

Annual Performance Report:

89ER60865

**Identification of Genes in Anonymous DNA Sequences**

Report Period:

1 February 1991 - 31 January 1992

Christopher A. Fields, PI

Computing Research Laboratory  
New Mexico State University  
Las Cruces, NM 88003-0001

505-646-2848 (office) 505-646-6218 (fax)  
cfields@nmsu.edu

**DISCLAIMER**

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

**MASTER**

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

OK

## **Summary.**

The objective of this project is the development of practical software to automate the identification of genes in anonymous DNA sequences from the human, and other higher eukaryotic genomes. A software system for automated sequence analysis, **gm** (gene modeler) has been designed, implemented, tested, and distributed to several dozen laboratories worldwide. A significantly faster, more robust, and more flexible version of this software, **gm 2.0** has now been completed, and is being tested by operational use to analyze human cosmid sequence data. A range of efforts to further understand the features of eukaryotic gene sequences are also underway.

## 1. Overview.

This Report summarizes work on the **gm** project during the third funding period, from 1 February 1991 to 31 January 1992. It supplements the previous Technical Progress Report, dated 29 July 1991, which was submitted as part of the Continuation Proposal for funding of the project from 1 February 1992 until 31 January 1993.

A major revision of the software to **gm** version 2.0 has been completed, and is available to researchers by anonymous ftp. **gm** version 2.0 incorporates a number of changes in both the pattern recognition modules and the underlying gene-assembly algorithm that render it faster, more general, and more easily modifiable than version 1.0. **gm** 2.0 is described in a paper, "**gm**: A tool for exploratory analysis of DNA sequence data," which is included in the Appendix.

An effort has been initiated to characterize differences in oligonucleotide word usage in exons and introns across a wide range of species, with the goal of understanding differences in word usage both between species and between gene families. These differences are responsible in large part for the relatively high error rates of gene prediction methods; hence understanding the range of variation, and if possible classifying genes into families based on word usage may lead to improvements in all composition-based gene prediction methods. Our previous effort to characterize sequence variability among splice sites has been extended from *C. elegans* to both *Drosophila* and plants. Papers describing this work are also included in the Appendix.

A collaboration has been initiated with the laboratory of Dr. J. Craig Venter, National Institute of Neurological Disorders and Stroke (NINDS), NIH, to analyze DNA sequence data from two large-scale human cosmid sequencing pilot projects. Dr. Fields has been resident in Dr. Venter's laboratory since 1 July 1991, and intends to remain through 31 July 1992 as described in the Continuation Proposal. This collaboration has allowed considerable testing of **gm** 2.0 on human cosmid sequence data, comparisons of **gm** with the CRM program developed at Oak Ridge National Laboratory (Uberbacher and Mural, *Proc. Natl. Acad. Sci. USA* 88: 11261 (1991)), and valuable direct experience in the operation and analysis needs of a large-scale genome sequencing laboratory.

Dr. Soderlund was awarded a Human Genome Distinguished Postdoctoral Fellowship in June, 1991. She has been resident at Los Alamos National Laboratory since 15 July 1991, in the Theoretical Biology and Biophysics group.

## 2. Current status of **gm**.

Replacement of the limited single- and dinucleotide frequency tests used in **gm** 1.0 for candidate exon evaluation with a general oligonucleotide word frequency test makes **gm** 2.0 adaptable to analyze sequences from any eukaryote. The word frequency test employs matrices of loglikelihood ratios for each n-mer word, with  $n = 1 - 6$ , occurring in either coding exons or introns; precalculated word frequency matrices for human, *Drosophila*, *C. elegans*, and both monocot and dicot plants are included with the software distribution. Alternative tables can be specified by the user.

**gm** 2.0 includes a function for initiating exon map construction from partial 3' cDNA sequence data. This function allows partial sequences of oligo-dT primed cDNAs to be used to calculate potential upstream exons. Partial cDNA sequences are being obtained very rapidly for humans (Adams *et al.*, *Nature* 355: 632 (1992)) and other organisms; hence this feature is expected to be of considerable utility for cosmid sequence analysis. **gm** 2.0 also includes a function for masking out *Alu* or other user-specified repetitive sequences, which increases its utility

for human cosmid sequence analysis.

The gm 2.0 user interface, shown in Fig. 1, provides facilities for redisplaying previously-calculated exon maps, comparing predictions with cDNA sequence data, displaying restriction maps, STS locations, and repetitive sequences, and accessing the menu suite of tools interactively without leaving a gm session. These changes considerably enhance the interface as a display tool, and obviate the need to restart gm during exploratory runs in which parameters are altered.

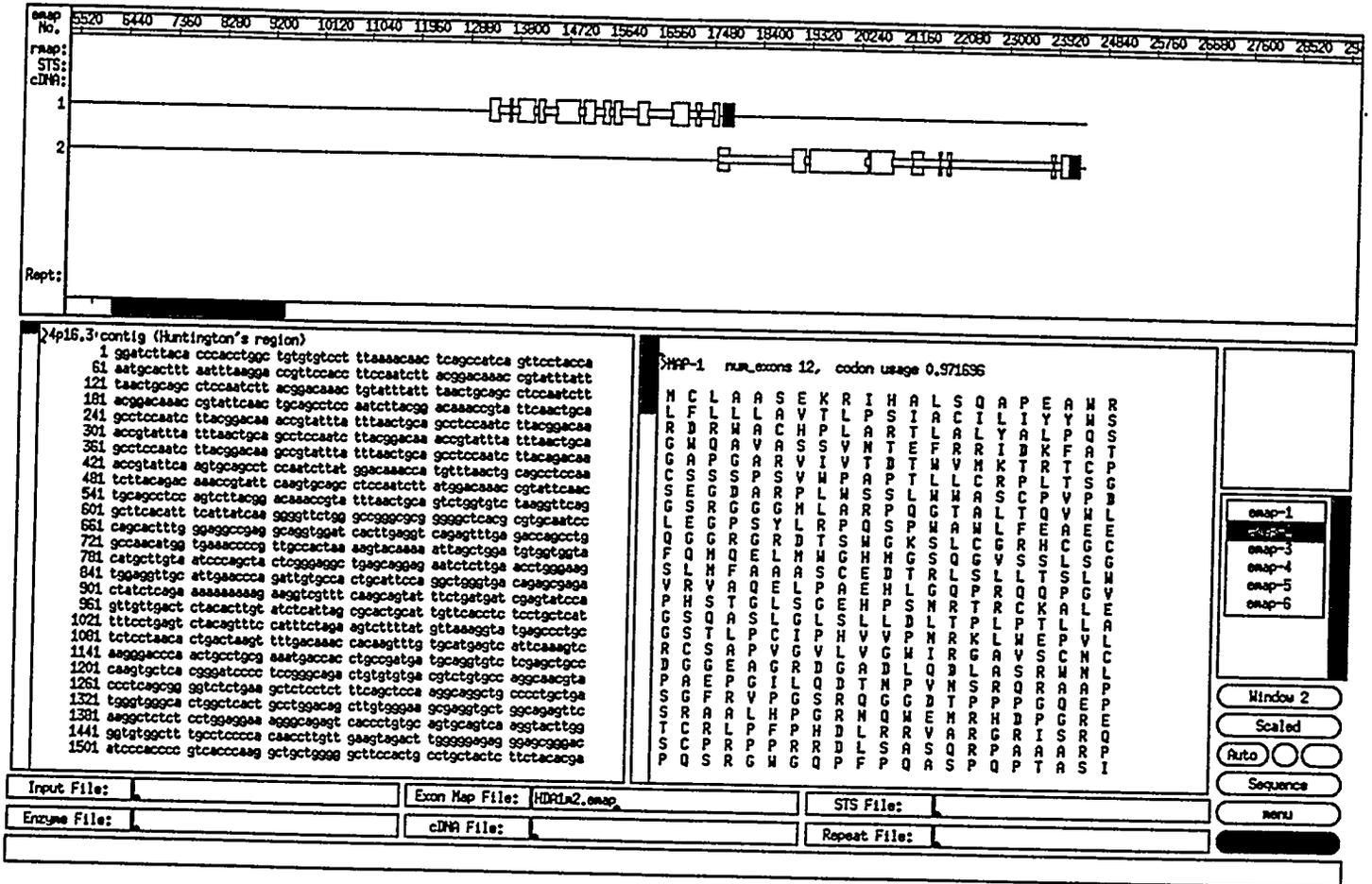


Fig. 1: Current gm graphic user interface. This interface provides four exon map displays, zoom and pan functions, interactive access to evaluation results, and a number of mapping tools. A prediction for a region of human cosmid sequence data from chromosome 4p16.3 is shown.

### 3. Distribution and user support.

The gm 2.0 system is available by anonymous ftp from haywire.nmsu.edu. Full source, executables, and standard input files are provided. Presentations describing the system have been made at five international meetings and several workshops during the last year.

#### 4. Sequence analysis.

Programs such as **gm** can only be adequately evaluated by operational use to analyze DNA sequence data. While retrospective analyses of characterized sequences can provide information on the accuracy *per se* of the program as a predictive tool, accuracy is only one component of utility in an operational setting. Speed, flexibility in different applications, and usability by the target user population are additional important components of utility. Operational use tests all of these components, and provides in addition valuable information on how the software is likely to be used by a typical large-scale sequencing laboratory.

**gm** has been used operationally since July 1991 in the Venter laboratory at NINDS, mainly by Dr. Fields, for analyzing human cosmid sequence data. The CRM coding-region prediction software and the Blast database comparison software (Altschul *et al.* *J. Mol. Biol.* 215: 403 (1990)) are also in operational use in this analysis effort. The sequence data are typically analyzed following contig assembly, but prior to completion of gap closure; the sequences may, therefore, contain errors and ambiguous bases when the initial analyses are performed. The goal of the analysis is not complete gene prediction, since predictions are not regarded by the biologists involved as sufficiently reliable to identify genes without further experimental evidence of transcription. Instead, initial predictions of exons are used to design PCR primers, which are used to amplify sequences from one or more cDNA libraries. The sequences of the amplified fragments are then compared to the genomic sequence, and predictions re-run using these new data to further define the regions containing genes. Both the predicted and experimentally determined exon sequences are simultaneously translated and compared to sequences of known proteins.

Several new human genes have been identified using this cyclic procedure. One of these, the *fosB* proto-oncogene, was initially recognized by a **gm** prediction (Martin-Gallardo *et al.* *Nature Genetics* 1: 34-39 (1992)). Other genes were initially recognized by CRM predictions, in which case **gm** was used to investigate possible splicing patterns (CRM only predicts coding regions).

None of the available gene prediction methods are 100% accurate. Fortunately, a prediction only needs to identify a single exon of a gene for the transcript to be amplified from a cDNA library by PCR. Improvements in accuracy would, however, increase the efficiency with which genes with unusual structure that encode proteins unlike any in the database can be detected. We are, therefore, continuing an effort to characterize both the structures of splice sites and the overall compositional properties of genes from a variety of organisms. A paper describing some of the work on splice sites, "Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin," is included in the Appendix.

In order to make full use of the exon and intron composition tests implemented in **gm** 2.0, we have initiated an effort to characterize n-mer word usage in a wide variety of species. Frequencies of all oligonucleotide words for  $n = 1 - 6$  have been measured for all human, *Drosophila*, *C. elegans*, monocot, and dicot sequences in GenBank for which the splice sites are unambiguously specified. The raw frequencies for each value of  $n$  have been corrected to remove the components purely due to the underlying distribution of  $(n-1)$ -mers. An example corrected frequency spectrum, for human 4mers, is shown in Fig. 2.

Comparisons of spectra like that shown in Fig. 2 show that the most frequent n-mers in exons and introns differ considerably between different organisms. Moreover, the values of  $n$  at which the greatest differences occur differ between organisms. Further work to characterize

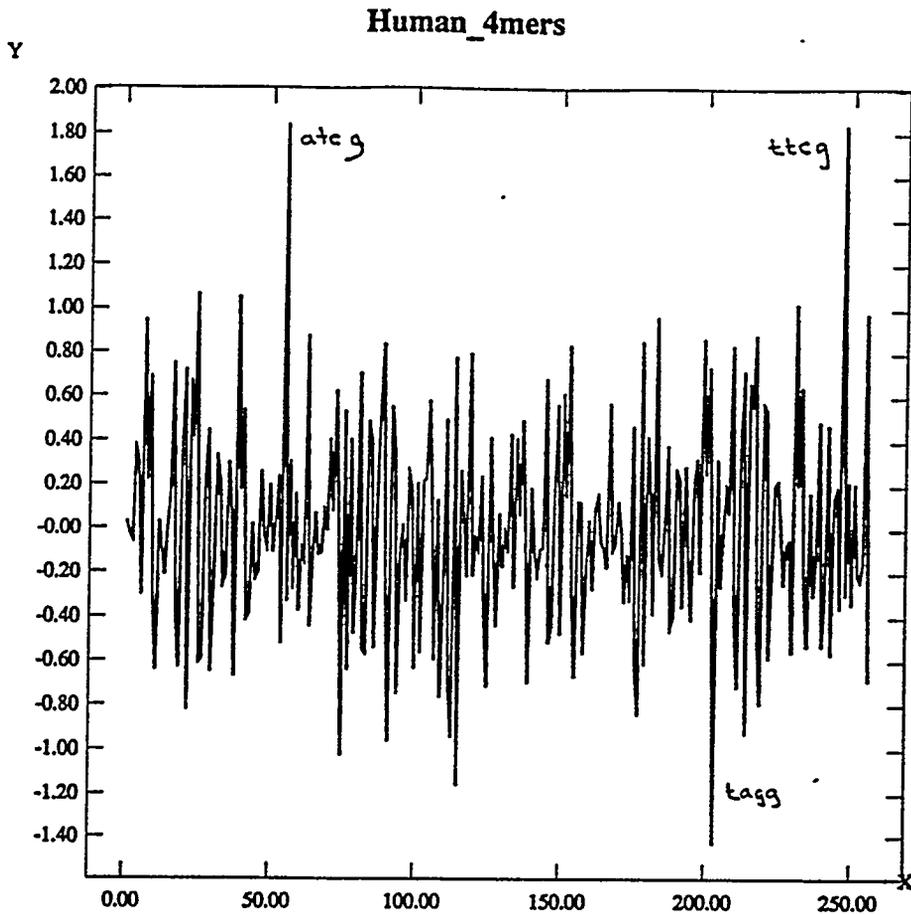


Fig. 2: Spectrum of human 4mer frequencies, corrected to account for the underlying frequencies of 3mers. The spectrum shows the loglikelihood ratio of 4mer frequencies in exons (values greater than 0.0) versus introns (values less than 0.0). Particularly asymmetrically distributed 4mers are labelled.

these differences quantitatively is underway.

### 5. Related projects.

A number of sequences similar to THE interspersed repeats were identified in the course of screening human genomic STS sequences generated at Los Alamos National Laboratory. Analysis of these sequences showed that the THE and *MstII* interspersed repeat families are very similar, and hence very likely to be subfamilies of a single family. A paper describing this work is included in the Appendix.

A set of Unix shell scripts and associated programs were developed in collaboration with intended users at LANL to analyze genomic STS sequences and select PCR primers. This work was supported separately by LANL.

Dr. Fields was appointed Informatics Coordinator for the DOE cDNA Program in May 1991. Fields is involved in informatics support and analysis for the cDNA sequencing project as part of his visit in the laboratory of Dr. J. Craig Venter, NINDS, NIH. This work is partially supported by the DOE Human Genome Program through an interagency agreement with

NINDS.

## 6. Personnel and facilities.

The project personnel and their principal activities during the project period are as follows:

*C. A. Soderlund, Computer Specialist:* Program specification and algorithm development, software design and implementation, management of software effort and releases, software and user support. Currently a Human Genome Distinguished Postdoctoral Fellow, T-10, LANL.

*P. Shanmugam, Research Assistant:* Graphic user interface development, graphics programming, information-theoretic and power-spectrum methods development. Completed MS degree in May, 1991; appointment ended June, 1991.

*O. White, Research Assistant:* Sequence analysis and methods development.

*Ted Slator, Research Assistant:* General programming.

*C. A. Fields, Project Manager:* Program specification and testing, analysis methods evaluation, sequence analysis applications, user support, PI. Currently visiting NINDS, NIH.

The Sun-4/60 workstation purchased in October 1989 remains the principal development and testing machine for the project.

## 7. Publications and presentations.

The following papers describing supported by this project have been published or accepted for publication:

- C. A. Soderlund, P. Shanmugam, O. White, and C. Fields (1992) gm: A tool for exploratory analysis of DNA sequence data. *Proc. Hawaii Int. Conf. on Systems Science*. IEEE Computer Society Press. pp. 653-662.
- C. Fields, D. Grady, and R. K. Moyzis (1992) The human THE-LTR(O) and *MstII* interspersed repeats are subfamilies of a single widely dispersed highly variable repeat family. *Genomics* 13: in press.
- O. White, C. Soderlund, P. Shanmugam, and C. Fields (1992) Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Molecular Biology* (accepted).

A paper reporting information-content analysis of *Drosophila* introns is in progress.

Project personnel have attended and presented talks or demonstrations at the following conferences and workshops:

- DOE Genome Program Contractor - Grantee Meeting, Santa Fe, 17-20 February 1991.
- Mathematics and Molecular Biology, Santa Fe, 24-29 March 1991.
- Genome Mapping and Sequencing, Cold Spring Harbor Laboratory, 8-12 May 1991.
- Recognition of Genes and Other Elements of Genomic Structure, Aspen Center for Physics, 20 May - 7 June 1991.
- International *C. elegans* Meeting, Madison, 1-5 June 1991.
- Human Gene Mapping Workshop, London, 18-22 August 1991.

- Genome Sequencing Conference III, Hilton Head Island, 22-25 September 1991.
- Workshop on Identification of Transcribed Sequences in the Human Genome, NIH, 4-5 October 1991.
- Second W. M. Keck Symposium on Computational Biology, Houston, 14 November 1991.
- Hawaii International Conference on System Science, 7-10 January, 1992.

Abstracts of papers presented at these meetings, where available, are included in the Appendix.

## Appendices.

### 1. Publications.

- **gm**: A tool for exploratory analysis of DNA sequence data. *Proc. Hawaii Int. Conf. on Systems Science*. IEEE Computer Society Press. pp. 653-662.
- The human THE-LTR(O) and *MstII* interspersed repeats are subfamilies of a single widely dispersed highly variable repeat family. *Genomics* 13: in press.
- Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Molecular Biology* (accepted).
- Splicing signals in *Drosophila*: Intron size, information content, and consensus sequences. *Nucleic Acids Research* (in preparation).

### 2. Abstracts.

- Automated prediction of priming sites for STS sequences. DOE Contractor-Grantee Meeting.
- Integration of automated sequence analysis into mapping and sequencing projects. DOE Contractor-Grantee Meeting.
- Contig assembly program (CA). DOE Contractor-Grantee Meeting.
- New features in **gm**, version 2.0. DOE Contractor-Grantee Meeting.
- Performance of **gm** version 2.0. Cold Spring Harbor.
- Automated prediction of priming sites for STS sequences. Cold Spring Harbor.
- New functions in **gm** version 2.0. Cold Spring Harbor.
- Fine-tuning **gm** version 2.0 for worm sequence analysis. *C. elegans* Meeting.
- The expressed sequence tag database: A resource for genomics and developmental biology. Genome Sequencing III.
- Gene identification using **gm**. Genome Sequencing III.
- Informatics support for large-scale sequencing. Keck Foundation Symposium.

### 3. **gm** 2.0 Distribution.

- README file.

Proceedings of the  
**Twenty-Fifth Hawaii International  
Conference on System Sciences**

**Volume I:**

**Architecture**

*Edited by Veljko Milutinovic*

**Emerging Technologies**

*Edited by Bruce D. Shriver*

**Sponsored by the University of Hawaii**

**In cooperation with ACM, the IEEE Computer Society, and  
the Pacific Research Institute for Information Systems and Management**

1951-1991



**IEEE Computer Society Press  
Los Alamitos, California**

**Washington • Brussels • Tokyo**

---

# gm: A Tool for Exploratory Analysis of DNA Sequence Data

C.Soderlund, P.Shanmugam, O.White, C.Fields  
Computing Research Laboratory  
Box 30001/3CRL  
New Mexico State University  
Las Cruces, NM 88003

## Abstract

*The gm (gene mapper) system provides a tool which automates the analysis and identification of candidate genes within an anonymous DNA sequence. Only a small, biased set of all genes are known; from this sampling it is not possible to know the composition of all genes. Consequently, it is not currently possible to develop software to detect all genes. gm is capable of detecting large regions of many genes and provides an exploratory tool for looking at anonymous DNA sequences. The first release of gm has been distributed to several dozen user sites, who have provided feedback which is being integrated into gm v.2. This paper discusses the problems encountered in gene identification software, the performance of gm v1.0, and the new features in gm v.2 that will enhance its use as an exploratory tool.*

## 1.0 Introduction

In consideration to the computer scientist who is not familiar with genetics, the structure of a gene will be reviewed in the first subsection. The second subsection describes various methods for detecting genes in DNA sequence. Section 2 describes the performance of gm v1.0, and section 3 describes the new features in gm v2.0 and the v2.0 algorithm.

### 1.1 Structure of a Gene

DNA sequence is a double stranded linear array of bases; a base can be A, C, G or T. Genes are located along both strands of the sequence. As is shown in Figure 1, a gene is composed of two types of entities: alternating protein coding exons and non-coding introns. A gene must have at least one exon, and has zero or more introns. The exons and introns are separated by functional sites called the 5' splice sites (5'SS) and the 3' splice sites (3'SS). The coding region of the first exon begins with the sequence ATG, and downstream from the coding region of the last exon is the sequence AATAAA.

As shown in Figure 1, a gene is expressed as follows: (a) the gene is first transcribed (copied) into RNA, (b) the introns are spliced out of the gene, resulting in the concatenation of the exons, and (c) the exons are translated into a protein. The translation mechanism converts each set of three bases, called a codon, into an amino acid; the string of amino acids creates the protein. The bases are translated into amino acids until a "stop codon" is encountered; the possible stop codons are TGA, TAA, or TAG. As shown in Figure 2, only an "in-frame" stop codon will stop translation since it is the only one that will be translated. An open reading frame (orf) is the stretch of bases from one stop codon in a given frame to the next stop codon in the same frame.

### 1.2 Methods of Gene Identification

Typically, the biologist looks for possible genes in a sequence by searching for the following: (i) patterns in the DNA that resemble known functional sites, (ii) base composition of the DNA between the functional sites that resembles that of known exons and introns, and (iii) long open reading frames which indicate the potential presence of an exon. Computer methods for identifying gene entities are reviewed by Stormo [1] and Doolittle [2]. The following discussion will review some of these techniques.

