene test set. They had mRNA sequences for these genes and these sequences were very unlikely to be part of the training set for the programs because they had been entered into GenBank after the programs were developed and trained. Of the programs, they found that overall HMMgene provided the greatest exon SN and SP with values of 0.76 and 0.77 respectively, with the next best program being GENSCAN.

The study in mammals (human, mouse, rat) included a large test set of 195 genes and the researchers took advantage of this to determine each program's performance relative to numerous gene factors including G+C content and exon length. They found that GENSCAN performed its best (in terms of exon SN and SP) when exons had a GC content of less than 40% whereas GeneMark.hmm performed its best for those exons with a GC content of 40-60%. They also found program specific differences in accuracy relative to exon length and exon type (initial, terminal, internal, and signal).[3] For example, some programs have a tendency to under-predict small exons while others tend to over-predict them. It is hard to pinpoint the basis for these differences. It seems likely that different algorithms provide better accuracy for different conditions, but these differences in accuracy might also be attributed to the actual training sets used in the optimization process.

Studies to evaluate ab initio prediction software have also been conducted in plant sequences. Pavy *et al* conducted a study in 1999 with *Arabidopsis* sequences, in which they discarded mRNA sequences that were publicly available during the training of the programs they were testing. Using the sequences for 168 genes, they found that the best ab initio program was GeneMark.hmm, which outperformed programs including GENSCAN, GRAIL, and FGENESP. GeneMark.hmm had an exon SN of 0.82 and an exon SP of 0.77 (see Tables 1-4 for more details). These values are therefore better than the best performing program in the 2001 Rogic *et al* study in mammals. In 2005, Yao *et al* evaluated five ab initio programs on eight maize genes for which they had obtained cDNA that was not publicly released prior to their study. Out of FGENESH,

Table 1: Nucleotide Accuracy

Organism:	Mammals (human, rat, mouse) (Rogic *et al* 2001)		*Arabidopsis* (Pavy *et al* 1999)		Maize (Yao *et al* 2005)	
Program	SN	SP	SN	SP	SN	SP
GENSCAN	0.95	0.90			0.81	0.95
GeneMark.hmm	0.87	0.89			0.92	0.93
FGENESH					0.97	0.94
FGENESP						
FGENES	0.86	0.88				

Table 2: Exon Accuracy

Organism:	Mammals (human, rat, mouse) (Rogic *et al* 2001)		*Arabidopsis* (Pavy *et al* 1999)		Maize (Yao *et al* 2005)	
Program	SN	SP	SN	SP	SN	SP
GENSCAN	0.70	0.70	0.63	0.70	0.54	0.81
GeneMark.hmm	0.53	0.54	0.82	0.77	0.69	0.80
FGENESH					0.86	0.88
FGENESP			0.42	0.59		
FGENES	0.67	0.67				

Table 3: Acceptor Site Accuracy

Organism:	Mammals (human, rat, mouse) (Rogic *et al* 2001)		*Arabidopsis* (Pavy *et al* 1999)		Maize (Yao *et al* 2005)	
Program	SN	SP	SN	SP	SN	SP
GENSCAN	0.87	0.80	0.73	0.78	0.53	0.86
GeneMark.hmm	0.81	0.75	0.90	0.84	0.71	0.85
FGENESH					0.91	0.93
FGENESP			0.55	0.70		
FGENES	0.80	0.77				

Table 4: Donor Site Accuracy

Organism:	Mammals (human, rat, mouse) (Rogic *et al* 2001)		*Arabidopsis* (Pavy *et al* 1999)		Maize (Yao *et al* 2005)	
Program	SN	SP	SN	SP	SN	SP
GENSCAN	0.90	0.84	0.77	0.82	0.56	0.93
GeneMark.hmm	0.82	0.78	0.93	0.81	0.77	0.92
FGENESH					0.91	0.91
FGENESP			0.58	0.72		
FGENES	0.85	0.82				

Tables 1-4. The above tables show results compiled from 3 different studies for select ab initio prediction programs. SN= sensitivity and SP = specificity. Note that Pavy *et al* did not provide this information at the nucleotide level. Also, in Rogic *et al*, HMMgene performed better than GENSCAN on some parameters and so the greatest exon SN and SP were actually 0.76 and 0.77.

GeneMark.hmm, GENSCAN, GlimmerR, and Grail, they observed the highest exon SN (0.86) and exon SP (0.88) with FGENESH (see Tables 1-4 for more details).

These three papers (Pavey *et al* 1999, Yao *et al* 2005, and Rogic *et al* 2001) all used GENSCAN, GeneMark.hmm, and a version of the FGENE program and this subset of results has been compiled into Tables 1-4. The different FGENES programs are listed separately because the most recent FGENESH uses a different algorithm (see Methods section above). Because these papers were published in different years, it is possible that updated versions of the programs were used in more recent studies. Overall, however, it does look like some programs may be best equipped for different organisms. The GeneMark.hmm in the earliest paper (on *Arabidopsis*) performed better on most parameters than it did on different species in later papers. The best exon SN and SP in the mammal paper was 0.76 and 0.77 (HMM gene) and in the *Arabidopsis* paper the best was 0.82 and 0.77 (GeneMark.hmm). In maize, the best exon SN and SP values (0.86 and 0.88) was with the HMM based FGENESH which was not used in the earlier papers.

An additional study in rice also found that FGENESH significantly outperformed GENSCAN and GeneMark (in addition to also outperforming RiceHMM and GlimmerM), although these researchers were not confident in whether their test sets were distinct from the training sets used in the development of the programs.[4] Rice and Maize are both monocots and potentially FGENESH may be better equipped to predict genes in these organisms, which the authors of the rice study think may be attributed to the fact that FGENESH puts more weight on signal terms than content terms (such as codon usage). Rice can be generally difficult for ab initio gene prediction because rice has a gradient of GC content from its 5' to 3' end which gene prediction programs generally do not model.[4] FGENESH also performed well in the 20kb *Arabidopsis* sequences that I analyzed, and so it may predict well in a wide number of organisms.

Overall, optimization of ab initio for a given species (or closely related species) is important. Rogic *et al* 2001 found that there were only marginal differences in SN and SP for human and murine exons when using human trained gene prediction programs. [3] On a more general level, however, phylogenetics may not predict the best gene prediction program for a new genome and optimization for a specific species will most often give the best results.[7] For example, *Arabidopsis* sequences may not be better predicted with a Maize optimized program than they are with a vertebrate optimized program, as shown previously with the 20kb region from *Arabidopsis* chromosome 2 run with GENSCAN. One would not expect all programs to perform equally well

across all species due to differences in genome composition and local nucleotide differences in signal sensor regions. There are differences in GC content and how this content is distributed throughout the genome. Typically the GC content is higher within coding regions. Monocots are distinct in that they not only have increased GC content within coding regions but this GC content also changes within the gene in a stereotyped fashion. Monocots, such as rice, have a negative GC content gradient within genes, meaning that within a typical rice gene, the 5' end typically has a much greater G+C content than does the 3' end (See Supplemental Figure 8). Monocots also have a similar gradient in terms of codon bias and amino acid usage. Gene prediction programs will need to consider organisms with gradients differently than organisms that do not have these gradients, such as animals and dicot plants.

Differences in eurkaryotic gene prediction then do not necessarily fit neatly into plant and animal categories. A study recently showed that *Arabidopsis* optimized GENSCAN outperforms vertebrate optimized GENSCAN in predicting exons in sponges.[17] The *Arabidopsis* optimized GENSCAN had very high SN and SP values for sponge exons (0.83 and 0.79 respectively), which is even better than the exon SN and SP that Pavy *et al* 1999 observed with GENSCAN using *Arabidopsis* sequences (0.63 and 0.70 respectively). In the sponge paper, however, it should be noted that only 18 known gene sequences were used and these were inputted into the program with little additional intergenic regions,[17] both of which could skew SN and SP values.

Conclusions:

It seems likely that different programs may fit different organisms better in terms of algorithms to identify signal sequences and open reading frames. There is also the issue of training sets and whether or not they were truly representative of the whole genome (with large training sets, this is likely to be less of an issue). On the flip side, there could also be caveats used in the test sets used to evaluate the programs. That is, it could be that genes used in the test set may not represent the average gene and thus do not reflect the overall accuracy of the program. The output for a given program depends both on the algorithms used as well as the training set used for optimization. On a general level, ab initio gene prediction does not appear to correlate well with phylogenetics[7], with the possible exception being very closely related organisms.