A typical way to find functional sites is with consensus matrices, as defined by Staden [3] and Stormo, Schneider, Gold and Ehrenfeucht [4]. They are built as follows: All 5'SS have a required GT pair and all 3'SS have a required AG pair. All the known 5'SS for an organism are aligned on their GT, and the consensus for the surrounding bases is calculated. For instance, the 5'SS consensus matrix for the nematode worm *C.elegans* is shown in Figure 3; the consensus was determined for 3 bases upstream from the GT and 8 bases downstream from the GT. The four rows show the frequencies of A, C, G, and T for each position. For example, the first column in the matrix shows the third base to the left of the GT is

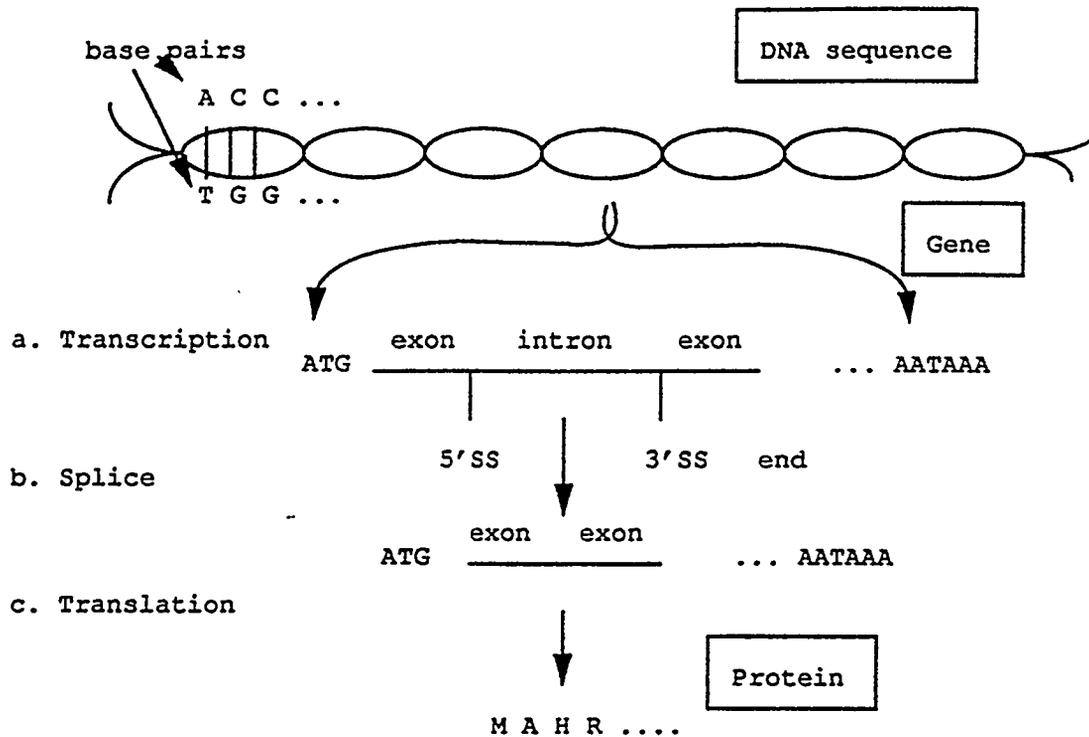


Figure 1. A DNA sequence is a double-stranded linear array of the bases. The gene in this example has two exons and one intron. A gene is expressed (converted to a protein) by transcription, splicing, and translation.

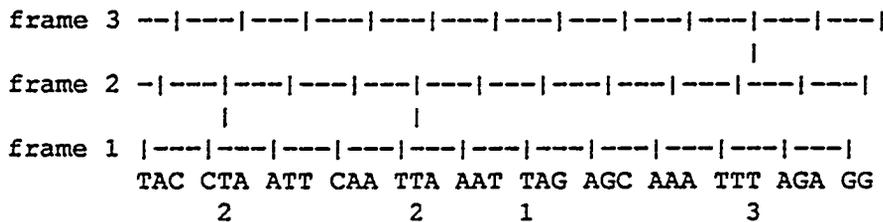


Figure 2. A sequence of DNA may be translated in one of three frames. In this example, translation occurs in frame 1, so only the TAG in frame 1 functions as a stop codon.

A	0.48	0.62	0.12	0.00	0.00	0.65	0.74	0.15	0.19	0.26	0.29
C	0.22	0.17	0.07	0.00	0.00	0.01	0.04	0.04	0.09	0.07	0.10
G	0.19	0.06	0.68	1.00	0.00	0.20	0.06	0.73	0.05	0.06	0.04
T	0.10	0.14	0.13	0.00	1.00	0.13	0.15	0.08	0.68	0.60	0.56

Figure 3. The consensus matrix for the *C.elegans* 5'SS uses 13 bases; this number was determined by evaluating the information content around the required GT. The GT is represented in the fourth and fifth columns.

0.48% A, 0.22% C, 0.19% G, and 0.10% T. This matrix is slid along a DNA sequence, and whenever the 13-base window scores above a cutoff, it may be considered as a potential 5'SS. The highest possible score for the example consensus matrix is 7.74 for the sequence AAGGTAAGTTT. The cutoff must be set such that the number of false-negatives (real splice sites that fail the consensus matrix test) and false-positives (regions that are incorrectly identified as splice sites) is low. We have found that the number of false-negative and false-positive splice sites is decreased by using the consensus matrix along with looking at a window on both sides of a potential splice site for asymmetric patterns in base composition (paper in preparation).

Composition of exons has been done in a variety of ways. A standard technique is by calculating the codon usage for all the known genes: there are  $4^3 = 64$  possible codons, and some codons are more likely to occur in exons. A user may check the codon usage of a sequence, and if the codons are similar to the standard codon usage, the interval may be marked as a potential exon. Shepard [5] has shown that identical bases are most often in identical codon positions in coding regions. Csank, Taylor, and Martindale [6] have shown that invertebrate exons tend to have high C+G content whereas introns tend to have high A+T content; this is not the case in vertebrates. Bougueleret, Tekaiia, Sauvaget, and Claverie [7] have observed that in vertebrates some 6-tuples are preferred in exons while a different set of 6-tuples are preferred in introns. We have observed that in invertebrates, exons can be distinguished from introns by looking at 1-tuples through 6-tuples, whereas it is necessary to look for tuples over 3 bases long in vertebrates in order to distinguish exons from introns (paper in preparation). A problem with all these techniques is that they do not work well for exons that are smaller than 200 base pairs; this is a significant problem considering that human exons tend to be smaller than 400 base pairs. A second unresolved problem, which none of the above techniques consider, is that a region of the DNA sequence may be used as an intron during one stage of a cell's life-time and may be translated as a gene in another stage of a cell's life-time; that is, a region of DNA may function as both a coding and non-coding region [8].

Neural networks have been used for composition analysis of genes. Farber, Lapedes, and Sirotkin [9] can identify coding and non-coding regions with the neural networks in conjunction with information theory. Brunak, Engelbrecht, and Knudsen [10] identify splice sites by a joint prediction scheme which uses the prediction of coding and non-coding regions (by standard composition methods described above) to regulate the cutoff value for

neural network splice site assignment. Uberbacher and Mural [11] use various composition methods to identify exons and introns, and input the various scores into a neural network to distinguish coding from non-coding regions.

The above techniques only provide ways to identify various entities of a gene; the biologist must assemble the entities from the output of the composition and pattern matching routines. Not only is this tedious work, but they tend to use the first syntactically correct solution they find. The gm system described in this paper aids the user by executing all the composition and pattern matching routines, assembling the results, and presenting the user with all possible exon maps (emaps); that is, all the syntactically correct combinations of exons and introns. Complementary research is being done by Knight and Myers [12], who are researching and developing a program which uses grammars to provide a flexible way to specify how the entities may be combined, and by Guigo, Knudsen, Drake, and Smith [13], who have recently developed software to aid the identification of genes in human DNA.

The problems encountered -- for both the biologist and gene identification software -- are: (i) very few genes are known so the sample set is small and probably biased, (ii) the composition of genes varies across gene families, and (iii) the composition of genes varies across organisms. In other words, we cannot predict all genes based on the knowledge we currently have of gene composition. The best we can do, at this point, is create software that allows the user to perform the following: (i) analyze the DNA sequence with known compositional methods, (ii) test hypothesis, and (iii) easily alter the compositional routines. Additionally, software needs to be continually updated and released to incorporate recent biological discoveries. These four points define the software methodology we adhere to for gm. gm v2.0 allows the user to tune the software with current knowledge about a specific organism, and to test hypothesis about gene structure, as follows:

- Interactive graphics allow the user to adjust parameters, run gm, and re-adjust parameters based on the output. The graphics allows the user to view each exon map along with its compositional measures. Relevant physical mapping data can also be displayed.
- A cDNA partially defines the coordinates of a gene. The user can input cDNA coordinates and gm will attempt to extend the cDNA; the results can be used to design laboratory tests to further probe for the gene.
- The user inputs organism specific consensus matrices for the 5' splice site, 3' splice site, and other functional sites.

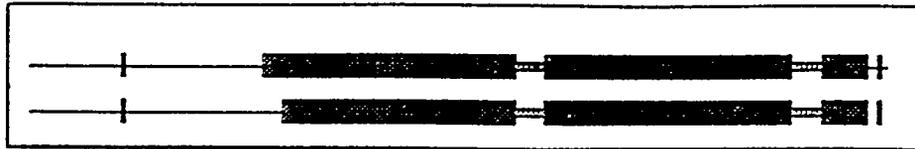


Figure 4. The top exon map is the correct *col-8* gene. The second exon map is a gm predication; it missed the 5' splice site in the first exon and detected an alternative splice site downstream.

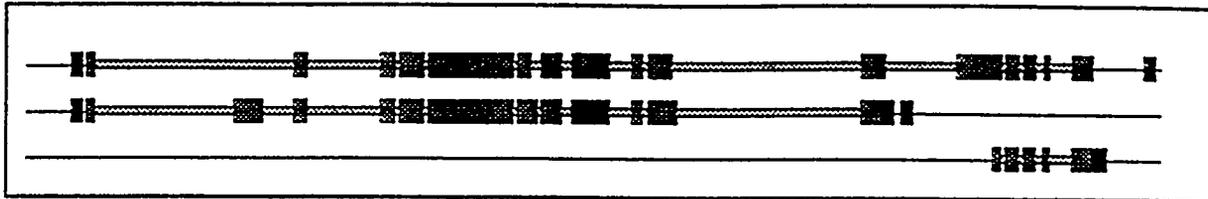


Figure 5. The top exon map is the correct *deb-1* gene. The second two exon maps are gm predictions; the two maps combined cover the majority of the correct map.

- The user inputs organism specific tuple preference matrices for exons and introns.
- The composition functions are separate modules that may be replaced with alternative methods.

## 2.0 Performance of gm v1.0

In the spring of 1990, gm Version 1.0 alpha was released [14]. It has been installed at approximately 50 laboratories around the world. We have demonstrated gm v1.0 at multiple conferences and visited multiple user sites. The response to gm has been very favorable; basically, users feel they really need a tool such as this one. On the other hand, gm has not received wide use for the following reasons: (i) Most laboratories are working on completing the physical map and do not anticipate doing large-scale sequencing for a while. (ii) gm v1.0 is tuned for invertebrates, specifically *C.elegans*. Many of our potential users are working on mammalian sequences. (iii) User feedback on gm v1.0 alpha shows some difficulty in using gm; the problems and their solutions will be covered in the section on gm v2.0.

### 2.1 Evaluation

We have tested gm v1.0 on several known genes from *C. elegans*. gm generally predicts short genes with short introns correctly. It has also correctly predicted long genes such as *myo-2* and *unc-15*. When gm does not get "the" correct solution, the incorrect solution is one of three types: (i) gm predicts most of the correct exons, with one or a few incorrect -- but close -- splice sites used. An example is shown in Figure 4, the top exon map is the

correct one, the second exon map is predicted by gm; it is correct except for the first exon. (ii) The correct gene is a composition of two or more predicted genes. An example is shown in Figure 5, the top exon map is the correct one, the second and third exon maps define the majority of the correct exon map. (iii) gm identifies little or none of the correct gene. We feel that the first two situations are acceptable; these predictions provide enough information for the user to design the appropriate experiment to verify the existence of a gene in the region shown in the prediction. We are currently investigating why some genes are difficult to identify (i.e. why some genes have uncharacteristic splice sites, exon composition, or intron composition).

We have used gm on new sequences obtained by collaborators. The 54 kb *unc-22* cosmid of *C. elegans* [15] was used for an initial large-scale test of gm v1.0. The locations of several new *unc-22* exons, and of five additional genes, were predicted using gm; this is shown in Figure 6. The existence of several of the predicted exons, and two of the five predicted genes has been independently confirmed by G. Benian and colleagues at Emory University by cDNA sequencing; the existence of an additional predicted gene has been confirmed by D. Baillie and colleagues at Simon Fraser.

### 2.2 Execution time

There is often a trade-off in computational genetic algorithms between fast execution time and accuracy. It is generally desirable to put accuracy first; that is, the user does not care if the program runs all night -- as long as the answer is as accurate as possible. The gm algorithm was

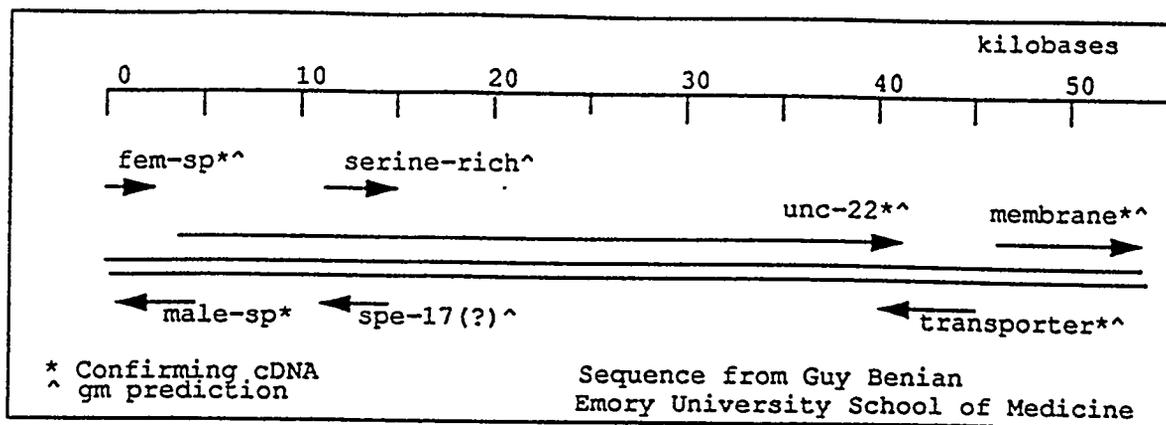


Figure 6. gm predicted the locations of several new *unc-22* exons and five additional genes. The "\*" indicate that the gene or some exons within the gene, were predicted by gm. The "^" indicates that the gene has been confirmed by cDNA.

written for accuracy, but was then tuned for speed. Table 1 (extracted from [16]) shows that gm can detect complicated genes in less than a minute.

We feel that software can be used more effectively as an exploratory tool if the user gets fast feedback. Unfortunately, we can not guarantee fast execution times for the following reasons: (i) The sequence may be inherently difficult; for example, it may be extremely long. (ii) The user may set the cutoff values very low which could create many entities, and cause an explosion in the number of exon maps generated. To prevent this, we advise the user to start with very high cutoffs and lower them as appropriate. Additionally, there is an option to

show only the longest non-inclusive exon maps. When using this option, the user is only shown one combination of splicing for a region. This is sufficient for determining if the region is interesting; if it is, the user may run gm with the option to look at all maps for that region.

### 3.0 gm Version 2

The following points describe changes made in gm v2.0 as a result of feedback from the users.

- **Ranked exon maps.** gm v2.0 ranks the exon maps according to length of coding region. Using the translation file (amino acid sequence corresponding to each

Table 1. Results of running gm on the sequences of five *C.elegans* genes.

Sequence	Length (kb)	Exons Actual	Number of Exon Maps	Exons Predicted	Protein Predicted	Execution Time
<i>unc-15</i>	12.1	10	4	10	100%	12 sec
<i>unc-54</i>	9.0	9	16	7	96%	10 sec
<i>myo-1</i>	12.1	9	3	4	84%	7 sec
<i>myo-2</i>	10.8	12	32	11	98%	24 sec
<i>myo-3</i>	11.6	7	2	6	100%	8 sec

"Exons Actual" is the number of exons in the known gene. "Number of Exon Maps" is the total number of candidate genes predicted. "Exons Predicted" is the number of exons predicted correctly, up to small insertions in the exon map that best predicts the structure of the known gene (in no case are extra exons predicted). "Protein Predicted" is the percentage of the known amino-acid sequence predicted correctly by the best exon map. Execution times are for a Sun 4/60.

exon map), the user can run the amino acid sequence against the PIR [17] or a translation of Genbank [18] using a program such as blast [19] to see if it resembles any existing sequence; a hit against a known protein may indicate an exon map of interest.

- **Extend exon maps from known cDNAs.** A cDNA gives the coordinates of the right end of a gene. The user often has a cDNA for a sequence, and wants to know the rest of the gene. We have implemented an option which allows the user to input a file with the cDNA coordinates and gm produces all candidate exon maps which extend the cDNA.
- **Vertebrates.** In gm v1.0, the testing of splice sites was based on the organism specific consensus matrix; therefore, splice site recognition could be tuned for each organism. But the intron and exon detection routines worked only for invertebrates. As is described in the next section, a flexible organism specific test has been implemented in gm v2.0 for detecting exons and introns for either invertebrates or vertebrates.

We have noticed during site visits that users often have trouble adjusting the input parameters. In order to alleviate the problem in gm v2.0, we have provided an interactive graphical method for changing parameters and displaying the results. We have also added a graphical display of physical map data so that the user can relate candidate exon maps to known physical maps. These features will be covered in the discussion on graphics.

As was previously mentioned, the user has the option of viewing (i) all the predicted exon maps or (ii) only the longest non-inclusive exon maps. In the first version of gm, the inclusive maps were deleted late in processing; this option was mainly to limit the amount of output. In the second version of gm, a greedy algorithm has been implemented for determining the longest non-inclusive maps. The greedy algorithm only extends maps that have the longest coding region; this causes the filtering of exon maps to occur early in the processing which greatly reduces execution time. For example, gm generates 200 exon maps and takes three minutes to run on the *C.elegan* sequence *deb-1*; it takes 41 seconds to generate two exon maps when run with the greedy algorithm.

### 3.1 Algorithm

Both version 1 and version 2 share the following structure: they have a set of pattern matching routines and a main engine that assembles the exon maps based on the analysis of the pattern matching routines. Both the pattern matching routines and the main engine have been changed in gm v2.0 for better functionality, speed, and ease of

extensibility. This subsection describes the version 2 algorithm.

The gm software contains three executables: (i) gm executes the gene identification software on a text terminal, (ii) gmwin executes gm and displays the results on a graphic terminal, and (iii) menu provides options to run each compositional function stand-alone, and assorted other useful functions.

gm takes as input a file of parameters, this is shown in Figure 7. The file may be created and altered using a custom menu-driven editor included in the menu program.

```

Sequence
  file name: daf-1.dna
  starting position: 1
  finishing position (enter zero to compute length): 0
Output Files
  exon map file: daf-1.emap
  translation file: daf-1.trans
cDNA Input (Optional)
  cDNA input (on/off): on
  cDNA file name: daf-1.coord
  (0) cDNA maps only      (1) All maps: 1
5' Splice Site Identification
  consensus matrix file name: worm5-13.log
  cutoff value: 2.000
Branch Site Identification (optional)
  branch site dinucleotide (on/off): off
  branch site consensus (on/off): off
3' Splice Site Identification
  consensus matrix file name: worm3-13.log
  cutoff value: 1.500
Splice Site Composition (optional)
  Splice Site Composition (on/off): on
  5' cutoff: 0.500
  3' cutoff: 0.500
Initiation Context (ATG) Identification
  consensus matrix file name: wormatg-9.log
  cutoff value: -1.000
PolyA Signal (AATAAA) Identification
  consensus matrix file name: aataaa
  cutoff value: 5.000
  maximum stop - AATAAA separation: 350
Promoter Signal Identification
  consensus matrix file name: tataa
  cutoff value: 5.000
  maximum promoter - ATG separation: 250
Intron Evaluation
  mask matrix: intron-mask
  cutoff: 2.000
  minimum length: 30
Repeat (optional)
  test exons against repeat coordinates (on/off): off
Exon Evaluation
  mask matrix: exon-mask
  cutoff: 2.000
  codon asymmetry file name: wormcodon-10
  codon asymmetry cutoff: 2
Sliding Window
  size: 50
  overlap: 0
  cutoff: 0.400
Protein Coding Capacity
  minimum coding length (in bases): 300
Analysis Options
  (0) Complete maps only
  (1) Complete maps and partial maps: 0
  (0) All maps
  (1) Longest noninclusive maps only: 1
  (0) Exon-biased evaluation
  (1) Brute force intron evaluation: 0
Trace Option
  (0) No trace      (1) Trace: 0

```

Figure 7. Example gm input file. These files are created and edited using a menu-driven editor.

```

--> MAP-2 size 5 total-length 1318 coding-length 838
--> TATAA Box
--> BASE 336 TATAA 5.000
--> EXON 395 ATG 2.500
--> BASE 556 -0.130
--> INTRON 5.863
--> BASE 557 5-SS 5.640
--> BASE 658 3-SS 3.120
--> EXON 2.637
--> BASE 659
--> BASE 1336 END
--> AATAAA
--> BASE 1654 AATAAA 6.000

```

Figure 8. An example of gm textual output. These files may also be displayed with the gmwin graphical interface.

gm outputs zero or more candidate exon maps (emaps). These emaps are written to a textual file, and may also be displayed with graphics. A sample of the textual output is shown in Figure 8; the values on the left of the various entities (exon, introns, 5'SS, 3'SS, ATG, and AATAAA) indicates how each entity scored in its respective test.

The flowchart for gm v2.0 is shown in Figure 9. The composition routines are shown on the left side of the figure; they return the locations of entities along the sequence. The compositional functions are as follows.

**Orf.** This routine returns the interval of each open reading frame, along with the reading frame number.

**Window.** This routine checks for intervals which look like coding regions. It uses the same coding region test used for exons, but with a different user supplied cutoff (it should be lower than the one supplied for the exon test). Within the intervals, this routine calls the splice site routines, and returns the list of splice sites to the main engine.

**Splice site test.** This routine takes as input the coding region intervals identified by the window routine, and the following user supplied parameters: (i) the 5'SS consensus matrix and a cutoff value, (ii) the 3'SS consensus matrix and a cutoff value, and (iii) a flag to indicate whether to look at the composition on either side of a potential splice site and a compositional cutoff value. Points in the interval that match one of the consensus matrices above the associated cutoff are marked as potential splice sites. If the composition flag is off, all splice sites are returned. If the composition flag is on, the composition of the flanking regions of each potential splice site are evaluated, and only the ones with composition scores above the cutoff will be returned.

**Intron test.** This routine takes as input an interval and the following user supplied parameters: a n-tuple preference matrix, where n may be from 1 to 6, and a cutoff. A n-tuple matrix is calculated for the interval, where n is the same as the input matrix. The score is the dot product of the input matrix and the interval matrix. If

the score is above the user-supplied cutoff, the routine will return success, otherwise it returns failure.

The n-tuple preference matrix represents the preferred tuples in introns for the specific organism. We have created preference matrices for tuples of size 1 through 6 for *C.elegans* and humans. The matrices were generated as follows: (i) all the complete Genbank genes for the organism were extracted, (ii) the frequency of each tuple was calculated for introns and for exons, and (iii) the value for the *i*th element of the n-tuple matrix was calculated by

$$\log(\text{freq\_intron}_i^n / \text{freq\_exon}_i^n), \quad (\text{EQ 1})$$

where  $\text{freq\_intron}_i^n$  is the frequency of the *i*th tuple of size n in introns and  $\text{freq\_exon}_i^n$  is the frequency of the *i*th tuple of size n in exons. An exon preference matrix is also generated, where  $\text{freq\_exon}_i^n$  is divided by  $\text{freq\_intron}_i^n$ .

**Exon Test.** This routine is exactly like the intron test except the input matrix is the exon preference matrix.

**Codon Usage.** This routine takes as input an exon map, and the following user supplied parameters: a codon usage table and a cutoff. It checks the codon usage for the concatenation of all exons in an exon map.

The left side of Figure 9 shows the main engine which builds the exon maps based on the analysis of the composition routines. The main loop is the traversal of the 5' splice sites in ascending order, so maps are constructed from left to right; that is, at any point in map building, all maps are complete to the left of the current 5'SS and no maps extend beyond the current 5'SS - 3'SS pair. Frame consistency is maintained as entities are added to each map. Codon usage is checked when a map is complete; this is so the usage may be checked as if the introns had been spliced and the exons were concatenated together. The final set of exon maps are sorted according to the longest protein coding regions (i.e. the sum of the exons).

The algorithm for the generating only the longest non-inclusive maps is the same as is shown in Figure 9, except: (i) when a new intron is found, a new exon map is only created if the intron can not be appended to an existing map, and (ii) if the intron can be appended to one or more existing maps, it is only appended to the map which has the longest coding region, and (iii) during the final sort, inclusive maps are deleted.

### 3.2 Graphics

Figure 10 shows the gmwin graphical display. The display is initially empty. The user can either display a previously calculated exon map file by entering the name in the "Exon Map File" slot, or generate a new exon file by

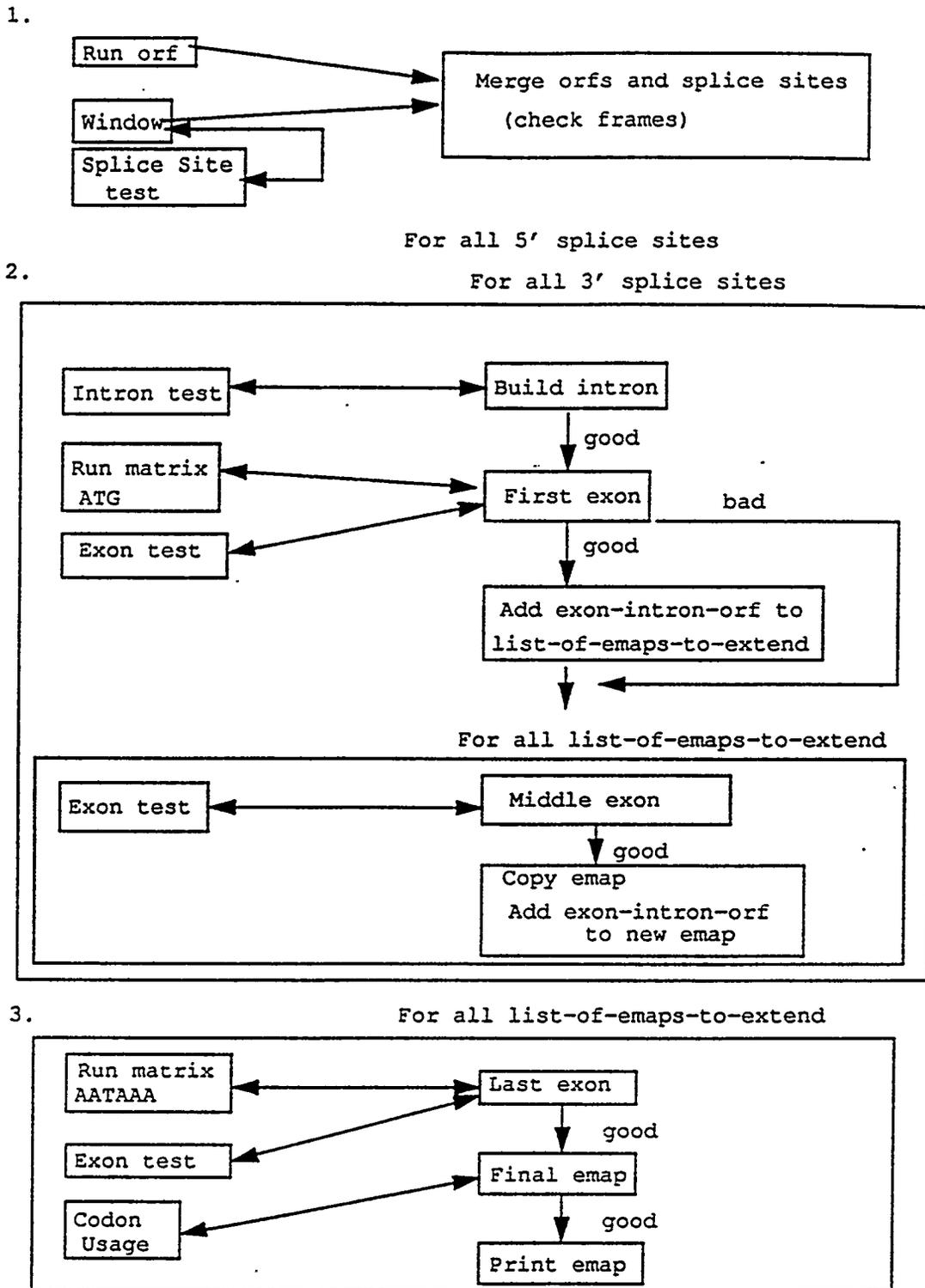


Figure 9. gm v2.0 flowchart: the left side of the diagram shows the pattern matching routines, the right side of the diagram shows the main engine that assembles the exon maps.

entering the input file name (i.e. the file containing the list of parameters) in the "Input File" slot. In either case, the DNA sequence is displayed on the left and the amino acid sequence is displayed on the right. The user may display an exon map by mousing the appropriate button on the right. Up to four exon maps may be displayed at a time

When a user clicks on an entity in an exon map, the DNA sequence for the entity will be highlighted and the coordinates and scores for the entity will be displayed in the middle box on the right. In the example, the second exon of the first map is highlighted. The user may toggle between highlighting the DNA sequence and the amino acid sequence by mousing the "Sequence" button.

When the user enters an input file, the gm algorithm is executed. The number of occurrences of the each entity is printed in the message bar on the bottom of the screen. If the solution set is unsatisfactory, the information in the

message bar will aid the user on what parameters need changing in order to approach a better solution. For instance, if no exon maps are generated and the number of 5'SS is zero, the 5'SS cutoff is too high and should be lowered. The user can mouse the "menu" button, change the appropriate parameter, write the file, and run the input file again. This scenario may be executed any number of times without exiting the graphics. This design allows the user to easily explore the consequence of using different parameter settings.

The user may view how a solution may integrate with the physical map by displaying: (a) a restriction map for an enzyme, (b) the exons of a cDNA, (b) the locations of STSs, and (d) the location of repeats. Examples of a restriction map and cDNA coordinates are displayed in Figure 10.

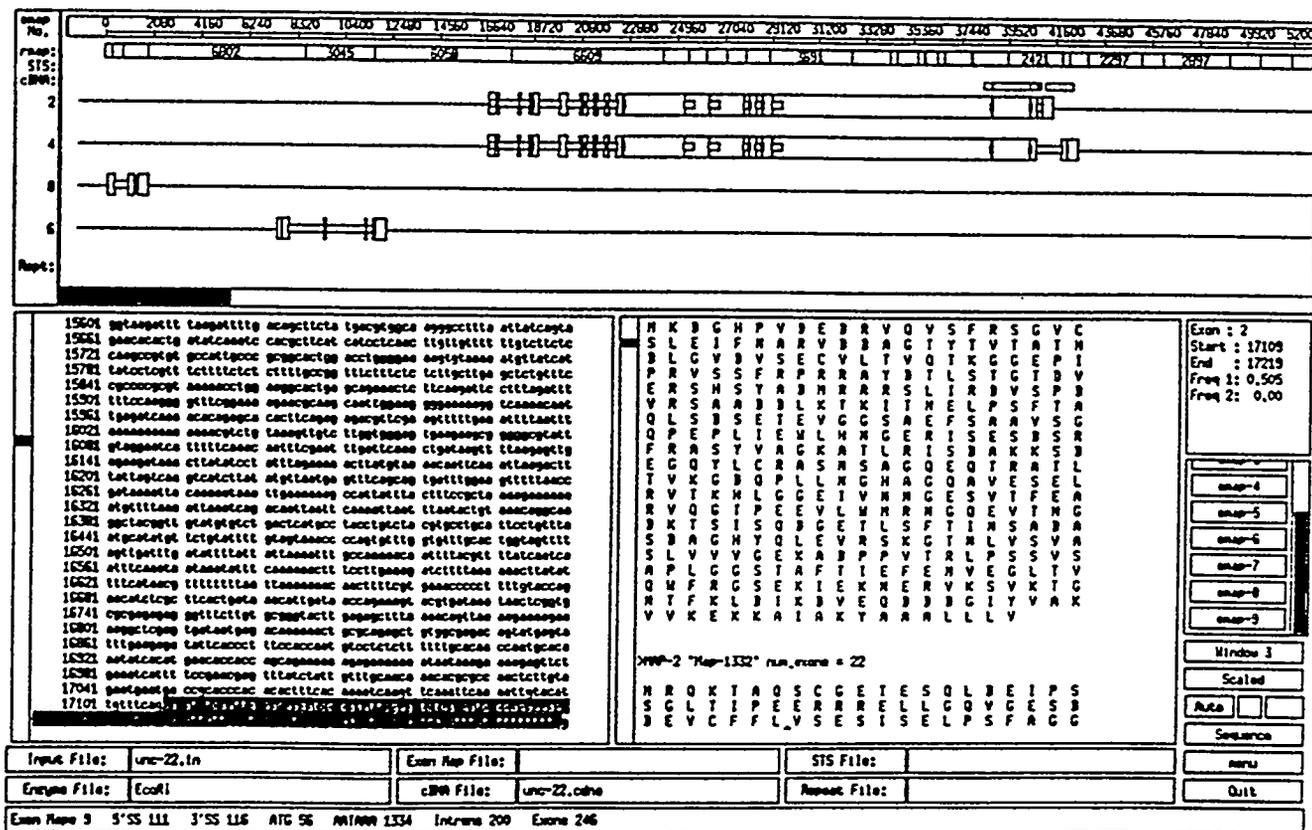


Figure 10. The gmwin v2.0 graphical interface.

## 4.0 Discussion

**Testing.** It was discussed at the Aspen Workshop on Recognizing Genes that there is a need for test data sets for all genomic computational problems; this would facilitate testing and comparing gene identification systems. The only other system for detecting genes -- which is complete enough to have results -- is the one developed by Guigo, Knudsen, Drake, and Smith [13]; their system has only recently produced results, consequently, we have not yet compared results.

**Release 2.** The new graphics has been demonstrated at the DOE Human Genome Workshop at Santa Fe, the Genome Mapping and Sequencing Meeting at Cold Spring Harbor, and the Aspen Workshop on Recognizing Genes. It received favorable comments at all three meetings. Initial testing has shown that the new technique for recognizing splice sites has allowed gm to generate the correct exon maps for genes that could not be detected in gm v1.0. The success of analyzing human exons successfully using preferences matrices is still in the experimental stage.

**Future plans.** As is shown in the *unc-22* example (Figure 6), two genes can occupy the same space of DNA; a gene lies in the middle of a 9 kb intron of *unc-22*. This makes detecting some introns difficult if not impossible. Our current research is investigating ways to detect this situation and better characterize gene entities.

The future plan for gm is to automate parameter setting. Very often, the correct exon map is not created due to one bad splice site, or an exon with poor codon usage, or an intron has a region that looks like coding region, etc.; gm will build exon maps allowing a threshold of "bad spots" which are corrected on a second pass.

## Acknowledgements

Part of this manuscript was written at the Aspen Center of Physics during a Workshop on Recognizing Genes. This research was supported in part by US Department of Energy, Genome grant 89ER60865 to C.A.F and C.A.S.

## References

1. G. Stormo, "Computer methods for analyzing sequence recognition of nucleic acids," *Ann. Rev. Biophysics. Chem.*, vol. 17, pp. 241-263, 1988.
2. R. Doolittle, ed., "Molecular evolution: computer analysis of protein and nucleic acid sequences," *Methods of Enzymology*, vol. 183, Academic Press, Inc., 1990.
3. R. Staden, "Computer methods to locate signals in nucleic acid sequences," *Nucleic Acids Res.*, vol. 12, pp. 505-519, 1984.
4. G. Stormo, T. Schneider, L. Gold, and A. Ehrenfeucht, "Use of the 'Perceptron' algorithm to distinguish translation initiation site in *E. coli*," *Nucleic Acid Res.*, vol. 10, pp. 2997-3011, 1982.
5. J. C. W. Shepherd, "Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification," *Proc. Natl. Acad. Sci. USA*, vol. 78, pp. 1596-1600, 1981.
6. C. Csank, F. Taylor, and D. Martindale, "Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes," *Nucleic Acids Res.*, vol. 18, no. 17, pp. 5133-5141, 1990.
7. L. Bougueleret, F. Tekaia, I. Sauvaget, and J. Claverie, "Objective comparison of exon and intron sequences by the mean of 2-dimensional data analysis methods," *Nucleic Acids Res.*, vol. 16, no. 5, pp. 1729-1738, 1988.
8. R. Breitbart, A. Andreadis, and B. Nidal-Ginard, "Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes," *Annu. Rev. Biochem.*, vol. 56, pp. 467-495, 1987.
9. R. Farber, A. Lapedes, and K. Sirotkin, "Determination of eucaryotic protein coding regions using neural networks and information theory," Preprint.
10. S. Brunak, J. Engelbrecht, and S. Knudsen, "Prediction of human mRNA donor and acceptor sites from the DNA sequence," *J. Mol. Biol.*, vol. 220, pp. 49-65, 1991.
11. E. Uberbacher and R. Mural, "Locating protein coding regions in human DNA sequences using a neural network - multiple sensor approach," Preprint.
12. J. Knight and E. Myers, 1991. Personnel communication and presentation at the Workshop on Recognizing Genes at the Aspen Center of Physics.
13. R. Guigo, S. Knudsen, N. Drake, and T. Smith, "Prediction of Gene Structure", Submitted.
14. C. A. Soderlund, P. Shanmugam, and C. A. Fields, *gm User's Manual*, 1989. Technical Report MCCS-89-158, Computing Research Laboratory, New Mexico State University.
15. G. Benian, J. Kiff, N. Neckelmann, D. Moerman, and R. Waterson, "Sequence of an unusually large protein implicated in regulation of myosin activity in *C.elegans*," *Nature*, vol. 342, pp. 45-50, 1989.
16. C. A. Fields and C. A. Soderlund, "gm: a practical tool for automating DNA sequence analysis," *CABIOS*, vol. 6, no. 3, pp. 263-270, 1990.
17. D. George, W. Barker, and L. Hunt, *Nucleic Acids Res.*, vol. 14, no. 1, pp. 14-15, 1986.
18. H. Bilofsky, C. Burks, J. Fickett, W. Goad, F. Lewitter, W. Rindone, C. Swindell, and C. Tung, "The Genbank genetic sequence databank," *Nucleic Acids Res.*, vol. 14, no. 1, pp. 1-9, 1986.
19. S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403-410, 1990.

# The Human THE-LTR(O) and *Mst*II Interspersed Repeats Are Subfamilies of a Single Widely Distributed Highly Variable Repeat Family

C. A. FIELDS,<sup>\*,1</sup> D. L. GRADY,<sup>†</sup> AND R. K. MOYZIS<sup>†</sup>

<sup>\*</sup>Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico 88003-0001; and  
<sup>†</sup>Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Received October 10, 1991; revised January 9, 1992

Fifteen examples of the transposon-like human element (THE) LTR and thirteen examples of the *Mst*II interspersed repeat are aligned to generate new consensus sequences for these human repetitive elements. The consensus sequences of these elements are very similar, indicating that they compose subfamilies of a single human interspersed repetitive sequence family. Members of this highly polymorphic repeat family have been mapped to at least 11 chromosomes. Seven examples of the THE internal sequence are also aligned to generate a new consensus sequence for this element. Estimates of the abundance of this repetitive sequence family, derived from both hybridization analysis and frequency of occurrence in GenBank, indicate that THE-LTR/*Mst*II sequences are present every 100–300 kb in human DNA. The widespread occurrence of members of this family makes them useful landmarks, like *Alu*, L1, and (GT)<sub>n</sub> repeats, for physical and genetic mapping of human DNA. © 1992 Academic Press, Inc.

## INTRODUCTION

The abundant interspersed repetitive sequences *Alu*, L1, and (GT)<sub>n</sub> have proven useful in genomic mapping, both as targets for PCR amplification of adjacent unique sequences (Nelson *et al.*, 1989) and as hybridization tags for fingerprinting random clones (Stallings *et al.*, 1990). These elements exist throughout the human genome and are distributed nonrandomly on each chromosome (Korenberg and Rykowski, 1988; Moyzis *et al.*, 1989). Approximately 10,000 copies of the transposon-like human element (THE) and 30,000 copies of solitary THE long terminal repeats [THE-LTRs, originally called "O elements" (Sun *et al.*, 1984)] have been estimated to exist in the human genome (Paulson *et al.*, 1985). Five complete or nearly complete THEs have been sequenced; aside from terminal deletions in the LTRs, all

are very similar (Paulson *et al.*, 1985; Willard, 1987; Deka *et al.*, 1988).

A number of sequences similar to the THE-LTR, the related *Mst*II repeat (Mermer *et al.*, 1987), or the THE internal sequence have been identified in the course of screening sequences of random genomic clones from human chromosomes 5 and 7 for suitability as sequence tagged sites (STSs; Olson *et al.*, 1989; Green *et al.*, 1991). To improve the efficiency with which additional sequences that are similar to these elements can be identified, we have constructed new multiple alignments of human sequences belonging to these repeat families. The consensus sequences for the THE-LTR and *Mst*II repeats derived from these alignments are similar, suggesting that the THE-LTR and *Mst*II repeats are subfamilies of a single family of interspersed repeats. The alignment of sequences similar to the THE internal sequence suggests that at least 400 nucleotides of this sequence are relatively common in the human genome and that it may be associated with enhanced recombination.

## METHODS

The similarity-searching code FASTA (Pearson and Lipman, 1988) was used to search the primate section of GenBank (Bilofsky and Burks, 1988) release 66 (January 1991) for human sequences similar to the THE-LTR HUMRSO5C (clone O-5 of Sun *et al.*, 1984), the *Mst*II repeat HUMAIGRA (clone 4.1 of Mermer *et al.*, 1987), or the internal part of the THE sequence HUMRSOLTR (clone THE 1-A of Paulson *et al.*, 1985). Sequences were aligned using the multiple-alignment code MALIGN (Sobel and Martinez, 1986) or the pairwise local-alignment code LFASTA (Pearson and Lipman, 1988). The alignments were refined by hand to maximize the number of aligned nucleotides.

## RESULTS AND DISCUSSION

### THE-LTR and *Mst*II Repeats

The alignment of the THE-LTR and *Mst*II sequences is shown in Fig. 1. The *Mst*II alignment is similar to that reported by Mermer *et al.* (1987), who noted the similarity between the five HUMAIGR sequences and the im-

<sup>1</sup> To whom correspondence should be addressed at present address: Receptor Biochemistry and Molecular Biology Section, Park Bldg, Room 405, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD 20892.





munoglobulin  $\epsilon$  heavy chain (IGHE) associated sequences HUMIGCD3 and HUMIGCC5 (Hisajima *et al.*, 1983). Sequences of *Mst*II repeats present in the human insulin receptor (HUMINSRD; Elbein, 1989), serum amyloid A (HUMSAA1A; Sack and Talbot, 1989), histocompatibility HLA-SB (HUMHLASBA; Lawrance *et al.*, 1985), thyroid peroxidase (HUMTPO04; Kimura *et al.*, 1989), and cytochrome P450 (HUMCYP450; Jaiswal *et al.*, 1985) genes have been added in the present alignment. The sequence of one (7-87) of four very similar *Hind*III-*Bam*HI fragments of chromosome 7 sequenced at LANL (Green *et al.*, 1991; GenBank Accession Nos. M74209, M74210, M74211, M74212) that appear to be partial *Mst*II repeats is also included in Fig. 1.

The present *Mst*II consensus length is 451 nucleotides (nt), 28 nt longer than that derived by Mermer *et al.* (1987) from the HUMAIGR sequences. The principal difference between the present alignment and that of Mermer *et al.* is the 11 nt spanning positions 298 to 308 in the present alignment, which do not occur in the HUMAIGR sequences. The sequences cannot, however, be divided cleanly into subfamilies based on the presence or absence of gaps. The *Mst*II sequences are most similar between positions 40 and 170 and between positions 320 and 400, which correspond roughly to the "left arm" and "right arm" described by Mermer *et al.* (1987).

The THE-LTR alignment shown in Fig. 1 includes the O-element sequences HUMRSO4C and HUMRSO5C (Sun *et al.*, 1984), the LTR sequences of a genomic THE (HUMRSOLTR; Paulson *et al.*, 1985) and an extrachromosomal THE (HUMEXCI5; Misra *et al.*, 1987), an unmapped genomic LTR with an *Alu* insertion (HUMTHEP2; Lloyd *et al.*, 1987), isolated LTRs near a 6-16 translocation breakpoint (HUMSATOD; Wong *et al.*, 1990) and a human papillomavirus insertion site (HUMHPP16; Baker *et al.*, 1987), and presumably isolated LTRs present in a U4 snRNA pseudogene (HUMUG4PB; Bark and Pettersson, 1989) and in the apolipoprotein B (HUMAPB03; Ludwig *et al.*, 1987), dystrophin (HUMDYSIN; Bodrug *et al.*, 1987), and S-protein (HUMPROSA; Lundwall *et al.*, 1986) genes. The LTR sequences of HUMRSOLTR are representative of those of the other known complete THEs (Willard, 1987; Deka *et al.*, 1988); thus, the latter have not been included in Fig. 1. Sequences of two random *Hind*III-*Bam*HI fragments of chromosome 5, 5-17, and 5-57 obtained at LANL (GenBank Accession Nos. M74207 and M74208) are also included in the alignment.