A complete set of cDNAs would allow one to determine the exact annotation for all genes in a genome. Establishing a complete cDNA library is not only very expensive, it is also likely to be technically impossible due to spatial and temporal expression differences among genes and the fact

that some genes may lowly expressed or only expressed under very specific conditions. Different programs have their caveats. Splice alignments rely on sporadic and perhaps improperly sequenced ESTs. Complete annotation with DNA sequence alignment depends on genes having recognizable homologs in other species. The ab initio programs use signal sensors combined and statistics to identify genes within a sequence, which conflates the complex processes of gene identification and transcription within the biological cell.[1] Notably, ab initio programs also rely on a representative training set which generally relies then on representative cDNA and EST data. These must then be acquired for proper optimization, or an ab initio program that has not been optimized for the specific genome must be used. One exception is the Viterbi training in the GeneMark.hmm which allows for self-training of a given sequence. For new genomes, it might be useful to combine the GeneMark.hmm self training programs with a BLAST search for ortholog alignment to allow for further refinment.

There are also methods implementing comparative genomics for gene identification. A 2005 study compared the *Arabidopsis* genome sequence to the partial genome draft of *Brassica oleracea*. They found numerous conserved regions in previously unannotated regions. They used this information to isolate cDNA from various regions and were able to confirm several sequences allowing for 21 novel gene models.[18] It seems likely that these methods (and others as well) will be predisposed to miss genes that are under positive selection in a given organism.

One common strategy to get around the limitations of individuals programs is to use the programs in combination. For example, when Pavy *et al* 1999 analyzed program on *Arabidopsis* sequences, they combined the highest HMM program, GeneMark.hmm, with the NetGene2 to allow for better splice prediction and were able to increase the splice site donor specificity to 0.94 (compared to 0.81 for GeneMark.hmm and 0.31 NetGene2 on its own).

Another possibility to improve gene prediction is to develop models on subsets of genes. It has been shown that different programs perform better or worse for particular genes. The Rogic *et al* 2001 paper documented this based on GC content and exon length and type. Brendel and Zhu have noted that differences in prediction power for different genes is a likely indication that current models are too general and that prediction could be improved if models were trained on subset of genes to optimize parameters for different gene classes.[8] Some of the major problems facing gene prediction are the presence of pseudogenes and alternative splicing. Overall, however, gene prediction at the exon level is very good and new methods promise to increase the accuracy further.

References:

1. Majoros, WH. *Methods for Computational Gene Prediction*. Cambridge: Cambridge University Press; 2007.

2. Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS. Unequal genetic redundancies in Arabidopsis--a neglected phenomenon? *Trends Plant Sci*. 2006;11(10):492-498.

3. Rogic S, Mackworth AK, Ouellette FB. Evaluation of gene-finding programs on mammalian sequences. *Genome Res*. 2001;11(5):817-832.

4. Yu J, Hu S, Wang J, et al. A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science*. 2002;296(5565):79-92.

5. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*. 2005;33(20):6494-6506.

6. Cho Y, Walbot V. Computational methods for gene annotation: the Arabidopsis genome. *Current Opinion in Biotechnology*. 2001;12(2):126-130.

7. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5(1):59.

8. Brendel V, Zhu W. Computational modeling of gene structure in Arabidopsis thaliana. *Plant Mol. Biol*. 2002;48(1-2):49-58.

9. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. 1997;268(1):78-94.

10. Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. *Genome Res*. 2000;10(4):516-522.

11. Schlueter SD, Dong Q, Brendel V. GeneSeqer@PlantGDB: Gene structure prediction in plant genomes. *Nucleic Acids Res*. 2003;31(13):3597-3600.

12. GeneSeqer. Available at: http://www.plantgdb.org/tool/GeneSeqer/help.php [Accessed December 8, 2009].

13. Sparks ME, Brendel V. Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics*. 2005;21 Suppl 3:iii20-30.

14. Ebersberger I, Strauss S, von Haeseler A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol. Biol*. 2009;9:157.

15. Brent MR, Guigó R. Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol*. 2004;14(3):264-272.

16. Wong GK, Wang J, Tao L, et al. Compositional Gradients in Gramineae Genes. *Genome Research*. 2002;12(6):851-856.

17. Stifanic M, Batel R. Genscan for Arabidopsis is a valuable tool for predicting sponge coding sequences. *Biologia*. 2007;62(2):124-127.

18. Ayele M, et al. Whole genome shotgun sequencing of Brassica oleracea and its application to gene discovery and annotation in Arabidopsis. *Genome Research*. 2005;15(4):487-495.

PROGRAMS ACCESSED:

GENSCAN:
http://genes.mit.edu/GENSCAN.html

FGENESH

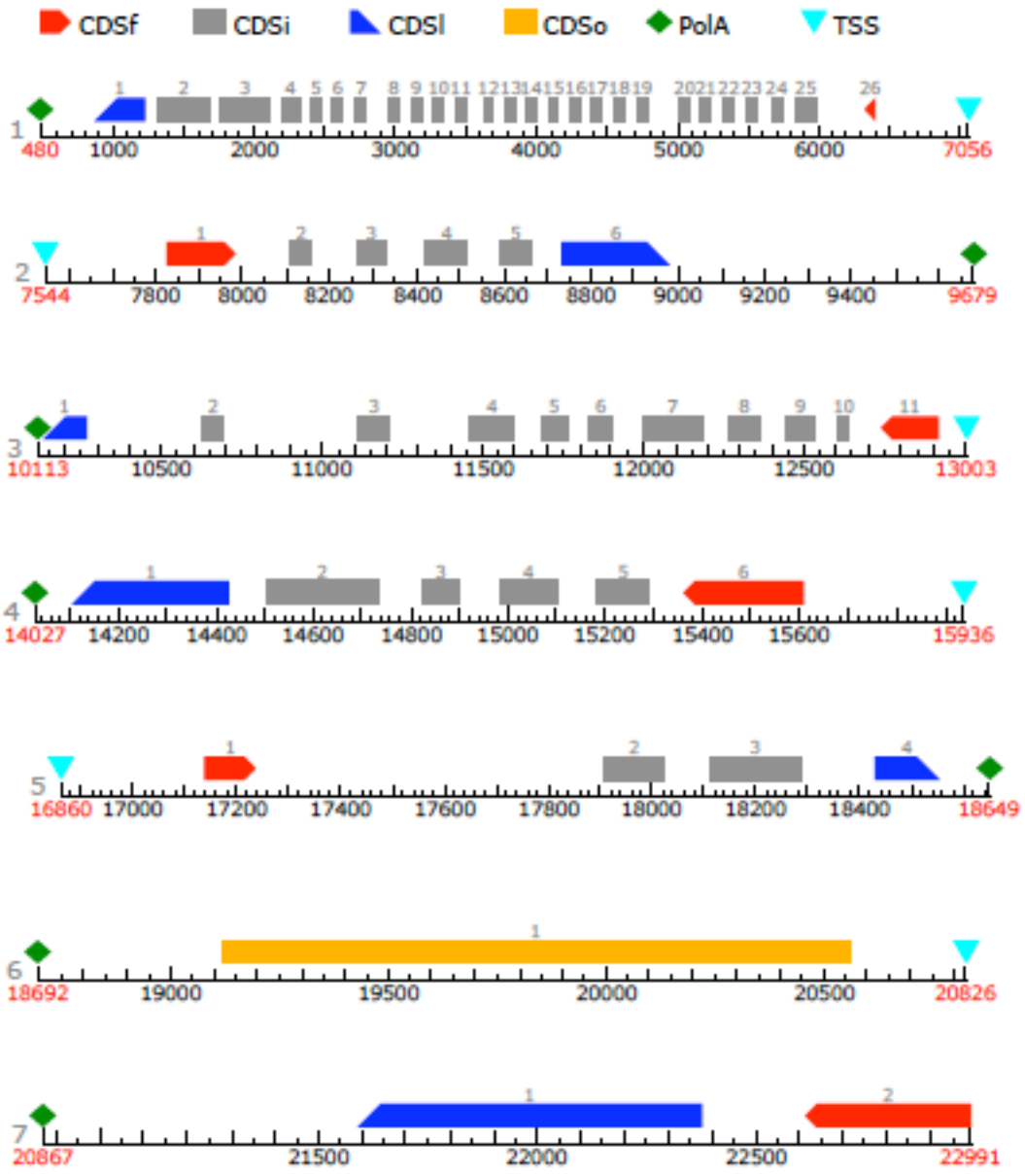
http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind

GenSeqer@PLANTGDP

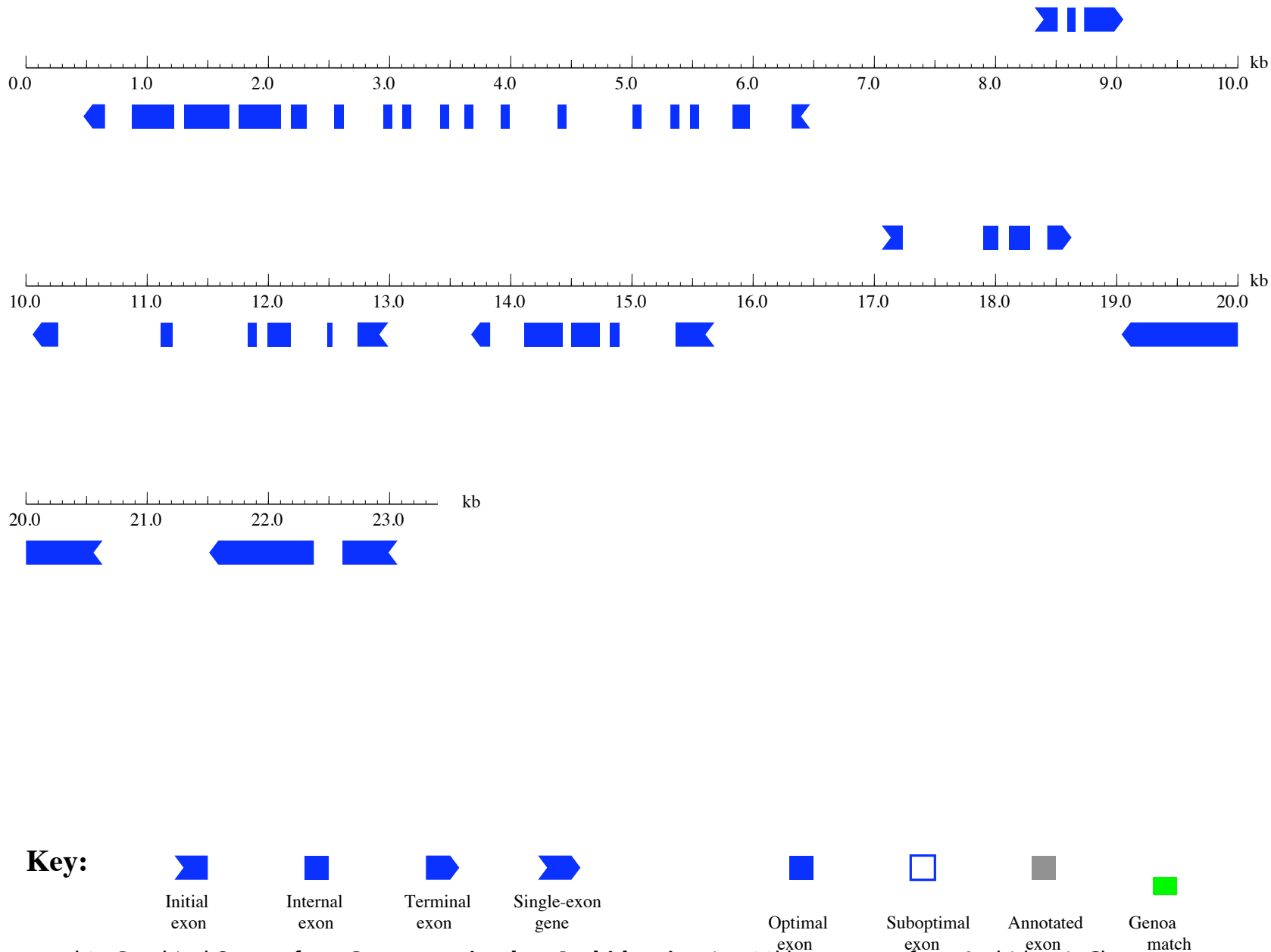
http://www.plantgdb.org/cgi-bin/GeneSeqer/index.cgi

GeneMark.hmm:
http://exon.biology.gatech.edu/eukhmm.cgi

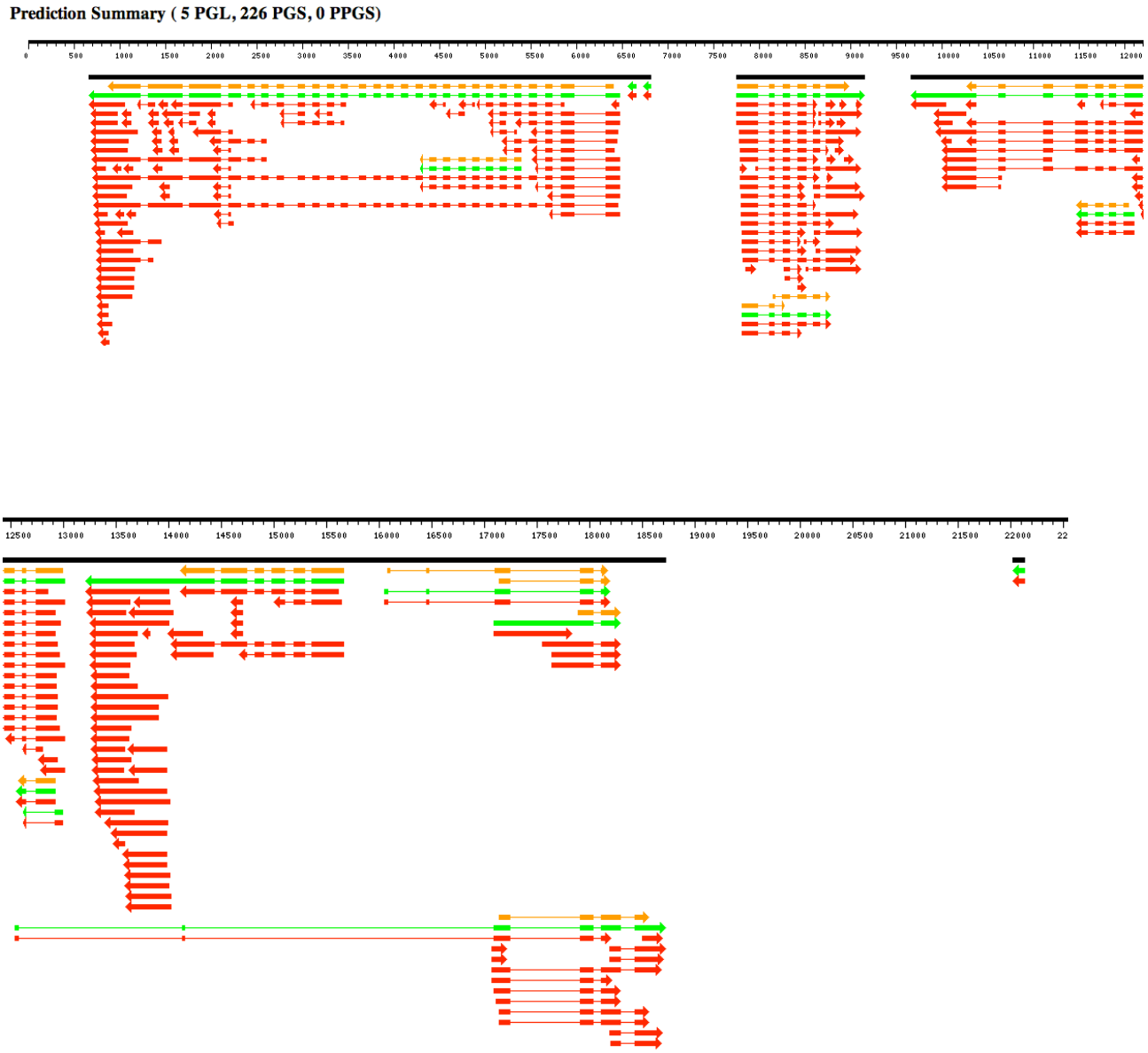
FGENESH 2.6:



Supplemental 1: Graphical Output for FGENESH analysis of a 20kb region on *Arabidopsis* chromosome 2. In addition to exons, FGENESH also provides information about the transcription start site and the Poly(A) tail.
Key:
CDSf - First (Starting with Start codon)
CDSi - internal (internal exon),
CDSl - last coding segment, (ending with stop codon);
TSS - Position of transcription start (TATA-box position and score)
FGENESH provides a text output of boundaries for each of the exons and a statistical weight for the features shown.



Supplemental 2: Graphical Output from Genscan **trained on Arabidopsis** using 20kb sequence from Arabidopsis Chromosome 2. The coding segments for predicted gene exons are shown in blue.



Supplemental 3: Graphical Output for GeneSeqer@PlantGDP analysis 20kb region on *Arabidopsis* chromosome 2. The orange corresponds to the predicted protein sequence and the green to the gene structure. Note that for many genes the green extends further for the initial or terminal exon, and thus the program shows some 5' and 3' UTRs. The red is the alignment from the various EST and cDNA in the database.
The primary predicted protein (orange) and gene (green) structures are shown along the top but for some alternative structures are also shown. The alternative structures also suggest the possibility of an overlapping gene, not present in the other prediction programs. For this input, GeneSeqer also supplied over 150 pages of alignments corresponding to the ESTs/cDNA and predicted structures with the scores. It also provided text output for the predicted boundaries of exons.