The present THE-LTR consensus length is 400 nt, 8 nt longer than that obtained by Willard (1987) for the LTRs of complete THEs. The THE-LTR sequences are most homogeneous between positions 60 and 110 and between positions 250 and 380. Gap polymorphisms occur spanning positions 49-59, 111-120, and 195-205; however, these do not divide the sequences cleanly into subfamilies.

The *Mst*II and THE-LTR consensus sequences align over their entire lengths, with no consistent gaps of over 30 nt. Although the THE-LTR consensus length is 51 nt

shorter than that of *Mst*II, the full-length alignment of the two sets of sequences suggests that they are subfamilies of a single family of interspersed repeats. The similarity between the *Mst*II sequences and the THE-LTR was noted by Mermer *et al.* (1987), who aligned the HUMAIGR sequences with the THE-LTR HUMRSO5C; however, their alignment assumes an 80-nucleotide gap in HUMRSO5C that is closed in the present alignment. The alignment of the THE-LTR and *Mst*II sequences includes numerous gaps and several regions, primarily between positions 200 and 350 of the *Mst*II sequence, containing multiple mismatches. The two sets of sequences are most similar in the first and last 100 nt.

The repeat family that includes the THE-LTR and *Mst*II sequences is evidently widespread in the genome. The current alignment includes sequences mapped to 10 autosomes and the X and to an extrachromosomal circle (Fig. 1). Estimates of the abundance of these sequences based on hybridization data range from 30,000 for the THE-LTR (Paulson *et al.*, 1985) to 5000 for the *Mst*II repeat (Mermer *et al.*, 1987); the high degree of polymorphism in the family suggests that the larger of these abundance estimates may be the more reliable. An abundance of roughly  $10^4$  copies per haploid genome is consistent with finding 25 examples in GenBank release 66, assuming that GenBank sequences are a representative sample of human DNA. A sample of 1 Mb of human cosmid ~~containing~~-DNA (Riethman *et al.*, 1989; Stallings *et al.*, 1990) was found to contain 12 sequences hybridizing to a THE-LTR/*Mst*II consensus oligomer (data not shown), indicating an average spacing of hybridizing sequences of once per 83 kb. Taken together, these data suggest that members of this repeat family are present, on average, once every 100 to 300 kb.

### THE Internal Sequences

The family comprising the THE-LTR and *Mst*II repeats is much more polymorphic than the *Alu* family, in which there are no large gaps, and subfamily divisions are made on the basis of a few nucleotide positions (Labuda and Striker, 1989; Jurka and Milosavljevic, 1991). It is thus far unknown whether THEs are transpositionally active, and if so, what the sequence requirements of active THE-LTRs or internal sequences are. The 1.6-kb internal sequences of the known complete THEs align over their entire lengths (Willard, 1987), suggesting that complete genomic THEs are relatively homogeneous. However, only the final 1 kb of the extrachromosomal THE HeLa5R (HUMEXCI5; Misra *et al.*, 1987), which consists of 1.6 kb followed by a single LTR, is similar to this conserved internal sequence. The final 673 nt of two 893-nt sequences previously identified as hot spots of recombination in the immunoglobulin heavy-chain (IGH) region on chromosome 14, HUMIGHHSC 3 and 4 (Keyeux *et al.*, 1989), are highly similar to the THE internal sequence, suggesting that the identified recombination event was mediated by two copies of THE. Three random genomic clones obtained

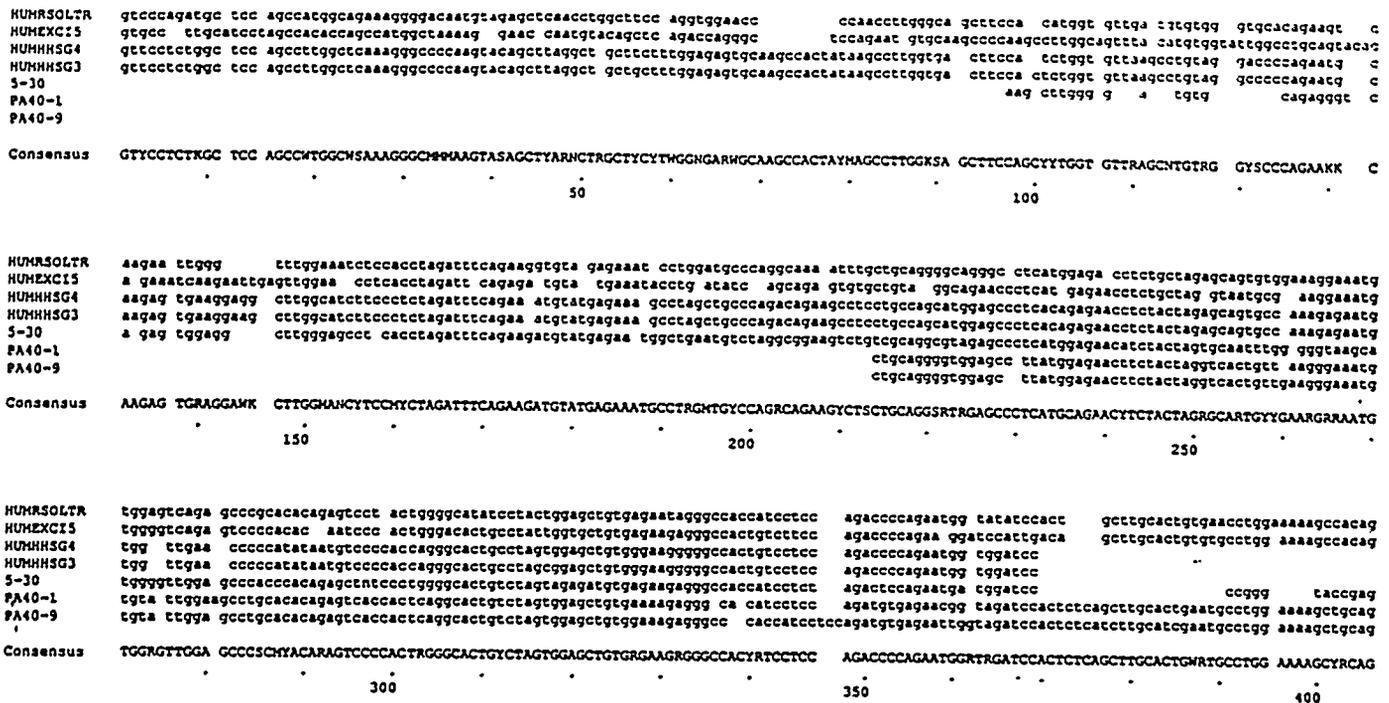


FIG. 2. Alignment of sequences similar to the THE 1-A internal sequence HUMRSOLTR (Paulson *et al.*, 1985), in the orientation of HUMRSOLTR in GenBank. The alignment starts at position 1103 of the HUMRSOLTR sequence, which corresponds to position 1111 in Fig. 6 of Misra *et al.* (1987), and extends to the end of the PA40 sequences. The THE 1-A and HeLa5R sequences are similar for 148 nt upstream of this aligned region, and dissimilar for an additional 600 nt upstream (Misra *et al.*, 1987); the HUMIGHHS3 sequences are similar to THE 1-A for 325 nt upstream of this aligned region, and dissimilar for an additional 220 nt upstream.

at LANL, a partial *Hind*III-*Bam*HI fragment of chromosome 5 and two unmapped *Pst*I fragments (GenBank Accession Nos. M74213, M74214, M74215), are also similar to THE internal sequences. These sequences are aligned in Fig. 2, and establish that THE internal sequences are present on at least two autosomes (5 and 14) as well as extrachromosomal circles.

In summary, consensus nucleotide sequences of the THE-LTR/*Mst*II repetitive sequence family have been generated, using both published and newly sequenced members of the family. These consensus sequences are valuable for screening genomic DNA sequences to identify regions where useful STSs can be constructed (Olson *et al.*, 1989; Green *et al.*, 1991). Due to their high abundance and diversity, members of this family will make useful landmarks for physical and genetic mapping of the human genome (Stallings *et al.*, 1990).

ACKNOWLEDGMENTS

This work was supported in part by U.S. Department of Energy Genome Program Grants 89ER60865 to C.A.F. and B04836/F137 and B04718/F518 to R.K.M. We thank an anonymous referee for bringing the work of Willard (1987) to our attention.

REFERENCES

Baker, C. C., Phelps, W. C., Lindgren, V., Braun, M. J., Gonda, M. A., and Howley, P. M. (1987). Structural and transcriptional analysis of human papillomavirus type 16 sequences in cervical carcinoma cell lines. *J. Virol.* 61: 962-971.  
 Bark, C., and Pettersson, U. (1989). Nucleotide sequence for and orga-

nization of full-length human U4 RNA pseudogenes. *Gene* 80: 385-389.  
 Bilofsky, H. S., and Burks, C. (1988). The GenBank genetic sequence data bank. *Nucleic Acids Res.* 16: 1861-1863.  
 Bodrug, S. E., Ray, P. N., Gonzalez, I. L., Schmickel, R. D., Sylvester, J. E., and Worton, R. G. (1987). Molecular analysis of a constitutional X-autosome translocation in a female with muscular dystrophy. *Science* 237: 1620-1624.  
 Deka, N., Wong, E., Matera, A. G., Kraft, R., Lienward, L., and Schmid, C. (1988). Repetitive nucleotide sequence insertions into a novel calmodulin-related gene and its processed pseudogene. *Gene* 71: 123-134.  
 Elbein, S. C. (1989). Molecular and clinical characterization of an insertional polymorphism of the human insulin receptor gene. *Diabetes* 38: 737-743.  
 Green, E. D., Mohr, R. M., Idol, J. R., Jones, M., Buckingham, J. M., Deaven, L. L., Moyzis, R. K., and Olson, M. V. (1991). Systematic generation of sequence-tagged sites (STSs) for physical mapping of human chromosomes: Application to the mapping of human chromosome 7 using yeast artificial chromosomes. *Genomics* 11: 548-564.  
 Hisajima, H., Nishida, Y., Nakai, S., Takahashi, N., Ueda, S., and Honjo, T. (1983). Structure of the human immunoglobulin Cε2 gene, a truncated pseudogene: Implications for its evolutionary origin. *Proc. Natl. Acad. Sci. USA* 80: 2995-2999.  
 Jaiswal, A. K., Gonzalez, F. J., and Nebert, D. W. (1985). Human P1-450 gene sequence and correlation of mRNA with genetic differences in benzo[a]pyrene metabolism. *Nucleic Acids Res.* 13: 4503-4520.  
 Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human *Alu* genes. *J. Mol. Evol.* 32: 105-121.  
 Keyeux, G., LeFranc, G., and LeFranc, M.-P. (1989). A multigene deletion in the human IGH constant region locus involves highly homologous hot spots of recombination. *Genomics* 5: 431-441.

- Kimura, S., Hong, Y.-S., Kotani, T., Ohkati, S., and Kikkawa, F. (1989). Structure of human thyroid peroxidase gene: Comparison and relationship to the human myeloperoxidase gene. *Biochemistry* 28: 4481-4489.
- Korenberg, J., and Rykowski, M. (1988). Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. *Cell* 53: 391-400.
- Labuda, D., and Striker, G. (1989). Sequence conservation in Alu evolution. *Nucleic Acids Res.* 17: 2477-2491.
- Lawrence, S. K., Das, H. K., Pan, J., and Weissman, S. M. (1985). The genomic organization and nucleotide sequence of the HLA-SB(DP) alpha gene. *Nucleic Acids Res.* 13: 7515-7528.
- Lloyd, J. A., Lamb, A. N., and Potter, S. S. (1987). Phylogenetic screening of the human genome: Identification of differentially hybridizing repetitive sequence families. *Mol. Biol. Evol.* 4: 85-98.
- Ludwig, E. H., Blackhart, B. D., Pierotti, V. R., Caiati, L., Fortier, C., Knott, T., Scott, J., Mahley, R. W., Levy-Wilson, B., and McCarthy, B. J. (1987). DNA sequence of the human apolipoprotein B gene. *DNA* 6: 363-372.
- Lundwall, A., Dackowski, W., Cohen, E., Shaffer, M., Mahr, A., Dahlback, B., Stenflo, J., and Wydro, R. (1986). Isolation and sequence of the cDNA for human protein S, a regulator of blood coagulation. *Proc. Natl. Acad. Sci. USA* 83: 6716-6720.
- Mermer, B., Colb, M., and Krontiris, T. G. (1987). A family of short, interspersed repeats is associated with tandemly repetitive DNA in the human genome. *Proc. Natl. Acad. Sci. USA* 84: 3320-3324.
- Misra, R., Shih, A., Rush, M., Wong, E., and Schmid, C. W. (1987). Cloned extrachromosomal circular DNA copies of the human transposable element THE-1 are related primarily to a single type of family member. *J. Mol. Biol.* 196: 233-243.
- Moyzis, R. K., Torney, D. C., Meyne, J., Buckingham, J. M., Wu, J.-R., Burks, C., Sirotkin, K. M., and Goad, W. B. (1989). The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* 4: 273-289.
- Nelson, D. L., Ledbetter, S. A., Corbo, L., Victoria, M. F., Ramirez-Solis, R., Webster, T. D., Ledbetter, D. H., and Caskey, C. T. (1988). Alu polymerase chain reaction: A method for rapid isolation of human-specific sequences from complex DNA sources. *Proc. Natl. Acad. Sci. USA* 86: 6686-6690.
- Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* 245: 1434-1435.
- Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schlinder, C. W., Rush, M. G., Kadyk, L., and Leinwand, L. (1985). A transposon-like element in human DNA. *Nature* 316: 359-361.
- Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.
- Riethman, H. C., Moyzis, R. K., Meyne, J., Burke, D. T., and Olson, M. V. (1989). Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. *Proc. Natl. Acad. Sci. USA* 86: 6240-6244.
- Sack, G. H., Jr., and Talbot, C. C., Jr. (1989). The human serum amyloid A (SAA)-encoding gene GSAA1: Nucleotide sequence and possible autocrine-collagenase-inducer function. *Gene* 84: 509-515.
- Sobel, E., and Martinez, H. M. (1986). A multiple sequence alignment program. *Nucleic Acids Res.* 14: 363-374.
- Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A., and Moyzis, R. K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci. USA* 87: 6218-6222.
- Sun, L., Paulson, K. E., Schmid, C. W., Kadyk, L., and Leinwand, L. (1984). Non-Alu family interspersed repeats in human DNA and their transcriptional activity. *Nucleic Acids Res.* 12: 2669-2690.
- Willard, C. (1987). "Structure and Evolution of Repeated DNA," Ph.D. dissertation, University of California, Davis.
- Wong, Z., Royle, N. J., and Jeffreys, A. J. (1990). A novel human DNA polymorphism resulting from transfer of DNA from chromosome 6 to chromosome 16. *Genomics* 7: 222-234.

## **Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin.**

Owen White, Carol Soderlund(†), Pari Shanmugan and Chris Fields(‡)

Computing Research Laboratory, Box 30001/3CRL  
New Mexico State University  
Las Cruces, NM 88003-0001 USA

† Present Address: Theoretical Biology and Biophysics (T-10), Los Alamos National Laboratory, Los Alamos, NM, 87545.

‡ Corresponding author

*keywords:* computation, dinucleotide, information, intron, splice site

## ABSTRACT

The DNA sequence composition of 526 dicot and 345 monocot intron sequences have been characterized using computational methods. Splice site information content and bulk intron and exon dinucleotide composition were determined. Positions 4 and 5 of 5' splice sites contain different statistically significant levels of information in the two groups. Basal levels of information in introns are higher in dicots than in monocots. Two dinucleotide groups, WW (AA, AU, UA, UU) and SS (CC, CG, GC, GG) have significantly different frequencies in exons and introns of the two plant groups. These results suggest that the mechanisms of splice-site recognition and binding may differ between dicot and monocot plants.

The mechanism of pre-mRNA splicing is now understood in considerable detail [1, 2, 3]; however, the molecular recognition of splice sites by spliceosome components is still not well characterized. Goodall and Filipowicz have identified a bias for A and U nucleotides in plant introns, and hypothesize that a high A+U content and consensus intron-exon borders may be the only sequence requirements for plant intron processing [4]. A minimal functional length of 70-73 nucleotides has been observed for introns in monocots and dicots [5], despite a heterogeneous distribution of intron lengths between the two plant groups. Total genome dinucleotide frequencies have been measured experimentally in a variety of organisms, including plants [6, 7]. Unique distributions of dinucleotides have been observed in DNA separated by evolutionary origin [8, 9, 10], organelle [11], and function [12].

The unique ability of monocots to process introns high in G+C suggests differences in dicot and monocot pre-mRNA processing mechanisms [13]. For example, pre-mRNA of 1,5-bisphosphate carboxylase from pea, a dicot, was efficiently spliced in transgenic tobacco plants but the same gene from wheat, a monocot, was not processed as efficiently in transgenic tobacco [14]. While experimental differences in dicot-monocot splicing specificities exist, both dicot and monocot genes are spliced with similar efficiency *in vitro* by HeLa cell extracts [15], and by an autonomously replicated vector in transient expression assays of tobacco leaf disks [16]. Stable incorporation of a phaseolin gene fused to the Cauliflower Mosaic Virus promoter resulted in equally efficient pre-mRNA processing in tobacco and rice cell lines [17].

Experimental evidence suggests that plant introns may have pre-mRNA recognition mechanisms that are different from those of vertebrate systems [13 and references therein]. Calculations of the information content of binding sites have provided some insight into site recognition [18, 19]. Information content analysis of intron splice sites in the nematode *Caenorhabditis elegans* has been previously described [20]; this analysis showed that splice sites in *C. elegans* vary with intron length. In this report we show that: 1) the information contents of splice sites differ between monocots and dicots; 2) the basal level of information is an average of 0.1 bit/base higher in introns than in exons for dicots; 3) dinucleotide usage differs between exons and introns, and between the two plant groups.

Plant DNA sequences were extracted from GenBank release 68 (June, 1991). Entries containing full or partial-length sequences, from 172 loci in dicots (Magnoliopsida) and 84 loci in monocots (Liliopsida) were used in our analysis. The coding portions of these sequences encode enzymes, structural proteins, storage proteins and peptides of unknown function. Sequences were

discarded in cases where intron/exon splice junctions were ambiguously determined. Overrepresenting of certain sequences due to oversampling of some gene families cannot be excluded. Sequences from 20 nucleotides upstream to 30 nucleotides downstream of the 5' splice site were aligned, and information content,  $I(n)$  in units of bits/position, was calculated as:

$$I(n) = \left( \sum_{B=A, C, G, U} F(B, n) \log_2 F(B, n) \right) - \left( \sum_{B=A, C, G, U} P(B, n) \log_2 P(B, n) \right)$$

where  $F(B, n)$  is the observed frequency of base  $B$  in position  $n$ , and  $P(B, n)$  is the prior probability of base  $B$  in position  $n$  [18]. The expression  $I(n)$  represents the information contained in a single nucleotide position, as the result of elevated usage of a particular nucleotide or nucleotides at that position in a DNA sequence. The prior base probability, as measured by base frequency, has been shown to vary between introns and exons, and between dicots and monocots [4]. To visualize the differences in base composition across the exon-intron boundaries in calculations of  $I(n)$ , the prior base probabilities were set to the equiprobable values (i.e.,  $P(B, n)=0.25$  for  $B=A, C, G, U$  in all positions). Similar information contents from sequences 30 nucleotides upstream to 20 nucleotides downstream were obtained from 3' splice site junctions. Standard deviations were calculated using the 'exact method' described in [18] and error bars for information plots consist of  $\pm$  two times the standard deviation. For basal information measurements, the 0.5% and 99.5% confidence limits were determined by numerical simulations, based on observed average nucleotide frequencies.

A dinucleotide ( $BB'$ ) is any two adjacent nucleotide bases. Dinucleotide frequencies were measured from both exons and introns. Two components of DNA composition are reflected in simple dinucleotide frequency measurements. One component of dinucleotide frequencies is that part strictly due to the underlying distribution of single nucleotide frequencies; e.g., increases in A and U would increase the random occurrence of AA, AU, UA, and UU. The second component of dinucleotide frequency is that part which reflects the correlation between the two nucleotides. To reduce the component of dinucleotide frequency that merely reflects single nucleotide distributions, the observed frequency of each dinucleotide was divided by the expected frequency of the dinucleotide, using the formula:

$$F'(BB') = \frac{F(BB')}{F(B) \times F(B')}$$

where  $F(BB')$  is the observed dinucleotide frequency and  $F(B)$  and  $F(B')$  are the observed single nucleotide frequencies for two single nucleotides. The logarithm of  $F'(BB')$  is referred to as the "mutual information" of  $B$  and  $B'$  [21]. Expected dinucleotide frequencies for introns were calculated using measured single nucleotide frequencies from intron regions, and expected exon dinucleotide frequencies were calculated using measured single nucleotide frequencies from exon regions. Log-likelihood ratios, demonstrating differences in non-independent dinucleotide usage between dicots and monocots, were derived by the formula:

$$R(BB') = \log_2 \left( \frac{F'(BB')_{di}}{F'(BB')_{mono}} \right)$$

where  $R(BB')$  is the log-likelihood ratio of the dinucleotide combination, and  $F'(BB')_{di}$  and  $F'(BB')_{mono}$  are the corrected dinucleotide frequencies for dicots and monocots, respectively. Our software to generate information contents, base frequency matrices, baseline frequency simulations, dinucleotide counts, corrected dinucleotide frequencies, and log-likelihood ratios is available on request.

Information contents of the splice junctions of both dicots and monocots are shown in Figure 1. The observed basal level of information across the intronic portion of dicot sequences differs from the exon basal level of information. This asymmetric distribution of basal information is not apparent in monocot sequences. We tested whether the elevation of basal information is due solely to an enrichment of A+U in dicot introns. Numerical experiments examining the range of possible baseline information values that can occur from a given average frequency of nucleotides demonstrated that baseline information is highly dependent on subtle changes in nucleotide frequency. Using the average nucleotide frequencies measured between positions 10 and 50 in monocot introns, which are 60% A+U, baseline information values are expected to range from 0.047 to 0.064, within 99.5% confidence limits. Nucleotide frequencies from the same portion of dicot introns (72% A+U), gave expected baseline information values of 0.155 to 0.180, within 99.5% confidence limits. Thus the observed A+U frequencies alone are sufficient to account for the observed basal information contents in introns of both dicots and monocots.

The information contents of dicot and monocot splice sites are similar, but not identical. A shoulder of the 5' peak extending into the coding portion of the splice site (positions -2 and -1) contains  $1.53 \pm 0.15$  and  $1.48 \pm 0.20$  bits of information in dicots and monocots, respectively. Not

including the conserved GU, the intronic portion of the splice site (positions +3 to +6) encodes  $1.73 \pm 0.18$  and  $1.59 \pm 0.21$  bits of information in dicots and monocots, respectively. In total (positions -3 to +6)  $7.26 \pm 0.24$  bits are observed in dicot 5' splice sites, while  $7.07 \pm 0.29$  bits are contained in monocot 5' splice sites. Statistically significant differences in information content between the two plant groups can be observed at specific nucleotide positions near the 5' splice site. At position 4 of the 5' splice site,  $0.25 \pm 0.11$  more bits of information are found in dicots than in monocots, while at position 5,  $0.21 \pm 0.14$  more bits are encoded in monocots than in dicots. Nucleotide counts at positions -3 to +8 of the 5' splice sites are presented in Table 1. The consensus sequence is AGIGUAAGU for both monocot and dicot 5' splice sites.

Similar differences in information contents are found in dicot and monocot 3' splice sites. Position +1, the exonic portion of the 3' splice site, contains  $0.39 \pm 0.09$  bits of information in dicots, and  $0.50 \pm 0.12$  bits in monocots. Significant information is encoded at positions -3 and -5 of plant 3' splice sites ( $1.30 \pm 0.26$  and  $1.53 \pm 0.28$  bits for dicots and monocots, respectively). This is primarily due to an enrichment of pyrimidine in position -3 and U in position -5. Monocot 3' splice sites contain fewer U at -3 than dicots, resulting in consensus sequences of UGYAGIGG for dicots and UGCAGIGG for monocots. In total, from positions -5 to +1,  $5.93 \pm 0.28$  bits of information are found in dicot 3' splice sites, and  $6.24 \pm 0.32$  bits of information are contained in monocot 3' splice sites. Discrete base frequency differences exist between the two plant groups in their 3' splice sites. Dicot splice sites use C in lowest abundance in positions -8, -7, -6, -5, -4, +2 and +3 around the splice site (the conserved AG at -1 and -2 were not considered). The nucleotide C occurs least frequently in only position -4 in monocots (A is as infrequent as C at position -6).

Lariat branch point sequences are in greatest abundance in the window between -50 and -1 of the 3' region in invertebrate, primate, plant and rodent introns [22]. Local information maxima are observed in dicot introns at positions -17 and -19 with respect to the 3' splice site; these do not appear in monocot sequences. These maxima are due to an increase of U at positions -17 and -19. We did not detect a significant association of the intron branch consensus URAY [23] with Us at these positions (data not shown). In dicots, 455 strict matches to the branch site URAY, out of a possible 526 sequences (86%), were detected in the -50 to -1 3' portion of introns. In the same region of monocot introns, 260 potential branch sites were detected in a total of 345 sequences (75%).

Raw dinucleotide frequencies and mutual information values are presented in Figure 2 for exons and introns of both plant groups. The overall frequencies of the nucleotides A+U in introns

are 71% and 61% for dicots and monocots, respectively. In exons, A+U occur at 55% and 42% in dicots and monocots, respectively. These values are in close agreement with those of Goodall and Filipowicz [4], as measured in a smaller data set. In the introns of both plant groups, SS dinucleotides (CC, CG, GC, GG) occur with the lowest frequency. The WW dinucleotides (AA, AU, UA, UU) are the 4 most frequent in dicot introns, while they are 4 of the 5 most frequent in monocot introns (UG is more common than AA or UA). Dinucleotide abundances are reversed in monocot exons, where WW dinucleotides become the 4 least frequent dinucleotides, while SS dinucleotides are 4 out of the 5 most common dinucleotides (CA is more common than CG). This abundance reversal does not occur in dicot exons, where SS dinucleotides remain as 4 of the 6 least common dinucleotides. Quite opposite from monocot exons, AA is the most abundant dinucleotide in dicot exons. The mutual information values (lower panel, Fig. 2) show that despite the overall consistency between dinucleotide frequencies of their component single nucleotides frequencies, some dinucleotide levels are lower than expected. The dinucleotide CG, which is a potential methylation site and is rare in vertebrate genomes [24] is under-represented in introns and exons of both plant groups. The dinucleotide UA also occurs much less frequently than expected in dicot and monocot exon sequences, perhaps because in-frame UAs either encode stops or tyrosine, which is a relatively rare amino acid. Other dinucleotides, such as CA and UG occur more frequently than expected considering their component single nucleotide frequencies. Mutual information for GC dinucleotides in monocot introns is higher than in exons and could serve as a possible distinguishing sequence feature for intron recognition.

Log-likelihood ratios for dinucleotides between exons and introns of the two plant groups are shown in Figure. 3. The largest differences are in the frequencies of the SS dinucleotides, with CC and GG preferred in dicots and CG and GC preferred in monocots. The SS and WW dinucleotides in exons, particularly CC, CG, UA and UU exhibit the greatest abundance differences of all the dinucleotides. The SS dinucleotides in introns demonstrate the largest extremes in abundance.

The differences in splice site structure and intron composition between dicots and monocots are similar to differences found between animal species. The SS and WW dinucleotides have different frequencies in exons and introns in many genomes [9, 25]. Intron sequences in *C.elegans* have high A+U content and elevated basal information content [20] similar to that found in dicots, while human intron sequences have basal information contents lower than that of monocots (data not shown). The information encoded in 5' splice sites varies widely between species, primarily due to differences in the information encoded at positions +4 and +5 (the AG of the universal

GIGUAAGU consensus). Position +5 encodes  $0.24 \pm 0.07$  bits in dicots,  $0.45 \pm 0.12$  bits in monocots,  $0.77 \pm 0.22$  bits in *C. elegans* [20],  $1.08 \pm 0.19$  bits in *Drosophila* [26], and  $1.15 \pm 0.08$  bits in primates (calculated from Table 1 in [27]). No significant differences between either dicot or monocot introns of different lengths, as reported for *C. elegans* introns [20], were observed.

The dinucleotide composition and information content differences reported here raise the possibility that alternate mechanisms for splice-site recognition or pre-mRNA processing may be used by dicots and monocots. The sequence variation observed at the exon-intron junctions suggests that U1 snRNA recognition may not always occur at the same nucleotide positions in the 5' splice site [28], or that additional undetermined factors may be involved in the recognition of either 5' or 3' splice sites. The compositional differences between monocots and dicots may also reflect the use of different mechanisms for exon or intron recognition between the two plant groups. A combination of specific splice-site recognition by snRNPs and recognition of exons by bulk composition is suggested, for example, by the exon-definition model of Berget et al. [29, 30]. A site- and composition-sensitive mechanism along these lines may be active in plants.

It has been shown that high mobility group (HMG) proteins bind to A+U-rich regions outside of plant genes [31]. Because of their presence in actively transcribed genes, HMGs have been implicated in transcriptional activation, perhaps by a mechanism involving conformational changes in chromatin [32]. In view of the elevated WW dinucleotide content in dicot introns, we suggest that HMG binding may not be exclusively confined to the flanking regions of genes, but may occur in intron sequences as well. Consistent with this possibility, HMGs have been demonstrated to bind to an intron portion of the N-20 gene, in soybean [33]. Whether compositionally-directed binding of proteins to bulk intron sequences plays any role in splice-site selection is unknown.

## Acknowledgments

This work was supported in part by U.S. Department of Energy Genome Program Grant number 89ER60865 to CA Fields and CA Soderlund. We also would like to thank Virginia Walbot for valuable discussion and Ted Dunning for technical assistance with the statistical modeling.

## References

1. Green MR: Pre-mRNA splicing. *Annu Rev Genet* 20:671-708 (1986).
2. Maniatis T, Reed R: The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* 325:673-678 (1987).
3. Sharp PA: Splicing of messenger RNA precursors. *Science* 235:766-771 (1987).
4. Goodall GJ, Filipowicz W: The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* 58:473-483 (1989).
5. Goodall GJ, Filipowicz W: The minimum functional length of pre-mRNA introns in monocots and dicots. *Plant Mol Biol* 14: 727-733 (1990).
6. Josse J, Kaiser AD, Kornberg A: Enzymatic synthesis of deoxyribonucleic acid. *J Biol Chem* 236: 864-875 (1961).
7. Swartz MN, Trautner TA, Kornberg A: Enzymatic synthesis of deoxyribonucleic acid. *J Biol Chem* 237: 1961-1967 (1962).
8. Ohno S: Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. *Proc Natl Acad Sci* 85: 9630-9634 (1988).
9. Kozhukhin CG, Pevzner PA: Genome inhomogeneity is determined mainly by WW and SS dinucleotides. *CABIOS* 7: 39-49 (1991).
10. Nussinov R: Doublet frequencies in evolutionary distinct groups. *Nucl Acids Res* 12: 1748-1763 (1984).
11. Boudraa M, Perrin P: CpG and TpA frequencies in the plant system. *Nucl Acids Res* 15: 5729-5737 (1987).
12. Bentler E, Gelbart T, Han J, Koziol JA, Beutler B: Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc Natl Acad Sci* 86: 192-196 (1989).
13. Goodall GJ, Filipowicz W: Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *EMBO* 10: 2635-2644 (1991).
14. Keith B, Chua N: Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco. *EMBO* 5:2419-2425 (1986).
15. Brown JWS, Feix G, Frendewey D: Accurate *in vitro* splicing of two pre-mRNA plant introns in a HeLa cell nuclear extract. *EMBO* 5:2749-2758 (1986).
16. McCullough AJ, Lou H, Schuler MA: *In vivo* analysis of plant pre-mRNA splicing using an autonomously replicating vector. *Nucl Acid Res* 19: 3001-3009 (1991).
17. Peterhans A, Datta SK, Datta K, Goodall GJ, Potrykus I, Paszkowski J: Recognition efficiency of *Dicotyledoneae*-specific promoter and RNA processing signals in rice. *MGG* 222:361-368 (1990).

18. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: Information content of binding sites on nucleotide sequences. *J Mol Biol* 188: 415-431 (1986).
19. Berg O, von Hippel P: Selection of DNA binding sites by regulatory proteins. *J Mol Biol* 193: 723-750 (1987).
20. Fields C: Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucl Acids Res* 18: 1509-1512 (1990).
21. Hamming RW: Coding and Information Theory. Prentice Hall, Inc, New York (1980).
22. Harris NL, Senapathy P: Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucl Acids Res* 18: 3015-3019 (1990).
23. Brown JWS: A catalogue of splice junctions and putative branch point sequences from plant introns. *Nucl Acids Res* 14: 9549-9559 (1986):
24. Aissani B, G Bernardi: CpG islands: features and distribution in the genomes of vertebrates. *Gene* 106: 173-183 (1991).
25. Fields C, Soderlund CA: gm: a practical tool for automating DNA sequence analysis. *CABIOS* 6:263-270 (1990).
26. Mount SM, C Burks, G Hertz, G Stormo, O White, C Fields: Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Manuscript in preparation.
27. Senapathy P, MB Shapiro, NL Harris: Splice junctions, branch point sites, and exons: sequence statistics, identification and applications to genome project. In: Doolittle RF (ed), *Methods in Enzymology* 183: 252-278. Academic Press, Inc, New York (1990).
28. Jacob M, H Gallinaro: The 5' splice site: phylogenetic evolution and variable geometry with U1RNA. *Nucl Acids Res* 17: 2159-2180 (1989).
29. Robberson BL, GJ Cote, SM Berget: Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol* 10: 84-94 (1990).
30. Talerico M, SM Berget: Effect of 5' splice site mutations on splicing of the preceding intron. *Mol Cell Biol* 10: 6299-6305 (1990).
31. Pedersen TJ, Arwood LJ, Spiker S, Guiltinan MJ Thompson WF: High mobility group chromosomal proteins bind to AT-rich tracts flanking plant genes. *Plant Mol Biol* 16: 95-104 (1991).
32. Spiker S: Histone variants and high mobility group non-histone chromosomal proteins of higher plants: their potential for forming a chromatin structure that is either poised for transcription or transcriptionally inert. *Physiol Plant* 75:200-213 (1988).
33. Gambliel H, Feder I, Sengupta-Gopalan, C: dAdT binding domains of soybean nodulin-20 and french bean  $\beta$ -phaseolin and phytohemagglutinin-L (Lec 2) genes are assembly sites of complex nucleoprotein structures *in vitro*. *Plant Cell* (in submission).

5' splice sites*Dicot introns:**5' position*

	-3	-2	-1	1	2	3	4	5	6	7	8
A	196(37)	321(61)	34(07)	0(0)	0(0)	370(70)	299(56)	116(22)	117(22)	200(37)	178(33)
C	170(32)	49(09)	15(02)	0(0)	2(0)	31(05)	75(14)	51(09)	68(13)	86(17)	98(18)
G	94(18)	47(08)	434(82)	528(100)	0(0)	44(08)	17(03)	261(49)	55(10)	47(08)	37(07)
U	68(12)	111(21)	45(08)	0(0)	526(100)	83(15)	137(26)	100(19)	288(54)	195(37)	215(40)

*Monocot introns:**5' position:*

	-3	-2	-1	1	2	3	4	5	6	7	8
A	147(43)	217(63)	18(05)	0(0)	0(0)	238(68)	152(45)	67(19)	65(18)	122(35)	81(24)
C	113(32)	54(15)	33(09)	0(0)	2(0)	28(08)	87(24)	38(10)	78(22)	47(14)	85(25)
G	62(18)	25(07)	278(80)	346(100)	0(0)	57(16)	23(07)	211(60)	31(09)	68(19)	37(11)
U	24(06)	50(14)	17(05)	0(0)	344(100)	23(07)	84(25)	30(08)	172(50)	109(31)	143(40)

3' splice sites*Dicot introns:**3' position:*

	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	140(26)	109(20)	121(23)	85(16)	163(31)	25(05)	526(100)	0(0)	101(19)	108(20)	156(30)
C	57(11)	63(12)	54(10)	28(05)	30(06)	314(60)	0(0)	0(0)	71(13)	87(16)	79(15)
G	82(16)	93(18)	86(16)	60(11)	223(42)	2(0)	0(0)	526(100)	306(58)	100(19)	145(27)
U	247(47)	261(50)	265(50)	353(67)	110(21)	185(35)	0(0)	0(0)	48(09)	231(44)	146(28)

*Monocot introns:**3' position:*

	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	62(18)	57(16)	63(18)	37(10)	55(15)	11(03)	345(100)	0(0)	50(14)	55(16)	69(20)
C	64(18)	71(21)	63(18)	53(15)	42(12)	281(81)	0(0)	0(0)	43(12)	71(20)	93(27)
G	70(20)	77(22)	71(20)	32(09)	167(48)	4(01)	0(0)	345(100)	220(64)	76(22)	85(24)
U	149(43)	140(40)	148(44)	223(64)	81(23)	49(15)	0(0)	0(0)	22(09)	143(42)	98(28)

**Table 1: Base number matrices for 5' and 3' splice sites. Base frequencies are given in parentheses.**

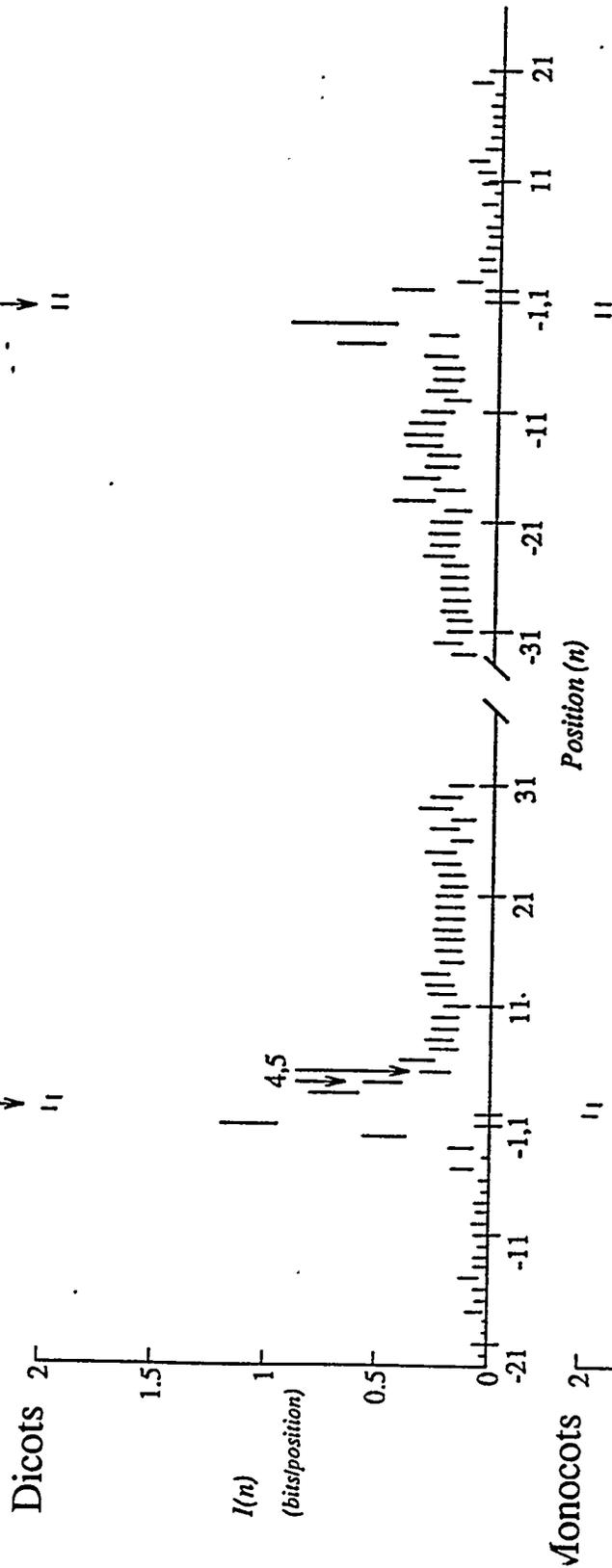
Figure 1: Single-nucleotide information encoded around 5' and 3' splice sites of dicot and monocot introns. Error bars are of +/- two times the standard deviation (within 95% confidence limits). Positions 4 and 5 of 5' splice sites are marked. The error bars are exaggerated in size for visibility on positions +1 and +2 for 5' splice sites and positions -1 and -2 for 3' splice sites.

Figure 2: (Upper panel) Raw dinucleotide frequency histograms for exons (left plot) and introns (right plot) from dicots and monocots. (Lower panel) Mutual information histograms for exons and introns. Corresponding dinucleotides are indicated below each bar. Dicot sequences contain 55% and 71% A+U in exons and introns, respectively. Monocot sequences contain 42% and 61% A+U in exons and introns, respectively. Note compressed y axis on the intron frequency histogram. A mutual information value of 0.0 indicates that the observed dinucleotide frequency exactly equals the dinucleotide frequency expected from the observed single-nucleotide frequencies. Mutual information values greater than zero indicate a greater than expected dinucleotide frequency; values less than zero indicate a less than expected dinucleotide frequency.

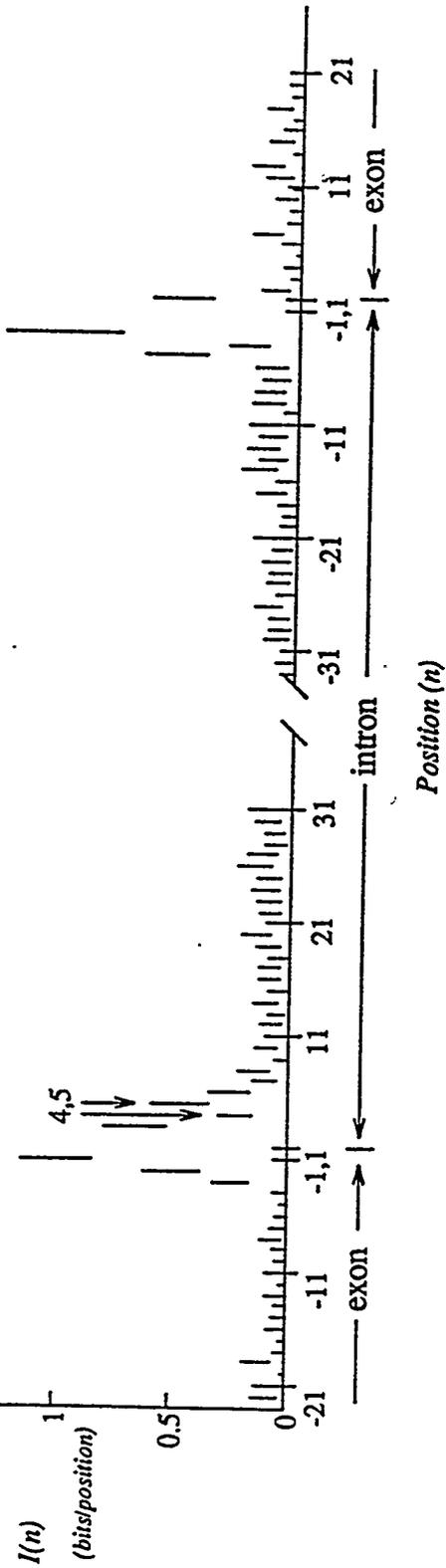
Figure 3: Dinucleotide frequency log-likelihood ratios of dicots to monocots for exons and introns. Corresponding dinucleotides are indicated below each data bar. Values that are more positive indicate higher frequencies in dicots, while values that are more negative indicate higher frequencies in monocots. Many dinucleotides in the SS and WW dinucleotide groups exhibit large asymmetries between exons and introns.