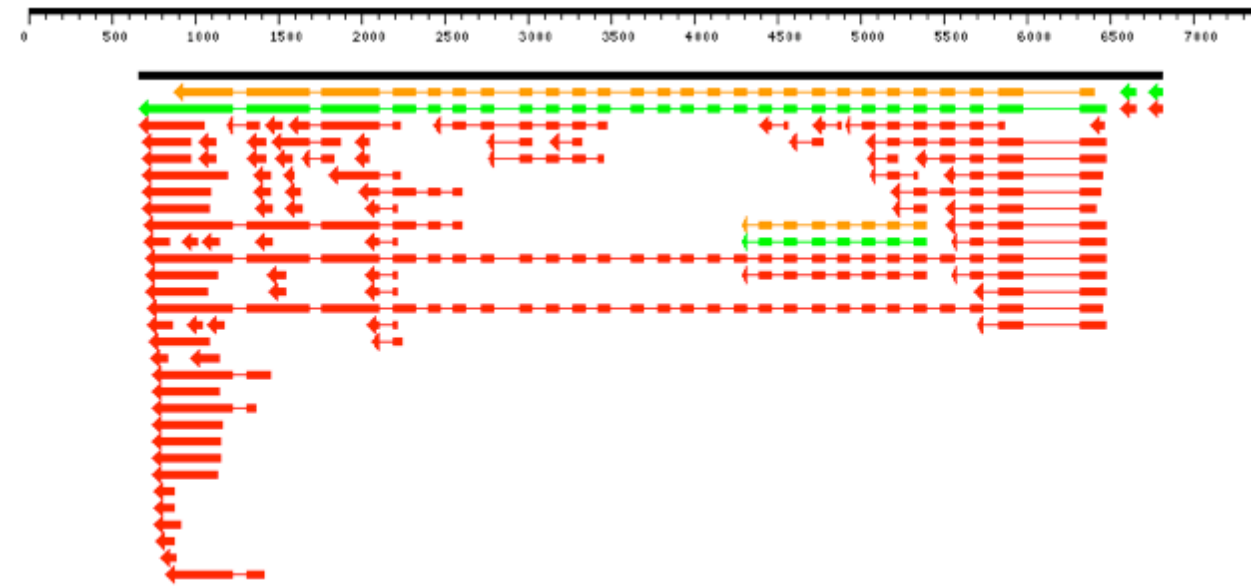
Gene:	FGENESH			GenScan			GeneSeqer		
1	6395-6317	5970-5832	5730-5659	6395-6317	5970-5832		6471-6317	5970-5832	5730-5659
	5554-5483	5391-5320	5229-5158	5554-5483	5391-5320		5554-5483	5391-5320	5229-5158
	5079-5008		4776-4705	5079-5008			5079-5008	4931-4860	4776-4705
	4611-4543	4457-4386	4309-4238		4457-4386		4611-4543	4457-4386	4309-4238
	4149-4078	3988-3917	3848-3777		3988-3917		4149-4078	3988-3917	3848-3777
	3689-3618	3491-3420	3337-3266	3689-3618	3491-3420		3689-3618	3491-3420	3337-3266
	3179-3108	3021-2950	2785-2717	3179-3108	3021-2950		3179-3108	3021-2950	2785-2717
	2621-2550	2471-2398	2315-2187	2621-2545		2315-2187	2621-2550	2471-2398	2315-2187
	2102-1755	1678-1308	1223-867	2102-1755	1678-1308	1223-875	2102-1755	1678-1308	1223-661
				652-549					
2	7823-7982	8109-8157	8259-8329				7752-7982	8109-8157	8259-8329
	8419-8513	8592-8659	8731-8980	8397-8513	8592-8659	8731-8990	8419-8513	8592-8659	8731-9147
3	12915-12735	12636-12605	12529-12440	12915-12735		12529-12487	13007-12735	12636-12605	12529-12440
	12357-12263	12183-11994	11899-11830		12183-11994	11899-11830	12357-12263	12183-11994	11899-11830
	11759-11681	11591-11458	11206-11112			11206-11112	11759-11681	11591-11458	11206-11112
	10692-10627	10266-10129			10266-10112		10692-10616	10369-9658	

Supplemental Table 1: Summary of exon boundaries for the first three genes from a 20kb *Arabidopsis* sequence as predicted by FGENESH, GENSCAN, and GeneSeqer@PlantGDP. All programs have been trained on *Arabidopsis*.

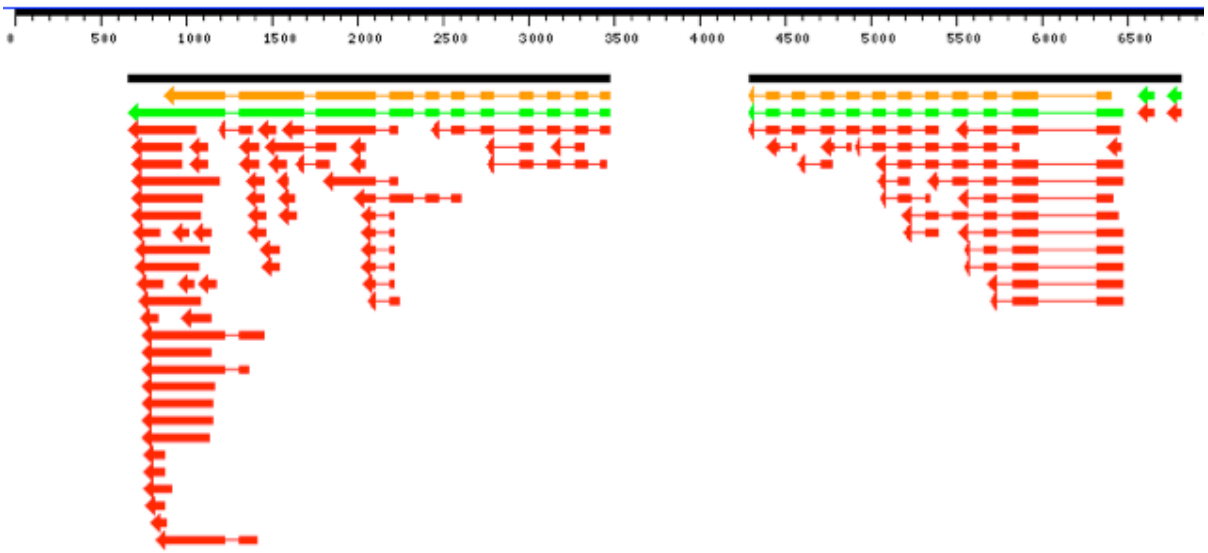
The exons are presented in order from the initial exon to the terminal exon. Predicted Genes 1 and 3 are on the reverse strand while Gene 2 is on the forward strand.

Note that the initial and terminal exons for GeneSeqer extent on the 5' and 3' boundaries respectively because the GeneSeqer provided these boundaries based on the full exon (including UTRs) and not just the coding segments.

A.

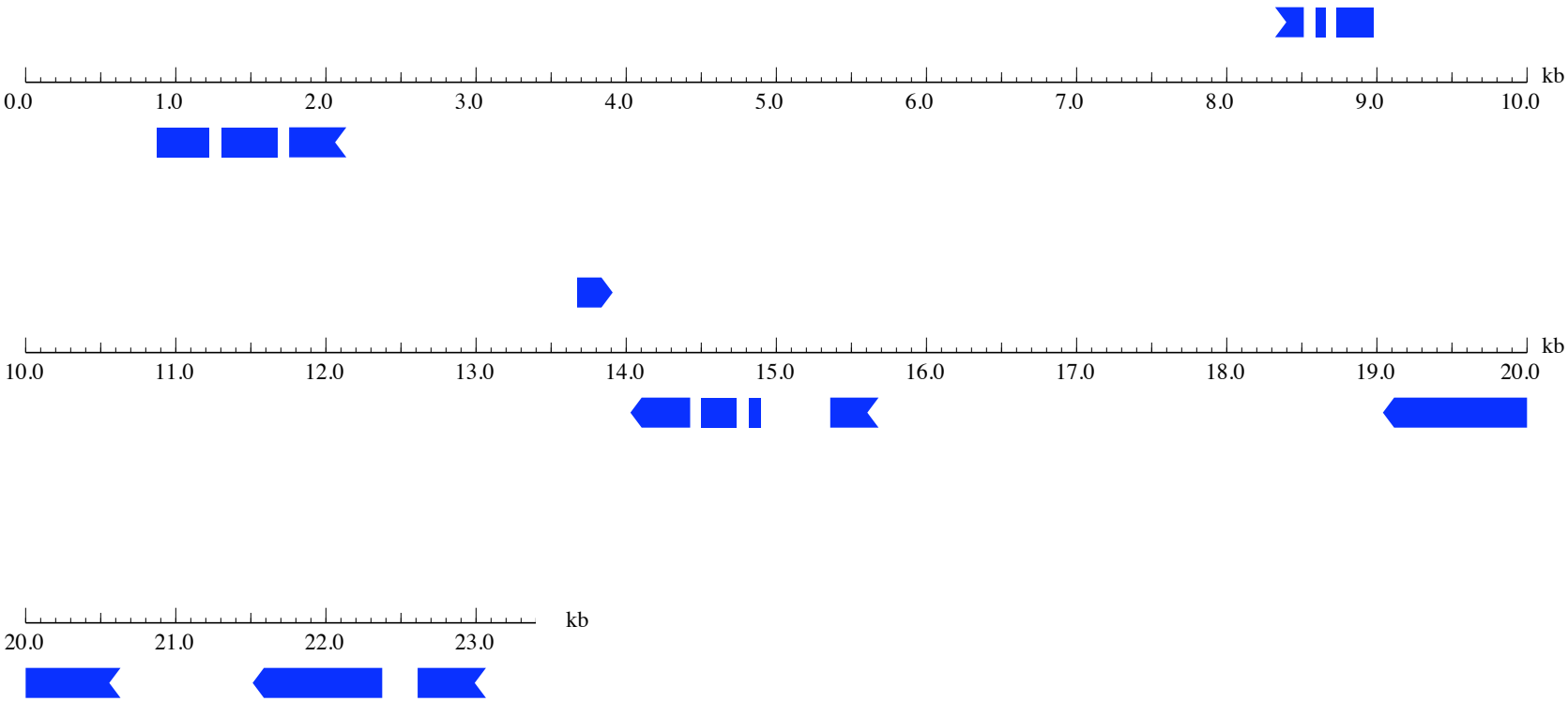


B.