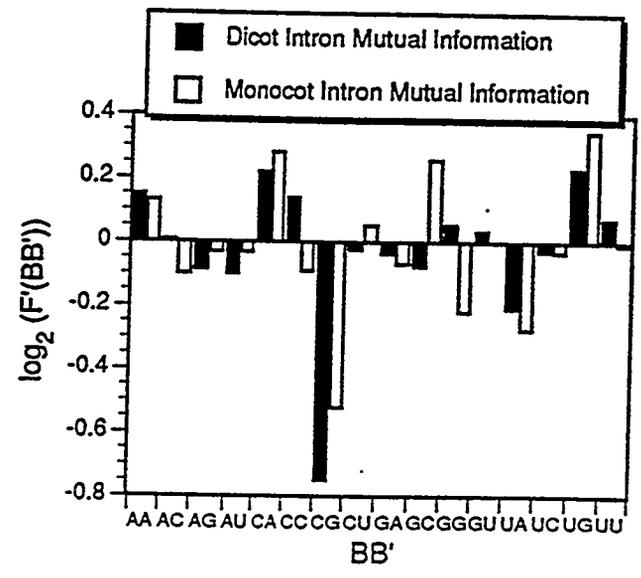
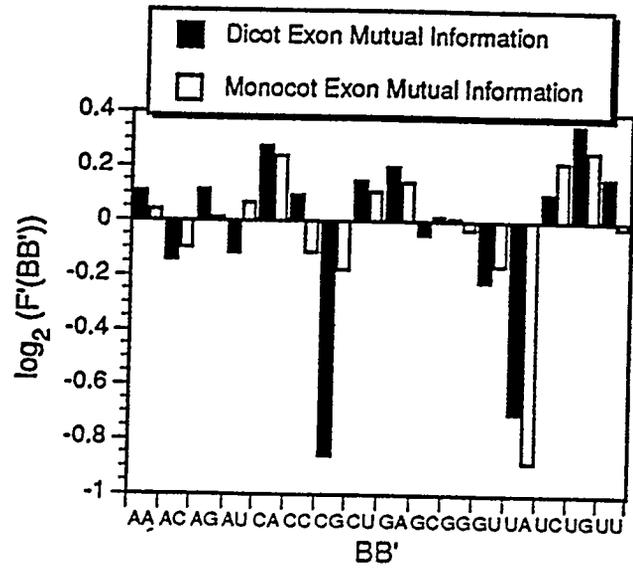
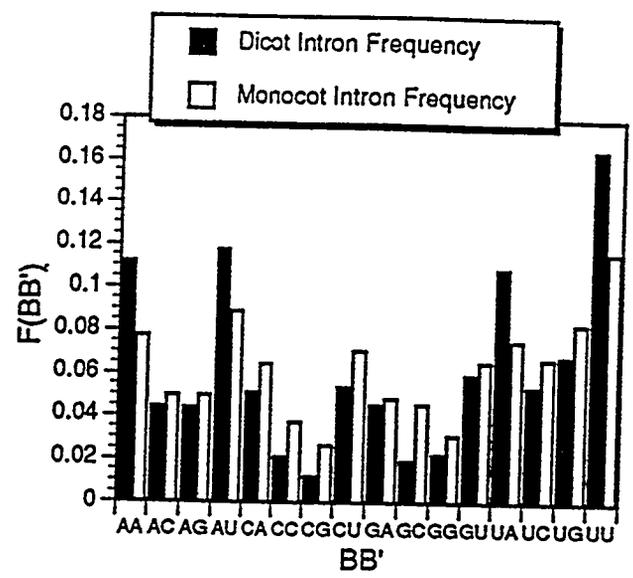
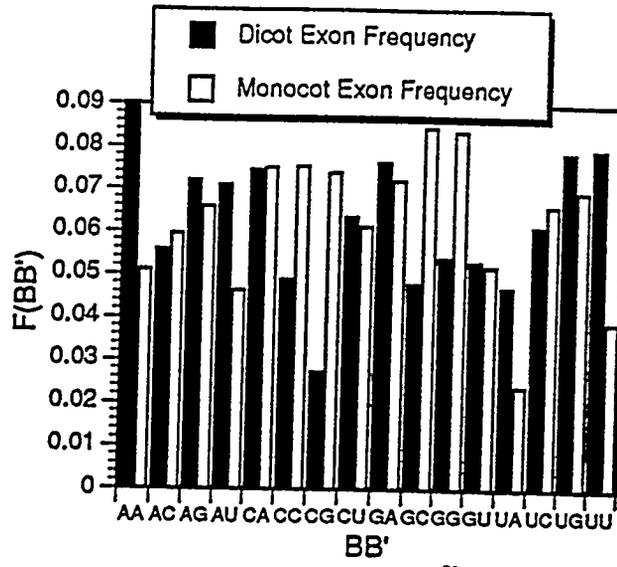
5' splice sites

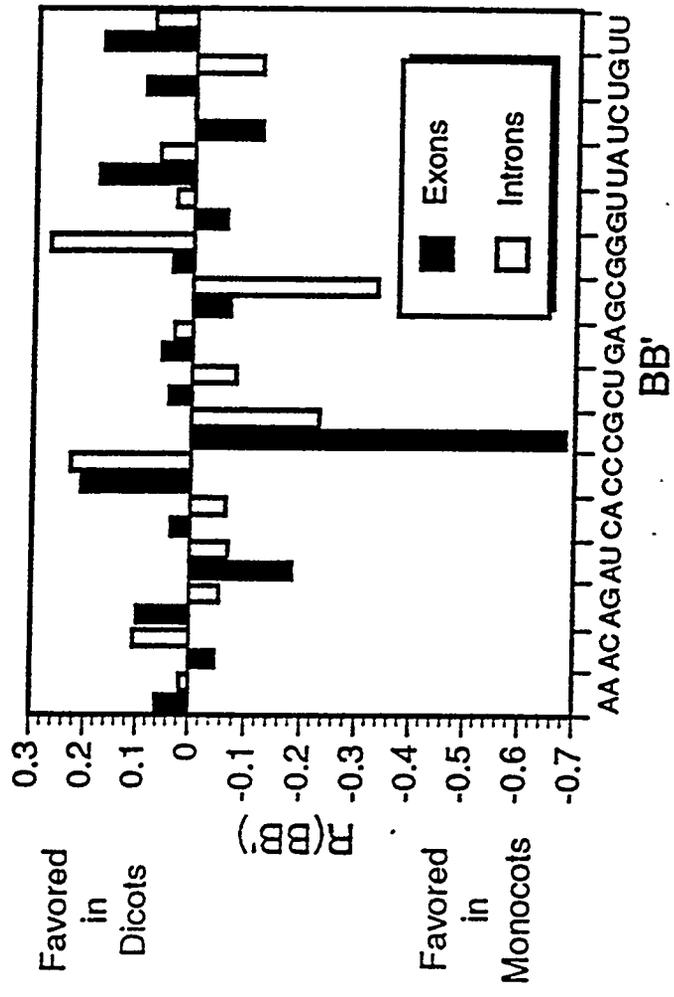
3' splice sites



Monocots







Splicing signals in *Drosophila*: intron size,  
information content, and consensus sequences.

Stephen M. Mount<sup>1,5</sup>, Christian Burks<sup>2</sup>, Gerald Hertz<sup>3</sup>,  
Gary D. Stormo<sup>3</sup>, Owen White<sup>4</sup> and Chris Fields<sup>4</sup>

1. Department of Biological Sciences, Columbia University, New York, NY 10027
2. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545
3. Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309
4. Computing Research Laboratory, Box 300001/3CRL, New Mexico State University, Las Cruces, NM 88003-0001
5. Author to whom correspondence should be addressed.

## ABSTRACT

A database of 209 *Drosophila* introns was extracted from Genbank (release number 64.0) and examined by a number of methods in order to characterize features that might serve as signals for messenger RNA splicing. A tight distribution of sizes was observed: while the smallest introns in the database are 51 nucleotides, more than half are less than 80 nucleotides in length, and most of these have lengths in the range of 59-67 nucleotides. *Drosophila* splice sites found in large and small introns differ in only minor ways from each other and from those found in vertebrate introns. However, larger introns have greater pyrimidine-richness in the region between 11 and 21 nucleotides upstream of 3' splice sites. The *Drosophila* branchpoint consensus matrix resembles C T A A T, and differs from the corresponding mammalian signal in the absence of G at the third position of these five, that immediately preceding the branchpoint. The distribution of occurrences of this sequence suggests a minimum distance between 5' splice sites and branchpoints of about 38 nucleotides, and a minimum distance between 3' splice sites and branchpoints of 15 nucleotides. The methods we have used detect no information in exon sequences other than in the few nucleotides immediately adjacent to the splice sites. However, *Drosophila* resembles many other species in that there is a discontinuity in A+T content between exons and introns, which are A+T rich.

## INTRODUCTION

The removal of introns from the mRNA precursors of higher organisms is a complex process involving many factors (for reviews see Green, 1986; Smith *et al.*, 1989; Guthrie, 1991). The splicing reaction occurs in two steps and proceeds via a branched or "lariat" intermediate in which the 5' end of the intron is joined, via a 2',5' phosphodiester bond, to a site within the intron, usually an A near the 3' splice site. The information required for splicing appears to be limited to sequences adjacent to the three sites involved in the actual chemistry of the splicing reaction (the 5' splice site, the 3' splice site, and the branch point), and a pyrimidine-rich region lying between the 3' splice site and the branchpoint. Factors responsible for the recognition of these sites have been identified. 5' splice sites in many species fit the consensus MAG|GTRAGT (M indicates A or C; R indicates A or G) or some closely related variant of it (Mount, 1982; Shapiro and Senapathy, 1986; Jacob and Gallinaro, 1989; Senapathy *et al.*, 1990). This conserved sequence is recognized by the U1 small nuclear ribonucleoprotein particle (U1 snRNP) through basepairing interactions between the 5' end of U1 RNA and consensus nucleotides (Zhuang and Weiner, 1986), and may also be recognized by additional factors (Zapp and Berget, 1989; Bruzik and Steitz, 1990; Séraphin and Rosbash, 1990; Newman and Norman, 1991). The branchpoint is recognized by the U2 snRNP (Black *et al.* 1985) in a manner that also involves basepairing (Nelson and Green, 1989; Wu and Manley, 1989; Zhuang and Weiner, 1989). In the yeast *Saccharomyces cerevisiae*, the branchpoint sequence UACUAAC is nearly invariant and plays a significant role in determining where and whether splicing will occur. While the identical sequence is preferred in mammalian splicing (Zhuang, Goldstein and Weiner, 1989), a looser consensus

sequence of UNCURAC can be derived for mammalian branchpoints (Nelson and Green, 1989). The interaction between U2 and the branchpoint requires additional factors, including the U1 snRNP (Barabino *et al.*, 1990; Rosbash and Séraphin, 1991), and at least one auxiliary factor, U2AF, that binds to the pyrimidine-rich region that lies between the branchpoint and the 3' splice site (Ruskin *et al.*, 1988). As yet, recognition of the short consensus sequence found at the 3' splice site (YAG|G, or simply AG) has not been definitively attributed to any factor.

There is some species specificity in the interpretation of splicing signals. For example, many introns from the worm *Caenorhabditis elegans* are significantly shorter than vertebrate introns, and this species also appears to have distinct splice site consensus sequences, particularly at the 3' splice site (Blumenthal and Thomas, 1988). As expected from these sequence features, short *C. elegans* introns are not spliced in nuclear extracts derived from human cells (Kay *et al.*, 1987; Ogg *et al.*, 1990). Furthermore, correlations exist between intron size and splice site sequences (Fields, 1990). The 5' splice sites of *C. elegans* introns greater than 75 nucleotides have significantly more information than those of shorter introns, primarily due to conservation at intron positions 4, 5 and 6. In addition, introns in many species, including plants, *C. elegans* and *Drosophila melanogaster*, but not mammals or the yeast *S. cerevisiae*, are significantly more A+T rich than flanking exons (Weibauer *et al.*, 1988; Csank *et al.*, 1990), and the inability of plant introns to splice in mammalian nuclear extracts has been attributed to differences in the A/T content between species (Goodall & Filipowicz, 1989, 1991).

*Drosophila* introns appear to have splice site consensus sequences much like those found

in vertebrates (Senapathy *et al.*, 1990) and a recognizable branchpoint consensus (Keller and Noon, 1984; Rio, 1988; Guo, Lo and Mount, 1992). In addition, known components of the splicing machinery, including U-RNAs (Mount and Steitz, 1981; Guthrie and Patterson, 1988) and snRNP proteins (Mancebo *et al.*, 1990), are highly conserved between *Drosophila* and humans. However, an earlier study of length distribution of *Drosophila* introns (Hawkins, 1988) found many *Drosophila* introns smaller than the minimum size for splicing in mammalian cells (Smith *et al.*, 1989). A cursory survey of *Drosophila* introns also appeared to indicate that a sizeable minority lack strongly pyrimidine-rich stretches adjacent to their 3' splice sites (see, for example, Falkenthal *et al.*, 1985; Bernstein *et al.*, 1986; Eveleth *et al.*, 1986). In addition, *in vitro* splicing results in *Drosophila* Kc cell and human HeLa cell nuclear extracts indicate that the sequence requirements of these two species differ (Siebel and Rio, 1989). In one case (Guo, Lo and Mount, 1992), a short (74 nucleotide) *Drosophila* intron was found to splice well in extracts from *Drosophila* cell nuclei, but showed no activity in extracts from human cells. Conversely, lengthened versions of the same intron (90 nucleotides) were observed to splice well in human extracts but were not active substrates for the *Drosophila* system.

We have undertaken a thorough examination of sequences from *Drosophila* introns and flanking exons using a number of computational methods, including information content analysis (Schneider *et al.*, 1986; Fields, 1990), two independent consensus-search methods, CONSENSUS (Stormo and Hartzell, 1989; Hertz *et al.*, 1990) and RTIDE (Galas *et al.*, 1985; Waterman and Jones, 1990), and simple assessment of the frequencies and distributions of subsequences from individual nucleotides to hexanucleotides. Our goal was to further

define the relevant consensus sequences and explore the possibility that there are size-dependent sequence features. We have confirmed the very sharp length distribution previously reported by Hawkins (1988). We have also observed that the highest information content is at the 5' and 3' splice sites, and that these sites differ in only minor ways between large and small introns and from vertebrate splice sites. However, large introns do tend to have a greater density of pyrimidines in the region upstream of the 3' splice site. A putative branchpoint consensus, C T A A T, differs from the mammalian signal primarily in the absence of G at the position immediately preceding the branchpoint. The distribution of occurrences of this sequence in *Drosophila* suggests a minimum distance between 5' splice sites and branchpoints of about 38 nucleotides and a minimum distance between 3' splice sites and branchpoints of about 15 nucleotides.

## DATABASE AND METHODS

### Data sets.

Database entries containing *Drosophila* nucleotide sequences were taken from the invertebrate division of Release 64.0 of GenBank (Burks *et al.*, 1990). This data set was manually edited to remove redundant copies of the same gene, tRNA-coding genes, and other questionable entries. Subsets of sequence data were automatically extracted from the collection of *Drosophila* entries using the program ExtractGenBank (R. Farber, Los Alamos National Laboratory), keying off the annotated IVS regions in the FEATURES table. These subset consisted of whole introns, 101 nucleotide-long windows centered on 5' splice sites, and 101

nucleotide-long windows centered on 3' splice sites. This procedure resulted in splice sites from introns whose sequences were determined in their entirety. The spans listed with the IVS annotation were used to calculate intron lengths. The data set was divided into short and long introns. 80 nucleotides was chosen as the cut-off value for this division after examining the distribution of lengths shown in Figure 1. This resulted in the data sets listed in Table 1. The number of large introns is greater than the number of 5' splice sites or 3' splice sites from large introns because of cases of alternative splicing in which the individual splice sites may be used to generate more than one intron.

### **Methodology.**

Frequency matrices and information content were determined as described in Fields (1990). Statistical uncertainties were calculated using the "exact" method described in Schneider *et al.*.

The identification of conserved sequences, with emphasis on definition of the branchpoint consensus, was performed with three different approaches. The RTIDE program (Galas *et al.*, 1985; Waterman and Jones, 1990) takes as input approximately aligned sequences and finds the most common "word", allowing for mismatches, within a specified "window" of sequence. The word length, window size and amount of allowed mismatch are all user specified parameters. The CONSENSUS program (Stormo and Hartzell, 1989; Hertz *et al.*, 1990) takes as input unaligned sequences and attempts to find an alignment that maximizes the information content of the identified sites. The output includes a "specificity matrix" which can be used by the program PATSER to find matches to the matrix in any sequence (Stormo,

1988,1990). Finally, occurrences of oligonucleotide sequences of various lengths within these data sets (or subsets of them), were determined.

## RESULTS AND DISCUSSION

### Size of introns

We first examined the distribution of sizes among the introns in our sample (Figure 1). As had been noted by Hawkins (1988), the majority of *Drosophila* introns are relatively small. In our data set, the median length is 79 nucleotides, and there is a sharp distribution of sizes around a modal length of approximately 63 nucleotides (Figure 1B). This is in marked contrast to the situation in mammals, where introns of less than 70 nucleotides are extremely rare and do not splice well in vitro (Fu *et al.*, 1988) or in vivo (Wieringa *et al.*, 1984). Our data sets were restricted to completely sequenced introns. This restriction introduces a bias against larger introns (Figure 1A). Many *Drosophila* introns larger than 6,000 nucleotides have been described, yet none have been completely sequenced. However, this bias should not greatly affect the apparent distribution of sizes within the set of small introns (Figure 1B). For example, the occurrence of 32 introns in the range of 61-65 nucleotides but only five introns in the range of 76-80 nucleotides cannot be explained by the difficulties associated with sequencing or reporting an additional 15 nucleotides. A similar distribution of intron sizes has been described for *C. elegans*, except that the modal size in the worm is approximately 50 nucleotides (Blumenthal & Thomas, 1988; Fields, 1990).

### Splice site consensus and information content.

A frequency matrix was prepared from 5' splice sites and 3' splice sites in each of the two size classes. These nucleotide frequencies are graphed in a number of ways in Figure 2, and values for nucleotides near the splice sites are reported in Table 2. We then applied the information content measure of Schneider *et al.* (1988; Fields, 1990), which reflects the extent of deviation from random base composition, to the data in the frequency matrices. The resulting distributions are shown in Figure 3.

A comparison of the nucleotide frequency distribution within 5' splice sites in *Drosophila* with all of those compiled by Senapathy *et al.* (1990), most of which are from vertebrates, shows considerable similarity (Table 2A). In each case the consensus MAG|GTRAGT is valid and position 5 is the most highly conserved nucleotide outside of the invariant GT. One minor difference between *Drosophila* and vertebrates is that position 6 is less variable in *Drosophila* (68%T) than it is in the total set (50%T). Short and long *Drosophila* introns are likewise similar. There are two cases in this data set of deviation from the rule that introns begin with GT. They are the *Gpdh* gene intron C, which begins with GC, in the sequence context AAGGCAAGT, and *rudimentary* intron E, which begins CT in GCGCTGAGA. The one case of deviation from the rule that introns end with AG is *perB* intron E, which ends CG. The phenomenon of rare nonconsensus sites has been described previously (Senapathy, Shapiro and Harris, 1990), but the frequency of exceptions is low in all data sets, and many apparent exceptions have proven to be errors in one way or another (Jackson, 1991). In each of the cases cited here, the exception was based on cDNA sequence information and the authors noted the exception in a refereed journal article.

As expected, there is a peak of information content at each splice site, corresponding to the traditional splice site consensus sequences, and the total information in both 5' and 3' splice sites from the large and small data sets are comparable. A large peak of information around position +40 of the short intron 5' splice sites probably corresponds to the branchpoint (see below).

### **Nucleotide frequencies.**

Both large and small introns are characterized by higher A+T content than exons. Base composition in the region between -51 to -42 relative to 3' splice sites and +21 and +30 relative to 5' splice sites (areas relatively devoid of information; see Figure 2) is 32% A, 19% C, 16% G and 33% T. This intron A+T content of 65% compares with an A+T content of 48% in the 50 nucleotide window of flanking exonic sequence contained in our database. This difference of 17% in A+T content between introns and exons is reminiscent of plant (particularly dicot) introns (Hanley and Schuler, 1988), in which high A+T content has been shown to be a signal for splice site recognition (Weibauer *et al.*, 1988; Goodall & Filipowicz, 1989,1991). More recently, Csank *et al.* (1990) have observed that high A+T content is a general property of introns in many species. Ironically, the two most highly studied groups, the yeast *Saccharomyces cerevisiae* and mammals are among the few exceptions.

Of particular interest is the pyrimidine-rich region associated with 3' splice sites, which has been shown to play a critical role in branchpoint recognition during mammalian splicing (Ruskin and Green, 1985; reviewed in Smith *et al.*, 1989). In our *Drosophila* data set, it appears that the pyrimidine-rich region of vertebrate introns is replaced by a large T-rich

region and a much smaller C-rich region. Examination of Figure 2 shows that the region between -30 and -10 relative to 3' splice sites of introns in both size classes is distinctive in that the frequency of A and G declines and the frequency of T increases as the 3' splice site is approached. Overall frequency of A+T does not change, but T increases to approximately 50%. In both large and small introns, the frequency of C is not significantly higher in the -30 to -10 region than it is in the -50 to -30 region, but C is very common at positions -9, -8 and -7. The position with maximal C content is -9 in the case of large introns (41%) and -7 in the case of small introns (44%). Interestingly, this C-rich region is in precisely the location of a peak of pyrimidine-richness in yeast introns (Parker and Patterson, 1987), where the pyrimidine stretch plays a less significant role than it does in vertebrates (Patterson and Guthrie, 1991). However, this small peak of C-richness is not seen in all *Drosophila* introns, and many introns lack C's in this region.

The most significant difference we have observed between large and small introns is that large introns have greater pyrimidine content near the 3' splice site (Figure 2D). This is particularly true in the -21 to -11 region, where large introns have a 10% higher pyrimidine content overall. Thus, only 27/107 (25%) of 3' splice sites from small introns have a stretch of 8 consecutive pyrimidines, while 50/98 (51%) of large introns do. Similarly, 55/107 (51%) of 3' splice sites from small introns have a stretch of 12 nucleotides with 10 or more pyrimidines, while 71/98 (72%) of large introns do. It is interesting to compare this result to that obtained by analysis of large and small introns in *C. elegans*. In each case, large introns were observed to carry more information. In the case of worms, that information was in the 5' splice site. In the case of flies, that information is in the pyrimidine stretch.

A number of *Drosophila* introns lack strongly pyrimidine-rich regions near the 3' splice site. For example, the second *white* intron has less than 50% pyrimidines (14/31), and no stretch of 12 consecutive nucleotides with more than seven pyrimidines between the 3' splice site and the branchpoint (Guo, Lo and Mount, 1992). However, the generality of this observation is hard to assess from the database, because shorter stretches, which could play the functional role of a pyrimidine stretch, can usually be found (although not in greater numbers than would be expected by chance). In our data set, only seven of 204 3' splice sites lack a stretch of 8 nucleotides with 6 or more pyrimidines in the -51 to -3 region. One possibility suggested by our data is that other aspects of base composition serve as a component of the splicing signal (conceivably, this could be reflected in species differences in U2AF specificity). For example, G-poorness contributes more to the information content of this region than does pyrimidine-richness, with the percentage of G residues being only 8.3% in the region between -5 and -17 (see Figure 2 and Figure 3).

#### **Branchpoint consensus.**

Because splice sites are readily identified by comparing genomic and cDNA sequences, the consensus sequences at the splice sites are easily identified and tabulated (see Table 2). However, definitive localization of a branchpoint requires analysis of splicing intermediates, or excised introns, which are generally obtained only through *in vitro* splicing. Only two wild type *Drosophila* introns have been analyzed in this way (see Table 3), and considerable experimental work will be required to establish a statistically meaningful data set of confirmed branchpoint sequences. However, computational analyses can yield information about the

branchpoint consensus. For example, a previous analysis using the consensus search program known as Makecons with a data set of 24 introns identified C T A A T as a putative *Drosophila* branchpoint consensus (Keller and Noon, 1984). We therefore applied a variety of computational methods to the problem of identifying signals within introns or within the flanking exons. Unlike the study of Keller and Noon (1984), which used a technique for refinement of a given consensus matrix, each of these methods involved an unbiased search for a consensus.

Initially, we analyzed our data set with the RTIDE program (Galas *et al.*, 1985; Waterman and Jones, 1990) and the CONSENSUS program (Stormo and Hartzell, 1989; Hertz *et al.*, 1990). The RTIDE method takes as input approximately aligned sequences and reports the most frequent "word" in a window of variable size, allowing for mismatches between the words. In addition, the frequency of the most frequent word is an additional measure of information content. We examined data sets consisting of sequences 101 nucleotides long centered on splice sites from large, medium and small introns (see Table 1). Word length was varied from 4 to 6, and window size from 8 to 15, allowing 0, 1 or 2 mismatches, with consistent results. The distribution of numerical scores that resulted when a window of eight nucleotides was used to look for 5 letter words, allowing one mismatch, is shown in Figure 4. As expected, these distributions are similar to those resulting from information-content calculations (Figure 3). The most frequent "word" at the 5' splice site, GTAAG for small (and all) introns and GGTAAG for large introns, agrees well with the consensus data above, and supports the observation that small introns show slightly more conservation in the intron portion of the 9 nucleotide consensus while large introns show

slightly more conservation in the exon portion of the consensus. At 3' splice sites, scores are lower, and the highest scoring words have the form TNCAG. The intron region adjacent to 3' splice sites (roughly -10 to -30) shows much higher scores than intron sequences in general. Analysis of the words that contribute these high scores is indicative of signals for the branchpoint and pyrimidine stretch. The highest scoring word for windows centered between -21 and -26 resembles the branchpoint consensus in each case (TAATT, TTAAT or CTAAT are observed). Windows centered between -8 and -15 have words consisting of entirely pyrimidines (TTTCT, TCTTT, TCCTT and CTTTC), and TACAG is the highest scoring word for each of the four eight-nucleotide windows that include the -5 to -1 region. In the region between the pyrimidine and branchpoint regions, words such as TTTAT, which can become either branchpoint (TTAAT) or pyrimidine stretch (TTTTT) with the variation of single nucleotide, score highest.

Unlike RTIDE, the program CONSENSUS looks for consensus patterns within sequences without regard to any pre-existing alignment. Rather, the matrix with maximal information content derived using one contribution from each of the sequences is determined. As expected, the 5' splice site data sets yielded a matrix that was essentially identical to the 5' splice site consensus. The 3' splice site data sets yielded matrices that were A+T rich, pyrimidine rich, resembled the branchpoint, or had some combination of these features. In order to look for a branchpoint consensus matrix without interference from other sequence elements (such as the pyrimidine stretch or splice site), a modified data set consisting of the 42-nucleotide region between -51 to -10 was used. The program was run with a window size of 8, using an order-independent algorithm, and saving a maximum of 1000 matrices per

cycle. The matrix initially derived by the CONSENSUS program was used to rescan the data and generate a new alignment. Information content was calculated as in Hertz *et al.* (1990) using as an *a priori* base probability the actual values from the modified data set (A=0.318, C=0.181, G=0.133 and T=0.368). The rescanning procedure was performed six times before the alignment became self-generating; the improvement in information content was 0.0795 bits. This pattern includes one position with very little information, and can be thought of as a seven nucleotide matrix beginning with position 2, roughly corresponding to the consensus WCTAATY (see Table 3). Because of its similarity to the branchpoint consensus sequence described in other organisms, it is extremely likely that this sequence is the *Drosophila* branchpoint consensus. However, this matrix is derived from the set of best matches to itself (through a collection of sequences) and it is different in a number of ways, most notable among them being the lack of G in the position immediately preceding the branchpoint (see Table 3).

In order to investigate the branchpoint consensus further, the distribution (relative to splice sites) of a large number of tetranucleotides was examined. These distributions were generally in agreement with the consensus sequence determined as described above. Frequency data for CTAA and CTGA are presented in Figure 5. As expected from its similarity to the branchpoint consensus, CTAA occurs frequently in the region between -50 and -18 relative to 3' splice sites. These distributions match those expected if CTAA were indeed a common tetranucleotide at *Drosophila* branchpoints, with larger introns showing a tendency for branchpoints slightly further upstream. The two occurrences of CTAA at -18 and three at -20 indicate use of a branchpoint A at -15 in two cases and -17 in three others. This suggests a

minimum distance between the branchpoint and 3' splice site of either 15 or 17 nucleotides, a number that is in reasonably good accord with results from mammalian systems, where a minimum distance of 18 nucleotides is indicated by the available data (Fu *et al.*, 1988b; Nelson and Green, 1989).

The distribution of CTGA tetranucleotides relative to 3' splice sites stands in sharp contrast to that of CTAA tetranucleotides. In fact, no significant enrichment of CTGA is observed in the branchpoint region. This is consistent with the consensus sequence derived above, and indicates a difference between *Drosophila* and mammals, where G is the predominant nucleotide in this position (Nelson and Green, 1989; Table 3).

The distribution of CTAA relative to 5' splice sites in small introns can be used as an indication of the minimal distance between 5' splice sites and branchpoints. In this case, there are two occurrences of CTAA at position 35 and seven at position 36, indicating possible use of a branchpoint A at position 38 in two cases and at position 39 in seven others. In fact, there are 27 occurrences of CTAA that would indicate branchpoint A's between position 39 and position 43 in this data set of 107 sequences, clearly indicating a very strong tendency for branchpoints to occur within this narrow range. Further support for this observation is found in the peak of information at this position in short introns (see Figure 3). This 5' splice site to branchpoint distance is considerably less than that in mammalian introns. For example, manipulation of the 5' splice site to branchpoint distance of the 66 nucleotide small-t intron (Fu and Manley, 1988) indicated that the wild type distance in that case (48 nucleotides) is minimal; an intron with a distance of 46 nucleotides showed no splicing, while a distance of 53 nucleotides showed significantly increased splicing. In a study of the  $\alpha$ -tropomyosin gene,

Smith and Nadal-Ginard (1989) found a 5' splice site to branchpoint distance of 51 nucleotides too short, but 59 sufficient, for *in vitro* splicing. In addition, a distance of 49 nucleotides between the 5' splice site and branchpoint was found too short to allow U4-U5-U6 binding to an adenovirus E1A pre-mRNA *in vitro* (Himmelspach *et al.*, 1991).

Other tetranucleotides from the branchpoint consensus (TAAT, TAAC, AACT and AATC) are distributed similarly to CTAA (data not shown). Thus, the distribution of branchpoint-related sequences indicates that the smaller size of *Drosophila* introns is associated with a decreased distance between 5' splice sites and branchpoints rather than a decreased distance between branchpoints and 3' splice sites.

### **Mechanistic implications**

Several of the results obtained here make it interesting to speculate that small and large introns may differ with respect to the mechanism by which branchpoints are recognized. First, our observations on intron size and the distribution of branchpoint-like sequences in *Drosophila* argue that mammals and fruit flies differ considerably with regard to the range of acceptable distances between the 5' splice site and the branchpoint. There is experimental support for this conclusion. For example, the second *Drosophila white* intron (74 nucleotides) has a 5' splice site to branchpoint distance of 43 nucleotides (less than the mammalian minimum) and is efficiently spliced in nuclear extracts from *Drosophila*, but not human, cells (Gou, Lo and Mount; 1992). It is likely that this minimum distance reflects the spatial requirements for spliceosome assembly in mammals, and indeed, Himmelspach *et al.* (1991) have observed defective spliceosome assembly on experimentally shortened introns. However,

because the size of U RNAs (Mount and Steitz, 1981) and snRNP proteins (Paterson *et al.*, 1991) is comparable in *Drosophila* and mammals, it seems unlikely to us that the shorter minimum distance could be explained by flies simply having smaller snRNPs. Second, it is striking that the separation between the 5' splice site and branchpoint in many short *Drosophila* introns is just over than the minimum distance. These observations indicate that branchpoint recognition in small introns may be facilitated by direct interaction between a factor at the 5' splice site (the U1 snRNP, a larger complex including the U1 snRNP, or some other factor) and a factor at the branchpoint (the U2 snRNP, a larger complex including the U2 snRNP, or some other factor). Normally, association between the U2 snRNP and the branchpoint is promoted by the binding of U2AF to the pyrimidine stretch. Thus, the reduced pyrimidine content of small *Drosophila* introns relative to large *Drosophila* introns (which are more like mammalian introns in their pyrimidine content) also supports the hypothesis that branchpoints in small introns might be recognized by a mechanism involving the 5' splice site. Such a mechanism of branchpoint recognition may be corroborated by the observation that, in yeast, the formation of early complexes including U1 snRNP has been observed to depend upon the branchpoint (S raphin and Rosbash, 1991). We (SM, unpublished results) have begun to explore these ideas experimentally.

#### ACKNOWLEDGEMENTS

We are grateful to G. Hartzell for systems support, to R. Farber for providing us with an updated version of the ExtractGenBank software, and to Nicole Kawachi for assistance with

tetranucleotide analysis. S.M. was supported by NIH grant GM 37991, by a NSF Presidential Young Investigator award, and by Basil O'Conner Starter Scholar Research award 5-630 from the March of Dimes Birth Defects Foundation. G.S. and G.H. were supported by NIH grants GM 28755 and HG 00249, O.W. and C.F. were supported by U.S. Department of Energy Genome Program Grant 89ER60865, and C.B was supported by NIH grant GM 37812. This work was in part done under the auspices of the Aspen Center for Physics under a grant from the NSF.

#### FIGURES LEGENDS

Figure 1. Size distribution of Drosophila introns. A. All introns with lengths less than 6,000 nucleotides are included, and the number of examples in each bin of 100 nucleotides is plotted. B. The distribution of sizes among introns of less than 250 nucleotides are plotted with a bin size of 3.

Figure 2. Nucleotide frequencies. A+T, C + T and G content is plotted for the 100 nucleotide window around each type of splice site (A: 5' splice sites; B: 3' splice sites). C: The frequency of A+T across 5' and 3' splice sites are superimposed to emphasize the uniformity of exonic and intronic A+T content. D: The frequency of pyrimidines is separately plotted for the intronic region adjacent to 3' splice sites in large (thin line) and small (thick line) introns.

Figure 3. Information content (Schneider *et al.*, 1986; Fields *et al.*, 1990) at nucleotide

positions between -50 and +50 surrounding the 5' splice sites (A and B) and 3' splice sites (C and D) of long (A and C) and short (B and D) *Drosophila melanogaster* introns.

Figure 4. RTIDE analysis of *Drosophila* introns. Scores from the program RTIDE (Galas *et al.*, 1985; Waterman and Jones, 1990) using 5 nucleotide words in an 8 nucleotide-long window are plotted vs. position, and the highest scoring "word" corresponding to each peak is indicated.

Figure 5. Occurrences of selected tetramers relative to splice sites. All occurrences of the indicated tetranucleotide in all introns within the indicated database are plotted relative to the splice site. See text.

## REFERENCES

- Barabino, S. L., B. J. Blencowe, U. Ryder, B. S. Sproat and A. I. Lamond. 1990. Targeted snRNP depletion reveals an additional role for mammalian U1 snRNP in spliceosome assembly. *Cell* 63: 293-302.
- Bernstein, S. I., C. J. Hansen, K. D. Becker, D. R. Wassenberg, E. S. Roche, J. J. Donady and C. P. Emerson. 1986. Alternative RNA splicing generates transcripts encoding a thorax-specific isoform of *Drosophila melanogaster* myosin heavy chain. *Mol. Cell. Biol.* 6:2511-2519.
- Black, D. L., B. Chabot and J. A. Steitz. 1985. U2 as well as U1 small nuclear ribonucleoproteins are involved in pre-messenger RNA splicing. *Cell* 42: 737-750.
- Blumenthal, T. and J. Thomas. 1988. Cis and trans mRNA splicing in *C. elegans*. *Trends Genet.* 4:305-308.
- Bruzik, J. P. and J. A. Steitz. 1990. Spliced leader RNA sequences can substitute for the essential 5' end of U1 RNA during splicing in a mammalian in vitro system. *Cell* 62: 889-899.
- Burks, C., Cinkosky, M.J., Gilna, P., Hayden, J.E.-D., Abe, Y., Atencio, E.J., Barnhouse, S., Benton, D., Buenafe, C.A., Cumella, K.E., Davison, D.B., Emmert, D.B., Faulkner, M.J., Fickett, J.W., Fischer, W.M., Good, M. Home, D.A., Houghton, F.K., Kelkar, P.M., Kelley, T.A., Kelly, M., King, M.A., Langan, B.J., Lauer, J.T., Lopez, N., Lynch, C., Lynch, J., Marchi, J.B., Marr, T.G., Martinez, F.A., McLeod, M.J., Medvick, P.A., Mishra, S.K., Moore, J., Munk, C.A., Mondragon, S.M., Nasser, K.K., Nelson, D., Nelson, W., Nguyen, T., Reiss, G., Rice, J., Ryals, J., Salazar, M.D., Stelts, S.R., Trujillo, B.L., Tomlinson, L.J., Weiner, M.G., Welch, F.J., Wiig, S.E., Yudin, K., and Zins, L.B. (1990) GenBank: Current status and future directions. *Meth. Enzymol.*, 183: 3-22.
- Csank, C., F. M. Taylor and D. W. Martindale. 1990. Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes.

- Nucleic Acids Res. 18: 5133-5141.
- Eveleth, D. D., R. D. Gietz, C. A. Spencer, F. E. Nargang, R. B. Hodgetts and J. L. Marsh. 1986. Sequence and structure of the dopa decarboxylase gene of *Drosophila*: evidence for novel RNA splicing variants. *EMBO J.* 5: 2663-2672.
- Falkenthal, S., V. P. Parker and N. Davidson. 1985. Developmental variation in the splicing pattern of transcripts from the *Drosophila* gene encoding myosin alkali light chain result in different carboxyl-terminal amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* 82:449-453.
- Fields, C. (1990) Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucleic Acids Res.* 18:1509-1512.
- Fu, X.-Y., J. Colgan and J. L. Manley. 1988. Multiple cis-acting sequence elements are required for efficient splicing of simian virus 40 small-t antigen pre-mRNA. *Mol. Cell. Biol.* 8: 3582-3590.
- Galas, D. J., M. S. Waterman and M. Eggert. 1985. *J. Mol. Biol.* 186: 117-128.
- Goodall, G. J. and W. Filipowicz. 1989. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell* 58:473-483.
- Goodall, G. J. and W. Filipowicz. 1991. Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *EMBO J.* 10:2635-2644.
- Green, M. R. 1986. Pre-mRNA splicing. *Annu. Rev. Genet.* 20: 671-708.
- Green, M. R. 1991. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu. Rev. Cell Biol.* 7: 559-600.
- Guo, M., P. Lo and S. Mount. 1992. Species-specific signals for the splicing of a short *Drosophila* intron in vitro. submitted.
- Guthrie. 1991. Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. *Science* 253:157-163.
- Guthrie and Patterson. 1988. Spliceosomal snRNAs. *Ann. Rev. Genetics* 22: 387-419.
- Hanley and Schuler. 1988. Plant intron sequences: evidence for distinct groups of introns. *Nucleic Acids Res.* 16:7159-7176.
- Harris, N. L. and P. Senapathy. 1990. Distribution and consensus of branchpoint signals in

- eukaryotic genes: a computerized statistical analysis. *Nucleic Acids Res.* 18: 3015-3019.
- Hawkins, J. D. 1988. *Nucleic Acids Res.* 16:9893-9905.
- Hertz, G. Z., G. W. Hartzell, III and G. D. Stormo. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 6:81-92.
- Hess, N. K. and S. I. Bernstein. 1991. Developmentally regulated alternative splicing of *Drosophila* myosin heavy chain transcripts: *in vivo* analysis of an unusual 3' splice site. *Dev. Biol.* 146: 339-344.
- Himmelsbach, M., R. Gattoni, C. Gerst, K. Chebli and J. Stevenin. 1991. Differential block of U small nuclear ribonucleoprotein particle interactions during *in vitro* splicing of adenovirus E1A transcripts containing abnormally short introns. *Mol. Cell. Biol.* 11: 1258-1269.
- Jackson, I. J. 1991. A reappraisal of non-consensus splice sites. *Nucleic Acids Res.* 19: 3795-3798.
- Jacob, M. and H. Gallinaro. 1989. The 5' splice site: phylogenetic evolution and variable geometry of association with U1 RNA. *Nucleic Acids Res.* 17: 2159-2180.
- Kay, R. J., R. H. Russnak, D. Jones, C. Mathias and P. Candido. 1987. *Nucleic Acids Res.* 9: 3723-3741.
- Keller, E. B. and W. A. Noon. 1984. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. *Proc. Natl. Acad. Sci. USA* 81:7417-7420.
- Mancebo, R., P. C. H. Lo and S. M. Mount. 1990. Structure and expression of the *Drosophila melanogaster* gene for the U1 small nuclear ribonucleoprotein particle 70K protein. *Mol. Cell. Biol.* 10: 2492-2502.
- Mount, S. M. and J. A. Steitz. 1981. Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing. *Nucleic Acids Res.* 9: 6351-6368.
- Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* 10: 459-472.
- Nelson, K. K. and M. R. Green, 1989. Mammalian U2 snRNP has a sequence-specific RNA-binding activity. *Genes Dev.* 3: 1562-1571.

- Newman and Norman. 1991. Mutations in yeast U5 snRNA alter the specificity of 5' splice site cleavage. *Cell* 65:115-123.
- Ogg, S. C., P. Anderson and M. P. Wickens. 1990. Splicing of a *C. elegans* myosin pre-mRNA in a human nuclear extract. *Nucleic Acids Res.* 18:143-149.
- Paterson, T., Beggs, J., Finnegan, D. and R. Luhrmann. 1991. Polypeptide components of *Drosophila* small nuclear ribonucleoprotein particles. *Nucleic Acids Res.* 19:5877-5882.
- Rio, D. C. 1988. *Proc. Natl. Acad. Sci. U.S.A.* 85: 2904-2909.
- Robberson, B. L., G. L. Cote and Susan M. Berget. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* 10: 84-94.
- Rosbash, M. and Seraphin, B. 1991. Who's on first? The U1 snRNP-5' splice site interaction and splicing. *Trends Biochem. Sci.* 16: 187-190.
- Ruskin, B., P. D. Zamore and M. R. Green. 1988. A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell* 52: 207-219.
- Schneider, T. D., G. D. Stormo, L. Gold and A. Ehrenfeucht. 1986. *J. Mol. Biol.* 188:415-431.
- Senapathy, P., M. B. Shapiro and N. L. Harris. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to the human genome project. *Methods Enzymol.* 183: 252-278.
- Seraphin, B. and M. Rosbash. 1990. Exon mutations uncouple 5' splice site selection from U1 snRNA pairing. *Cell* 63: 619-629.
- Shapiro, M. B. and P. Senapathy. 1986. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15: 7155-7174.
- Siebel, C. W. and D. C. Rio. 1989. Regulated splicing of the *Drosophila* P transposable element third intron in vitro: somatic repression. *Science* 248:1200-1208.
- Smith, C. W. J. and B. Nadal-Ginard. 1989. Mutually exclusive splicing of  $\alpha$ -tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell* 56: 749-758.
- Smith, C. W. J., J. G. Patton, and B. Nadal-Ginard. 1989. Alternative splicing in the control of gene expression. *Annu. Rev. Genet.* 23: 527-577.

- Stormo, G. D. 1988. Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Biophys. Chem.* 17: 241-263.
- Stormo, G. D. 1989. Consensus patterns in DNA. *Methods Enzymol.* 183: 211-221.
- Stormo and Hartzell. 1989. *Proc. Natl. Acad. Sci. U.S.A.*: 86:1183-1187.
- Talerico, M. and S. M. Berget. 1990. Effect of 5' splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol.* 10: 6299-6305.
- Waterman, M. S. and R. Jones. 1990. Consensus methods for DNA and protein sequence alignment. *Methods Enzymol.* 183: 221-237.
- Weibauer, K., J.-J. Herrero and W. Filipowicz. 1988. *Mol. Cell. Biol.* 8:2042-2051.
- Wieringa, B., E. Hofer and C. Weissman. 1984. A minimal intron length, but no specific internal sequence is required for splicing the large rabbit  $\beta$ -globin intron. *Cell* 37: 915-925.
- Wu, J. and J. L. Manley. 1989. Mammalian pre-mRNA branch site selection by U2 snRNP involves base pairing. *Genes Dev.* 3: 1553-1561.
- Zapp, M. L. and S. M. Berget. 1989. Evidence for nuclear factors involved in the recognition of 5' splice sites. *Nucleic Acids Res.* 17: 2655-2674.
- Zhuang, Y. and A. M. Weiner. 1986. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* 46: 827-835.
- Zhuang, Y. and A. M. Weiner. 1989. A compensatory base change in human U2 snRNA can suppress a branch site mutation. *Genes Dev.* 3: 1545-1552.