Supplemental Figure 4.
Geneseqer run for locus AT2g26330 (sequenced AT2g26330 cDNA has 27 exons). In A. the entire available *Arabidopsis* EST and cDNA databases were selected to run in the analysis while in B. only a subset of the database (*Arabidopsis* ESTs only and not full cDNAs) were included.
Key: (as described by Geneseqer): The orange bar corresponds to the predicted protein structure, the green bar to predicted gene structures, and the red bar corresponds the EST and cDNA spliced alignments.
Note that the predicted gene structures typically include UTR regions, as they extend further out than the predicted protein structure.

GENSCAN predicted genes in sequence /tmp/12_09_09-00:34:46.fasta



Key:

Initial exon

Internal exon

Terminal exon

Single-exon gene

Optimal exon

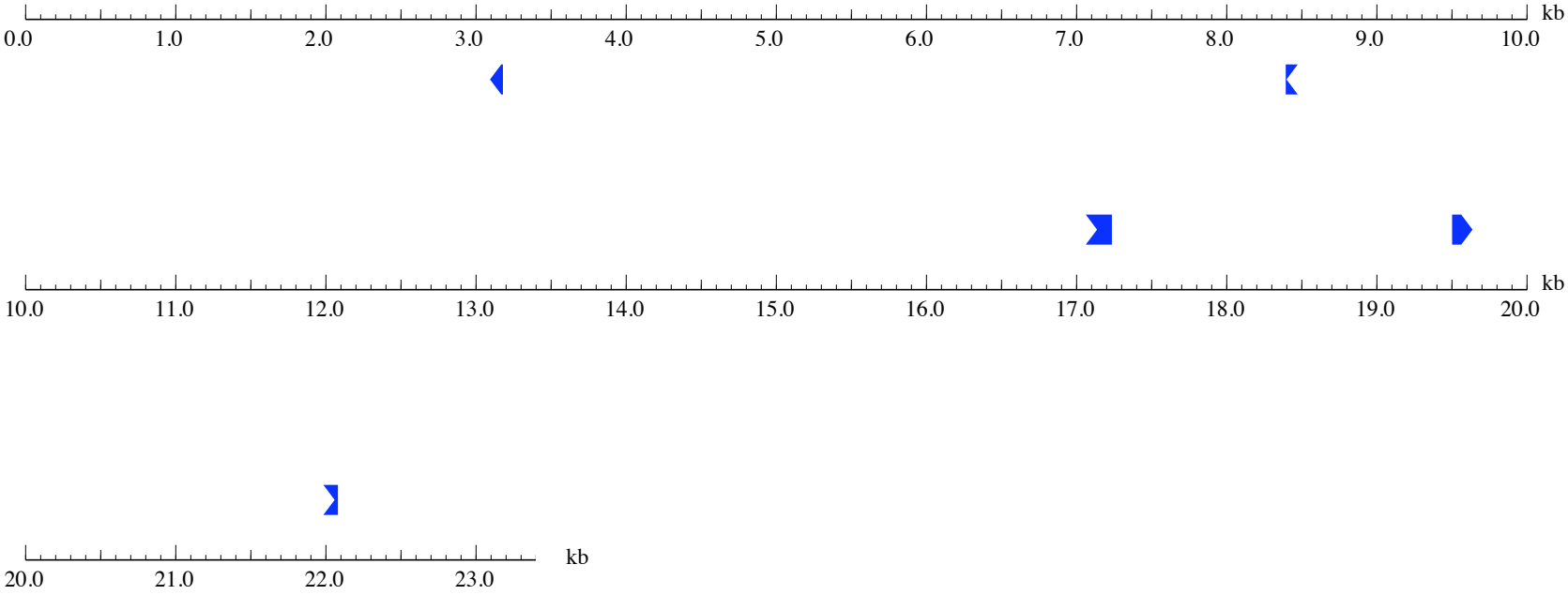
Suboptimal exon

Annotated exon

Genoa match

Supplemental 5: Graphical Output from Genscan **trained on vertebrates** using 20kb sequence from Arabidopsis Chromosome 2 (compare to Supplemental 2 showing results for same sequence run on Genscan trained for Arabidopsis)

GENSCAN predicted genes in sequence /tmp/12_09_09-18:06:34.fasta



Key:

Supplemental 6: Graphical Output from Genscan **trained on maize** using 20kb sequence from Arabidopsis Chromosome 2 (compare to Supplemental 2 showing results for same sequence run on Genscan trained for Arabidopsis)

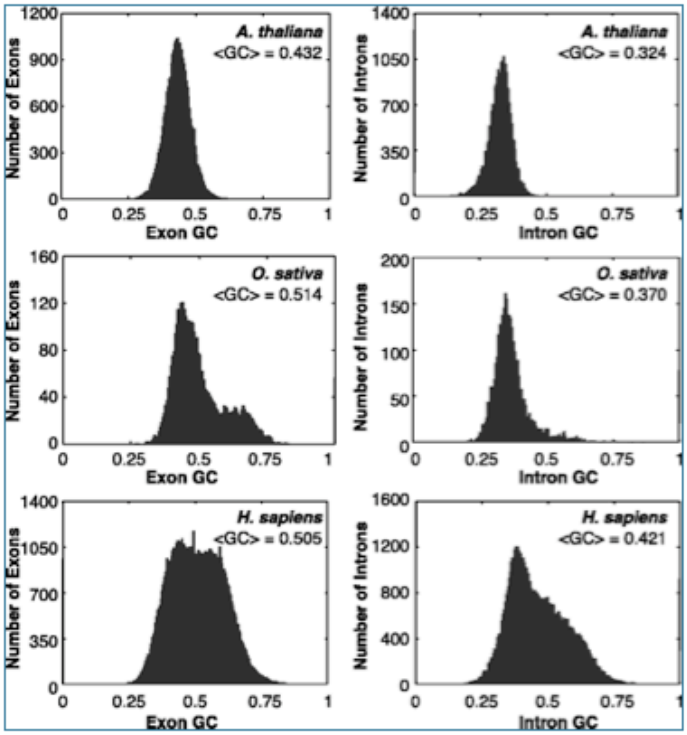


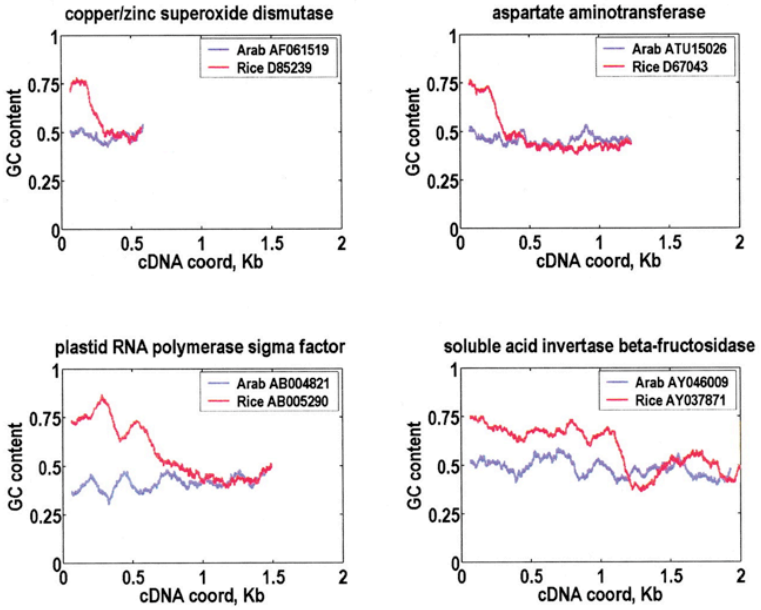
Figure 3. GC content distribution for exons and introns in *A. thaliana*, *O. sativa*, and *H. sapiens*. All exon and intron sequences were derived from cDNA-to-genomic alignments. Mean GC content is

$$\text{computed on a length} - \text{weighted basis as} <\text{GC}> = \sum_i \text{L}_i \cdot \text{GC}_i / \sum_i \text{L}_i, \text{ where } \text{GC}_i \text{ and } \text{L}_i \text{ are the GC content}$$

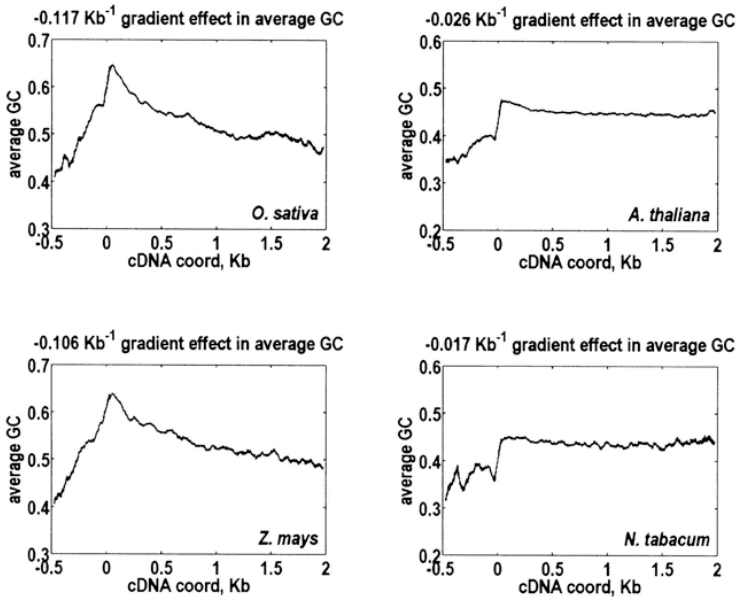
and length for the *i*th segment (exon or intron).

Supplemental Figure 7 (Reproduced from Yu *et al*. Science 296 (5565):79) GC content can vary from one organism to another and generally GC content is higher within exons than introns within a given genome. The authors note that in some organisms the average genome GC density is less than 22% (*Plasmodium falciparum*) and in others it is more than 68% (*Halobactrium*). There can also be tremendous various in local GC content within a genome.

A



B.

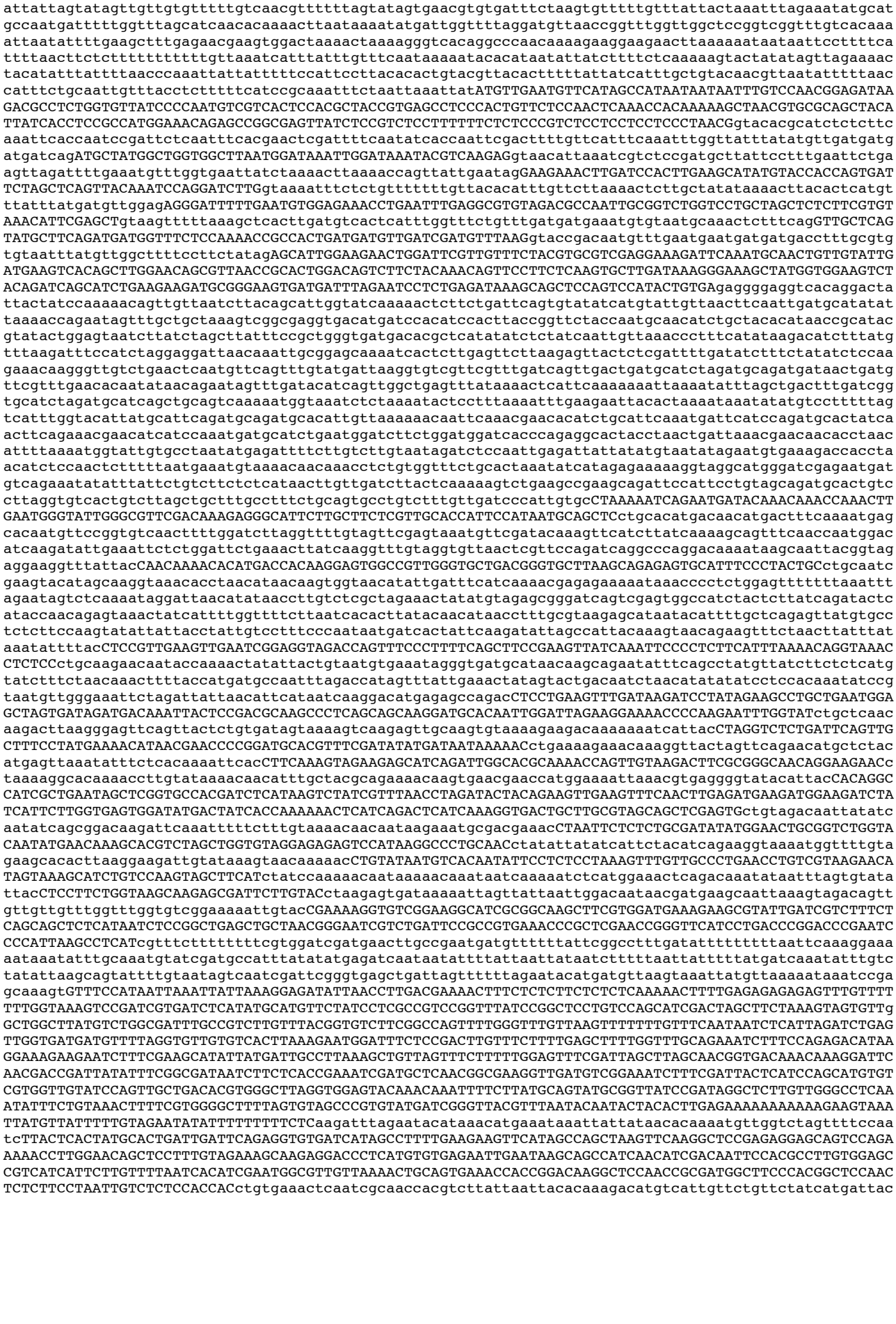


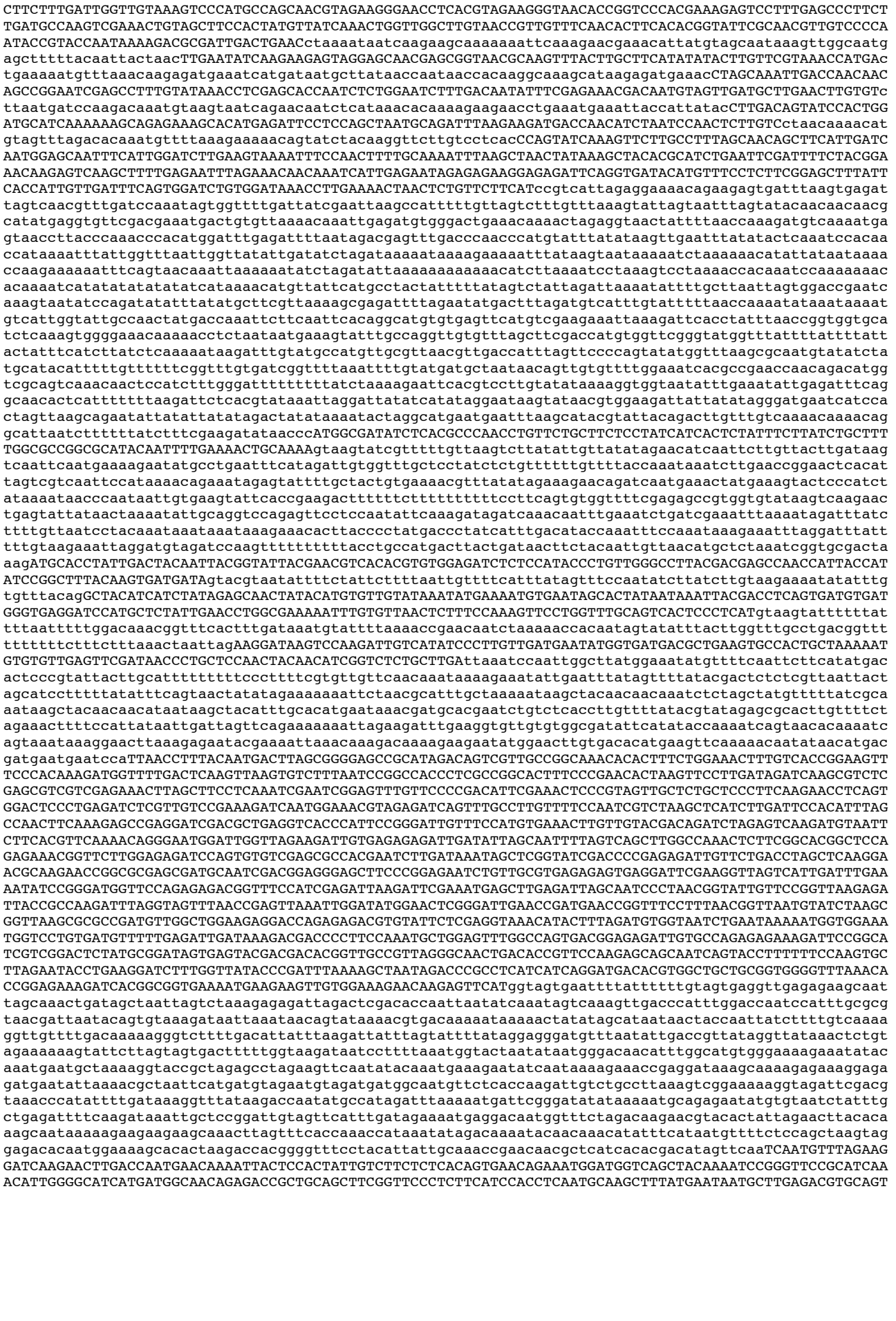
Supplemental Figure 8. GC gradient in plants. Reproduced from Wong *et al* 2002.
"A. GC content as function of position from start of coding region for four pairs of best available *O. sativa* and *A. thaliana* homologs (possible orthologs). A 129-bp sliding window, equal to the median size of a rice exon, was used to filter out the fluctuations in the sequence.
B. Overall GC content as a function of cDNA position, relative to the start of the coding region, and averaged over all cDNAs with a 51-bp sliding window. Shown here are *O. sativa, Z. mays*,*A. thaliana*, and *N. tabacum*. Negative coordinates are 5′-UTR. Positive coordinates are coding."

** Note that in B an increase in GC density occurs in the coding region in both monocots and dicots (which is typical for many organisms), but in monocots there is also a negative 5' to 3' gradient of GC density within the gene. **

Sequence: *Arabidopsis* thaliana chromosome 2 fwd strand (11207501 to 11230701)

tcaaaaatagaatttaaacttagaacaagttttgaaaagggtaggcttatcagtgggataaacaacaatcaagacgattttacaatcgcacaacagtgct
actcttctgccttcttgaaaagtcttttttttttacgtttaattatttactacttcttctttttatgtatttttttttgtagactctcgtaaatgtatac
ataaaaaccaaagttagtagattttagttgtagtaattagttcttgtctaaaatgccttttttgaaaaaaatcttaggaaaatgccatttttaaatggtg
tacggatttatctttttaacattttaaaatgtatgtatttacgattttgattttgcaaaaatacgtactttatactaaactagaaagtctcacgcgctg
tatgagttcgaattaatataagaaattataaaaaaactatataaaataaattaatcagacaatatatttgtttatcatttttattgccaagtgagaaaac
ttattaacctatcgtttgcactctcatagacgtggttgacatataacactacatagcaaaacgtttctcgtagtttagaaggttctttctatgtatatgt
ttttgagcataagaaatcttttgtggtggtgacgtggtgttgagtactagatctaagatatgttacaagaataacccaatttagaaaaacttagataggc
aagaacttataggaagtctaagtgccagtctgaagacatatttcacaccacacacttgtctaataataataaggcagtgcataatacttacatgagacat
taatttttctaacagcttaacgcaacgaaaagataccgttttaaagattctcctcctaacgaaaaaCTACTCACTGTTCTGAGAAATAACTTGTCCAAAC
CGAAGAAACAGTTGAGCATCAGAAGCACTCATGGAAGAGCAATTGACAGAATGAGGAGTCTTGAGATTTGCATACTCATCGACGTAGCACGAACCAGCCA
GCGTCGCTGACGTGTCAGTCGCAGCAGGTGGTTGTTCCGATAGCATAAAACTGCCGAGAACACGAGTCACCTGGTGCATTGTGGGTCGATCATTCGGCTG
TCTTTTGGTGCATAGGAGTGCCAGTTGGAAAACTTTCTTCACCACACCGAGATCTTTACACGTCGATGTGATGTCTGGATCTGCCATTTCCATCACTTCA
TTGTTCCCCGTCTTTGACATTATctgatgtaagtgaatcaatgaaaacaagattacggacttgacctaaacagagcagctgagagagataggcaagaaag
aacaaacCAGATGGTGGAGATTGGATTCGTCATCAACGGCTTTCCTTCGGGTTAACAACTCAAGAAGGACTATTCCATAACTGTAGACATCGGATTTCTC
AGTGAGCCGTGAAGTGCGAGCATACTCGGGGTCTATGTAACCTATCGTGCCCATCACGTAAGTTGAAGTATGTGACTTTGACACACACAAGCTTTTCGCT
ATTCCAAAATCTGTCAAACGAGCCTCTAAGTCTTTGTCCAAGAGAATGTTGGACGACTTCACGTCTCTGTGAATGATCCTTGGACTACAGTCATGGTGTA
GATAAGCTAAACCTTGTGCTGCACCATATGCTATCTTAAGCCGTGTGTCCCAATCAAGAGTCTTTTTCTTCGTAGGGCctttacaaaacacacaagcaag
ttgttatgtcacagaagattcaaataattttctatgtttggcgatgagacttacCATGAAGAAGATCCCAGAGGCTACCATTTTCCAAATAGTCATAGAA
CAGAAGACTCCCCAAGTGAGAGAGGGAATAAGCTTGTAGGCTCACAAGATTTCTGTGCTTGATGCTACTTAGCATCTCGAGTTCTGTTTCAAACTGTTTC
ATTGACTGTGGGTTGTGAGAGTAAAGCCGCTTAATCGCAACCGGTTTACAATTCTTCAAAACACATTTGTATACAGTGCTTGATGCTCCGTGCCCAATGA
TATACTTCTCACTTAGATTCTCTGTCATTCTCATGATATCCTCGTAAACGTGGAGTGCCATGTTCATATGAAGGATGACGAGCTTCGGTGTCGAATAAGT
TActgatgtcaccaggaaatttaaggattaactattggattacattgtaggtgaagaacattcgtaaagtggtttggagagtagaccTGGTTTGTCAAGT
GATCCATCAAGAAAAGGAGGAGGATTATGCGGTCGGCAAGCTGCTATTAAGACCATGAGAAGGATCACAAGTCCCCCAATAGCTATTCCAAGAATAGCTG
CTCTAGAGATTGACActgcagacaaaacatggatagaaacatgagctttaattatcttaggttttgtgaatggaagcttttagaagaatgtaatcacCTC
GTACAGTTCGACGAGAATCATGACACGGTGAGTTTAGCCAACTACCGCAAAGACCAGGATTGCCAATGAAGctttatccaaaaacataaacatagatatt
agcaaacaaaacaaaatgcacaatacccaaaacctgctaaattaccatacCTGTCTGGTGAAAATCTTGAGAAGTTATTGTTCTTAGGGATATCACCTACG
AGGTTGTTATGAGATACATTCctataagagttagtgaagattgcagggttatattcttagcaaaagatgcacaaagagggacaaaaaatgtttcagaacc
aaaggtgcctacttacAATACAGTGAGACTGAGACAGTTGGCTAATGAACCAACATTACCAGTCAGGTTATTATTTTCCAGTCTCctgcaagcaatttta
agaaacacaacattagatagatcaccaacccaatacatgttcctggatcacactatagtgacctcattattgtgtatacagagttttagtttcattctac
aaaaacaatttcttaacagcttggaagggataagaggaagattgcttacAGCAAAATTATGTTCTGTAATTGGTTAAGCTCTTCTGGAATTGGGCCAGAG
ATATCATTATTTGAAAGATCTctgcaaccacaaacacacaaattcagaacagtaattgagataaatgaaaaaaactctaaccaagcagatagaagttaac
ttcttacATTTCCATGATGCTTCTTAGATTTCCAAAGTCGCCTGGAACTACACCAGTTATATGATTTCTACTCAAGTTCCtataacaagaatgcaagagt
tcacaaaccattttcatctttgctagaaataaacaataaaactggaggtcttatggatgctcacATCTTGAGAAGATGCTCCAAATCACCAAGGGAAGA
AGGAATGATTCCATTTATCTTGTTGTTGGAAAGATCCcttcaaatgtcattaaatttaataaatgtgagacaaggcacatataatgatgtatacttcaca
gaaaacaagaaacacttacAATGTATCTAAGTTACCGATACGAGATAGCTCAACCGGGATTGGACCTTTGATATTGTTGCTGGACAGATTActggttgaa
ccgataaatcaaaggcaaccaaaaactcagtagttaatgacttgaaattcgaaaaaacaaacgctgtggaactatatattatataagattagagaaattc
tcagcgtcgatacttacAGGTAAGTCATACTTTCTAGCTTTTGAAATGCTCGGGGTATAGTGCCACTAAACTTGTTCCCATGAACATTTctacattacat
taatgagacacaaaatcatataaagcaccaatcttgaacattcaagtgagaagttttcagagaaagagatacatacAAGCTGTTTAGATTTGTGCAAGAG
CTCAGATGATCAGGTATAGGTCCTTCCAGATCATTGTTGGCCACATTCctatatgtacatagttcatataagcataagaacatcaaaacatgtcaagcat
aggaagaactacttacAGATCAAACAAGTCAGTAAGCTTCCCAAGCTCTGGTGGTATATGACCCGTGAGATGATTATCATTGAGTTCCctatgaatattt
tacaaatagttgataaggaaacacaacttagtagtgactgtatgcaaaaagagaagtgaagagaaggttggtcatacAGGTAATGGAGTTTTGACATGTT
TCCAAGCTCAGGTGGAATTGAACCAGTCAGCTTGTTACTGTGCAAATACctagagataaagccaaaaagataaatcactatcattaacatggtatgacat
ttgactccaaactgaaaacaaacaggtaaagaattacAATTTCTCGGTGAAAGTAAGATTTCCGAGAATCGGAGGAATAGATCCACTCAACAAGTTGCCA
CTTAGATCTctgtaaaagaaaacatataacctaagttcatgaaaatgttataagtagcttcattgattagaagaaaagtacttacAAGACTGCAAGGGCT
TGCATGAGACCAATCACTGATGGAATCTTCCCAGAGAGTTGATTGCCTTGCAATGATctaagataagaaaatgatcttctttaaaagaaaacaaaggcta
gccaaaatttaaagcaaaagattagtagaggtgagaactaacAATGTTGCAACTTGCAGGAAGCCGATGTCAAAAGGGATCTCACCAGTTAGCTGATTGT
AGGACAAGTCCctgagttcgtaataagcaaaccagttaagtagattttcttaagtaacccaaattacagtaatgatgtttagaagtagagaaagaggcac
atacAAAACCTGGAAGGCAGTGCAATTTCCTATCGTCTCAGGTATACTACCAGTCAAACTGTTGTTTCTTACGTCActgattcaatatcaaaacaaagtc
agatatttgatatttacaaaagaactcatcatactattcagatgtgcaagaagactcacAAATACCAAAGACCAGTCAGTTGACACAAATCTGGAGAAAT
GTTACCGACTAAGTTGTTTCCTCGCAACCCActgctcatcatagtacaacaattatcaccgggtaatgaagagacattgacagacttcaaaaaacattga
cacttacAGATACTGAAGAACTTCATTCCAGTAAATAAGTCTTGGTATCTCACCACTGAGTTTATTCTGTGCCAAGTCCctgcaattgaaattacacatg
tatgatcacctatatagctaaaagtagtaagacatgtaagaatcatggggaacatacAGAATTTTCAGGTTTGGAATCTGTGAAAGTGTTGAAGGGATCG
GTCCTATCAATTGGTTATTCTTCAGAATCctggacaaagtaaaaatccatattccacatggtgagtatattttagcaatgacatatgaagactagtagc
agaataactagctacttacAGCTGCTCAAGTTGCTTCAACTTCGAAATCGAAAACGGTATGTCACCACTTAATTCATTGAAGGATAAGTCTctgaataaa
aaacaacatgaacacacttcatcaaataaaacctctatgaacaaaagtttaatggtagaataacaccaatcactgttcttacAAGTTTTGCAAAGAAGAA
CAGTCACCAATCTCATCAGGGATTTGTCCAGACAAGCGATTACCTCGCAGATCActacattagaaaaaaacagaagacaaaagtcacaactcagagagtg
gaatagaattaaagtaagatactttatttcatgctacagtgaatgttgaaacagttacATTGACAAGAGACTCTTGAGATCTCCAATAGCAGGTGAGATT
TCTCCATCAAGATTCAAATCTGACAAATTActacaaaacatagaaaacacacacaaaacaaattacaacttttcccacaagaacaccaaaaaccactgt
aatagtaatctaaaggaatgaagaaacttacAGAGCAACAACATTGAAGGTGACATTTTCACAAGACACACCTCTCCAGACACAATAATCCGAAGAAGGT
GAAGTTGTCCAGTCATAAAGAACATTGTTCACATCTTTGAATGACTTCTTAATCTCCAGCAACGTTGCTCctttaattttgttttaaggatattttatta
acacacacgcacaaaaacaaaaaaaaatatttgtcagttcttgagaagattaaaaattcaacttccaagaaaacaaaataaaactcaaaccctaatttat
tcaaagcttcttttgattctatagaagacccaccaagtaaaaaaaaaccattacaaaaagcaaaaatagaaaataacaagtacccagatctcattttcac
caaatgaaacttcaaaaatcatataaaaatctcatctttaagttttttttcactaatataaaacaaaaagacgaaaatcttgaacttgaagaagcatgcat
cagtataataactgacCCTCCTCTGAAGTCACAGTAGCTACTAAGCTCAAGCAGAAGAGAAACCCAAGAAGAACAATATCTCTAAACAGAGCCATttctc
acacacagtcttaaaacgactgcgttttagatatacttaaaagctttcatgaagtaagaaacagaggatagagaaagaagaagaaagaaccgta
ttatgtaccgacattagatttttttttctttctttattatgatttctacagaaatttgtttttgttttttttacttaaccgagaaaaaatgatttttatc
gttccatcacatcaaattctggcggaaatttacagtggtttcattgagtgctcagacgacagaaatacagagtaagatagagatggagagagaaagagag
agagtctttaaagagaggagattgtacaaatgtagggacctgttgtgtaagagagagagatgaagggaagaagaattatttggggaggataaagttaaaa
accccatcatcttttatctctctgccagagttctgaagctttgcagaagagagagatagagtgagtgagagggagtgattttgtttctctctcttgaagg
gtaaaacggtcataaaatgtagaagaagaaggagaagaagaacatgttaatggcagtggtgaataaaggaggtcgctttcgttgcccaatacttacgagt
ctttgtctctctgtggcttgtttcacagatgacccatatcaataaatccatcttatgggtacttttgtaattatgtatttttatgattttagtgtaat
ttttttgtttgcttaacatgttcttcttcatcttcttacttcaattactaatttaaattcaggagtttacaaaaagtttggttcagttttctcttcttta





TTATCACCGTTCGAAGGCGAATCAACCATCTCAGTTAAATTACCTTTGGAAGTAAATGGAGAAGTAAGCCCCATGTCTTTCAAAACTTCTGAAGCTTTAA
ACTCAAAAGAGAAGTTCAACTTCGGAATTCTCAAAGCATCGACTGGTGTACGGTGAAGTGGAATATGGCTATCAAGAAATCCTGGTTCAGTGCTTATTTT
CTCAAGCAGAGCAGCTAGTCCGTCTTTGTCATTAGGAAGATAAATGTACATGGAGAAATGACGTTTATCCTCTACATAAGGTAGGCGCAAAACTTGGAAA
CCATCATAGCCTCTTAGATATTGGTCCTTATAACTCATCATGAAGGGGACTTTTACCGTGTTACCGTCAAGAAGATGAAAATCGTTATCTTTTGTTAGTT
TTGCATCAAATTTTCTACTCCATGCTGCTTTGAAGTACACTGCATTTGCTAGAATAAGTGTGCTATTACGGATCTCCTTGATAGTATCGGTACAATCGCG
TGAAAGAATCTGCTTGATGAGTCCATTTGTGTGGACGTCAGCCCATATATTTACCTCATCAATCACTTCAACAGGcttcatgaccaagcgataaaaattg
accaatcattagtgaagttggatgtcatatatacacaagaaattgaagaaataaaaaaagagaatgtgatatagccgtctccttcgtaaccacattaaac
gtcttaaataacacgaacattcacctttatatatctatgttatagggattagtctaagtttcggtcccataataattgaagcaaaatttatccctaacga
aaaaagttacCTTGGTTGCGAAGTCAACTTGACTACAAGAAGCCTTGTAGGAATTCTCCAAAAGCTCTTTAAAGGAAGGCTTCAAATAAGAAGACTTGTC
GATCCAGACACCGTGAGCCGTAGATAAGCACAAATCACTTCTCTCGGTGCCGCCGTCCGCGATCTTGGCCAAGACCGCGTTAAGGTGATCCGTAGACGGT
GACATGAGGAATGAGAGGATTTCTTCTTTGGTGACGGGATTAGAACCAGCGGCGATGAGGCTGAGCAAAACGTTGATTGACATCGGTGAAAAGACAACAT
TGGAGCCATTGGCGACATCTGTTTCAATGACTTTCTTTGCTAGTCTTGCCACGACGTTGTTTTGGTTCTCAATTGATTTTCCTAACTCCATttttttcca
agaagaggtgtacgggaagatataagaatggttctgaatatgcagatgtgatgattgatgaaatagcttgtttcaagtatactagtagctatttttttgtt
ttaaatatattttctgcacatgaattattcaatcgaaatgatttagaagtaataaagtaccgcgtgtggaaaagtcgtatcatggacttcgttggaagta
cttggcattaactataaactaggtcttgaagtttggttaatttatgccttatgggtcaatttacttggaacaactctctttttcatccttttacccataa
ggcatatgcataaaccaattgttccaaatcgattttataactctgggagttgccatttggaacaactctctttttcatcattttatatctcaatctagga