Table 1: Drosophila intron data sets

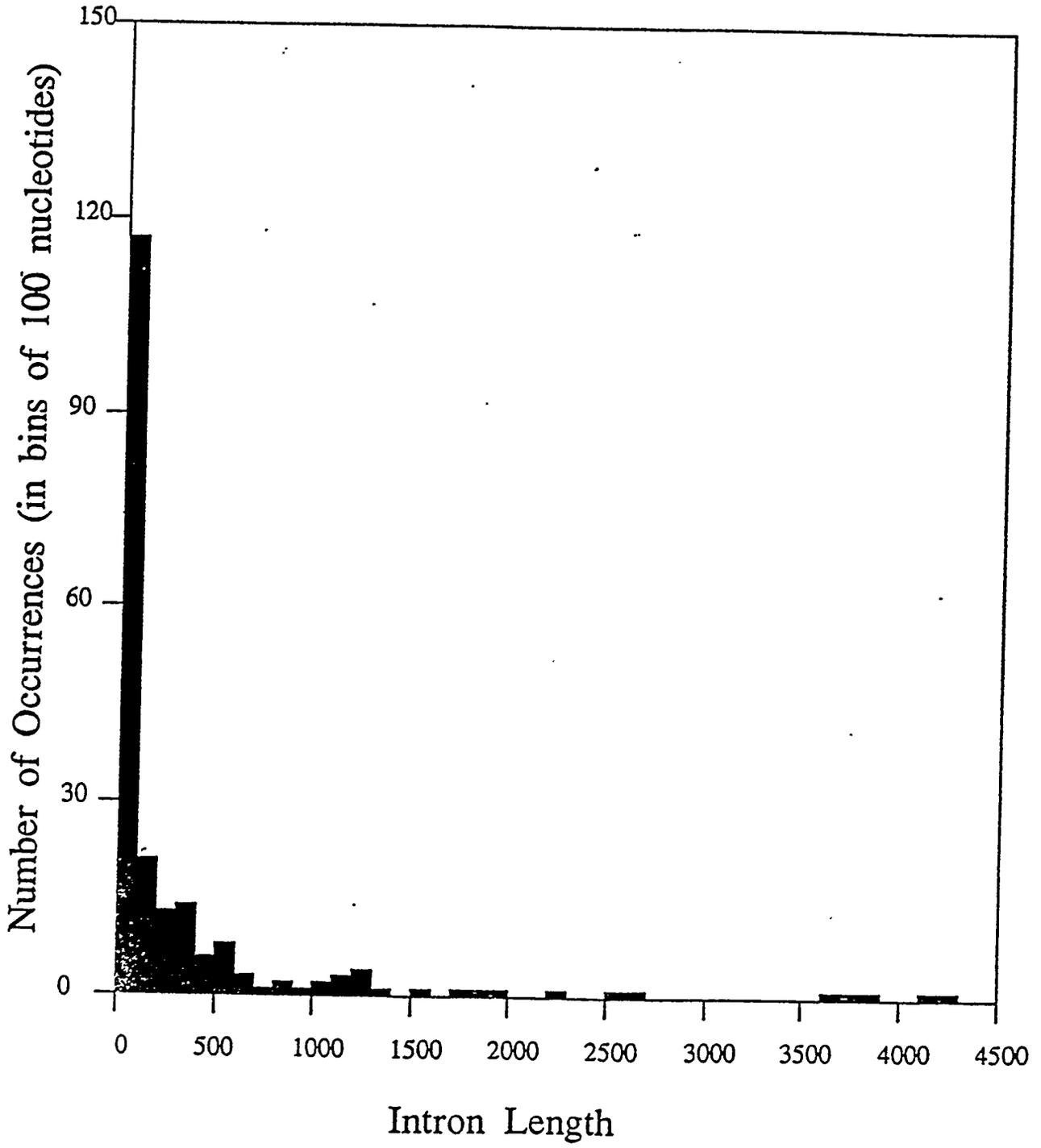
category	parent intron sizes	number of examples
complete introns	small (51-80)	107
	large (81-5392)	102
5' splice sites	small (51-80)	107
	large (81-5392)	99
3' splice sites	small (51-80)	107
	large (81-5392)	98

Table 3: Branchpoint sequences

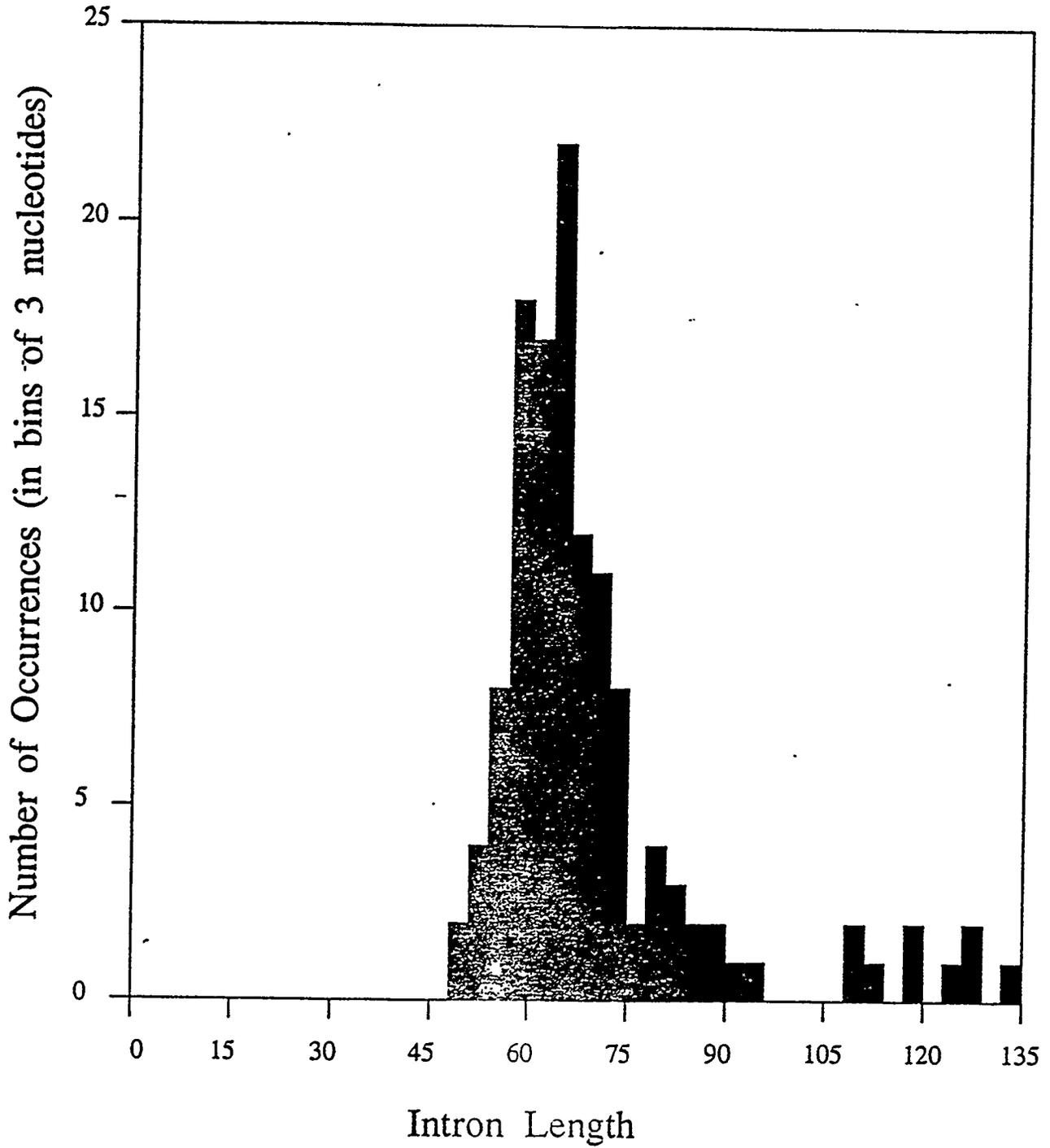
position	a	c	g	t	information
1	87	38	17	65	0.05056
2	81	0	43	83	0.32832
3	0	148	56	3	1.55082
4	50	51	5	101	0.13923
5	152	36	19	0	0.85662
6	183	16	0	8	1.13131
7	33	24	0	150	0.46963
8	18	77	40	72	0.26587

Figure 6. The "mammalian examples" matrix comes from a published compilation of experimentally determined branchpoints (Nelson and Green, 1989). The sequence TACTAAC is nearly universal in *Saccharomyces cerevisiae* and has been shown to be a preferred sequence in mammalian systems (Zhuang *et al.*, 1989).

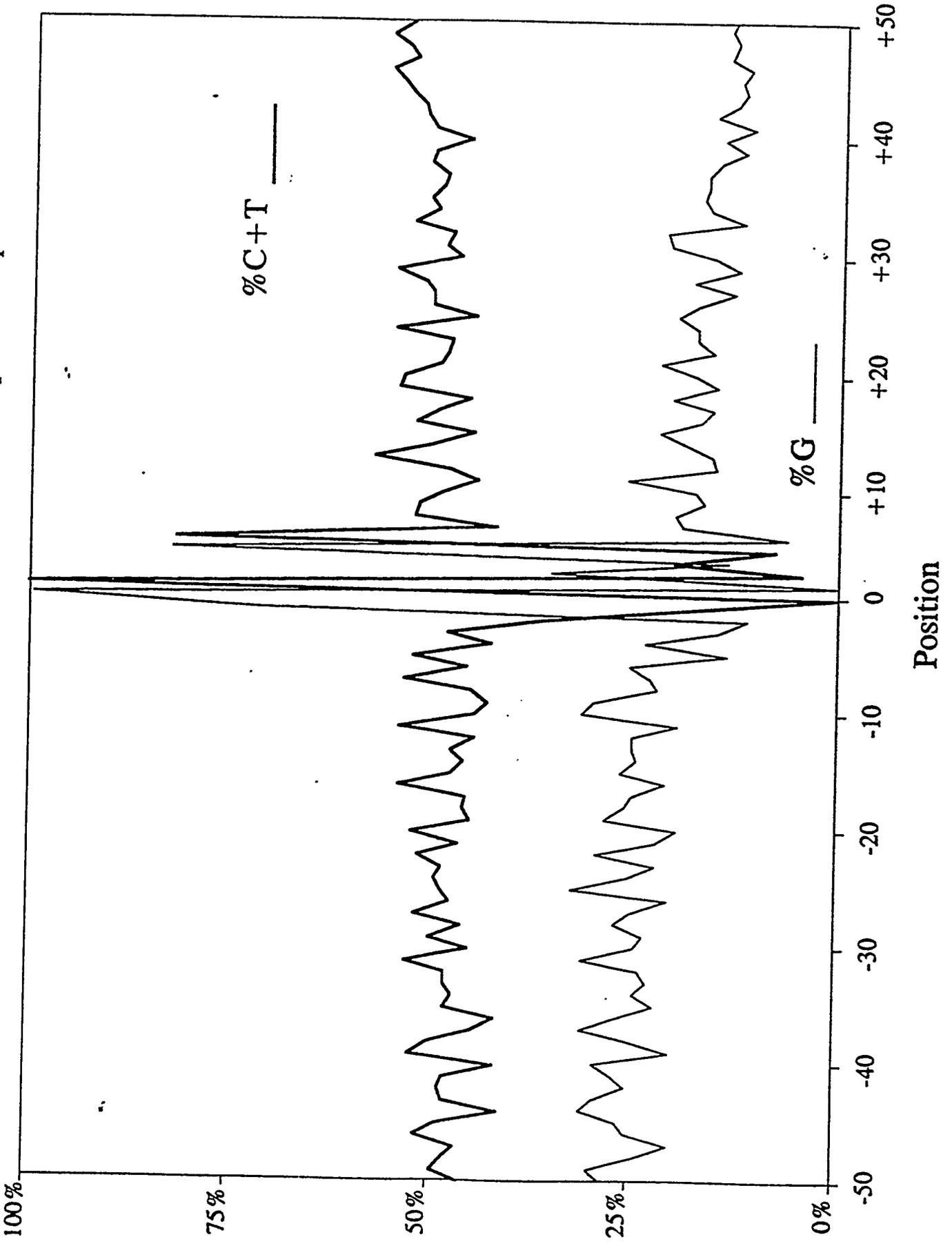
**A**



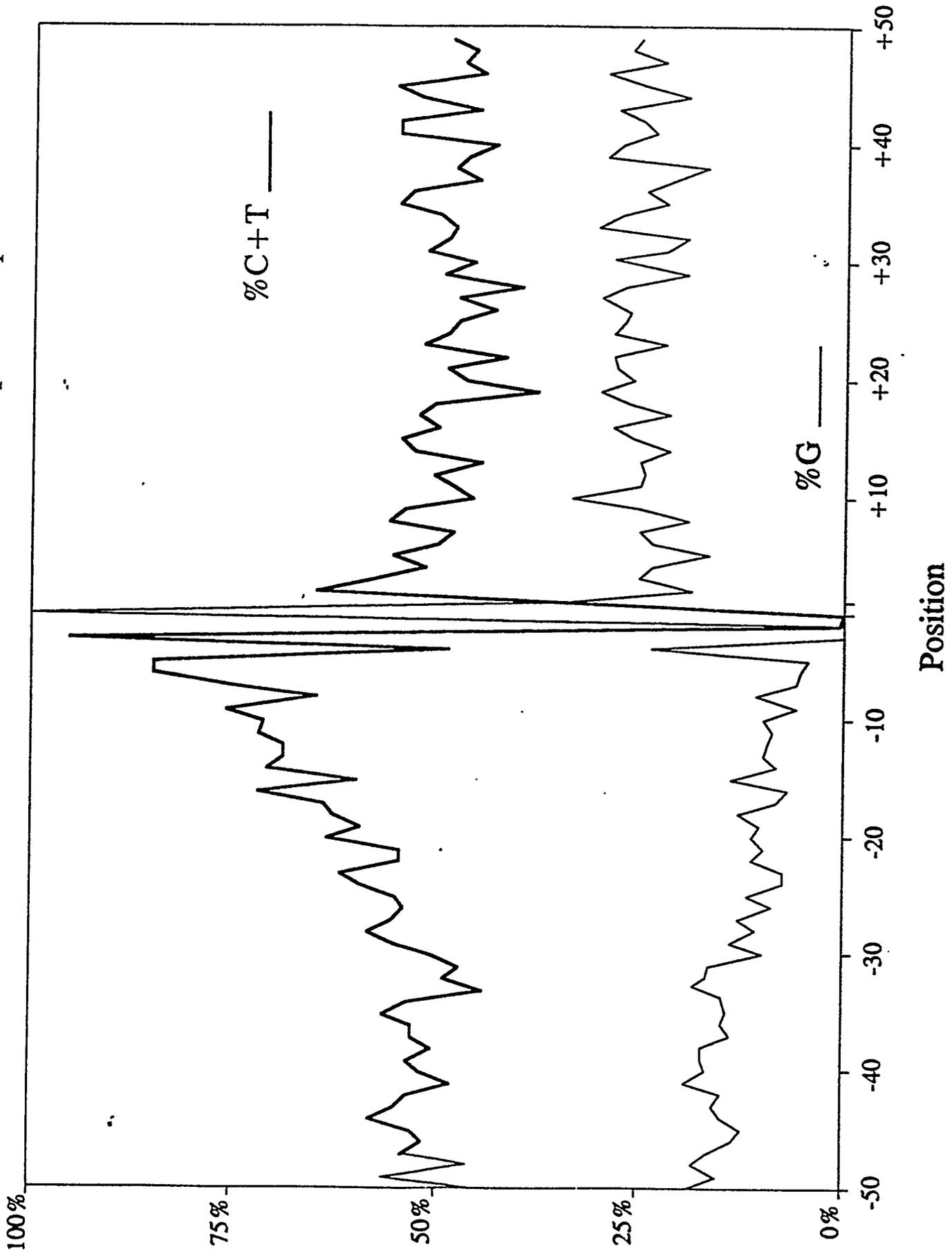
**B**



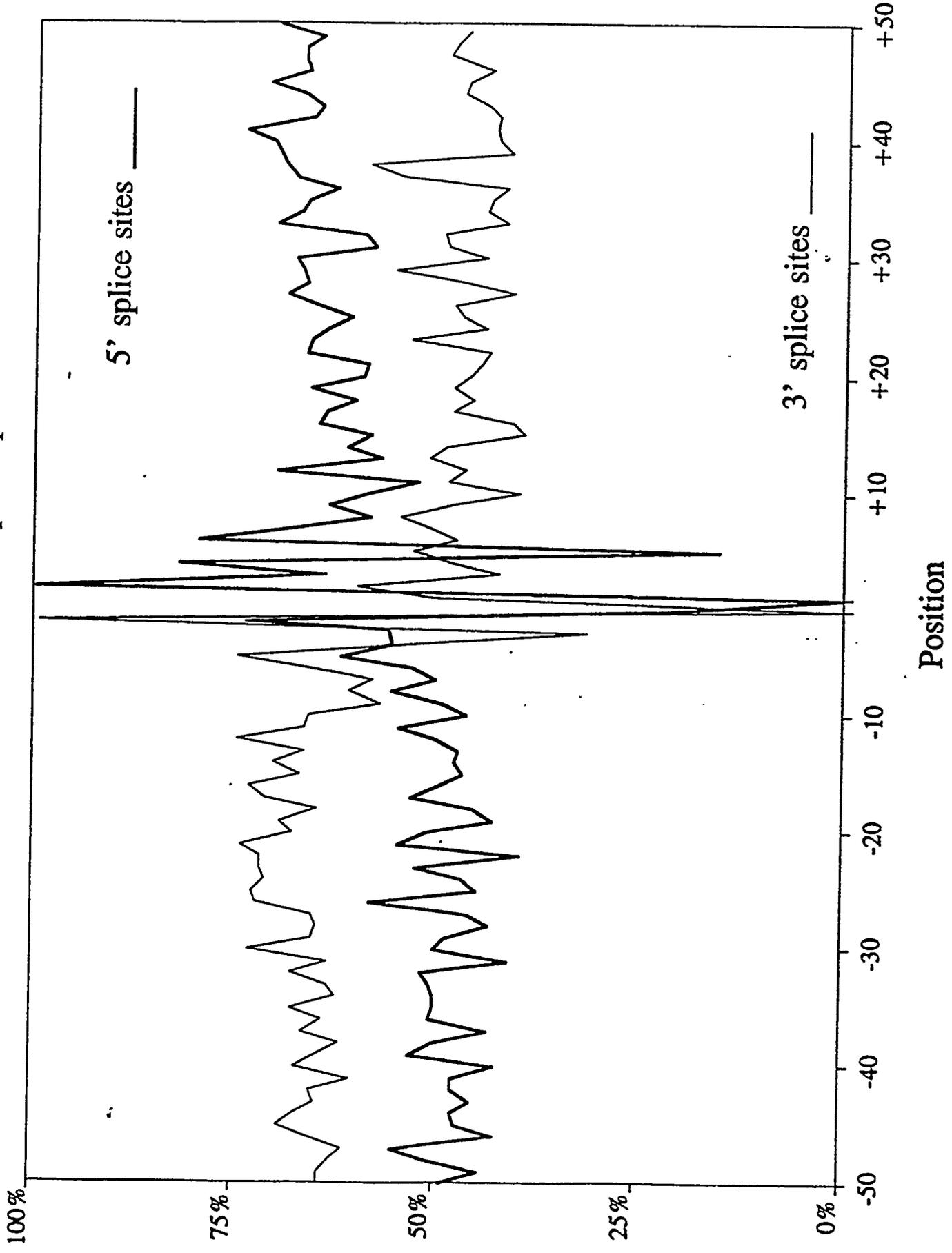
Base Composition (G and Pyrimidine content) - All Drosophila 5' splice sites



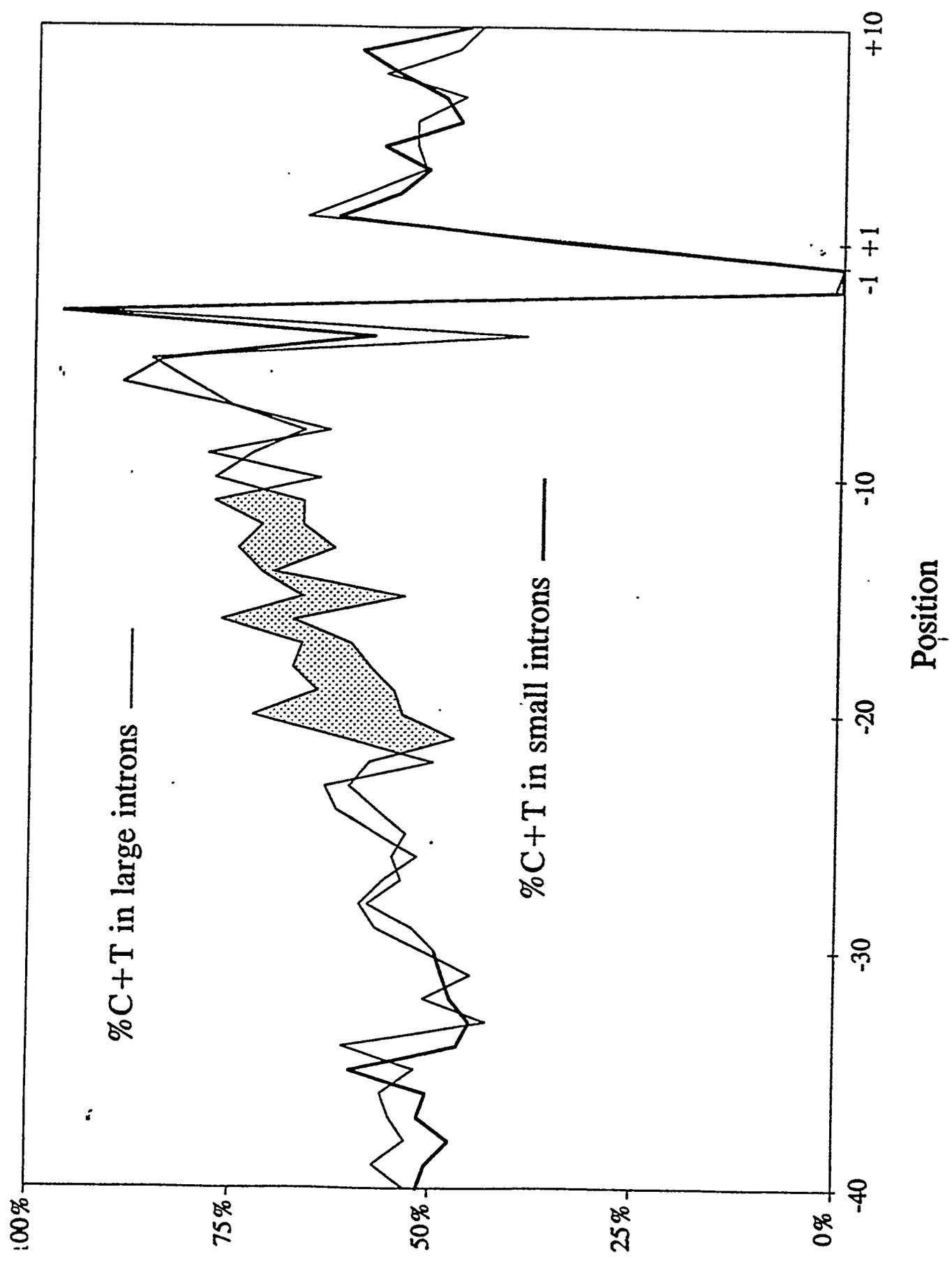
Base Composition (G and Pyrimidine content) - All Drosophila 3' splice sites



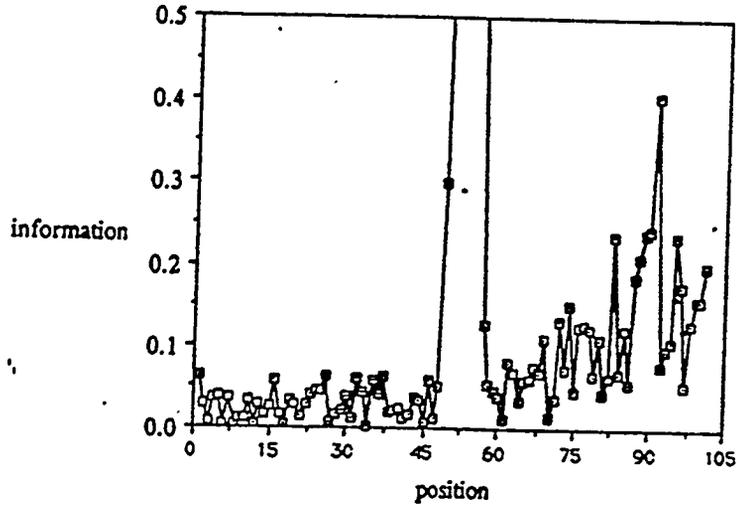
Base Composition (A+T content) across *Drosophila* splice sites



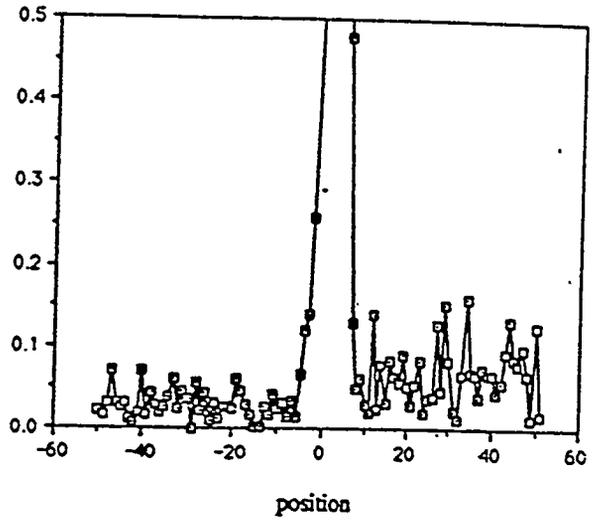
Pyrimidine content near Drosophila 3' splice sites - large introns vs. small introns



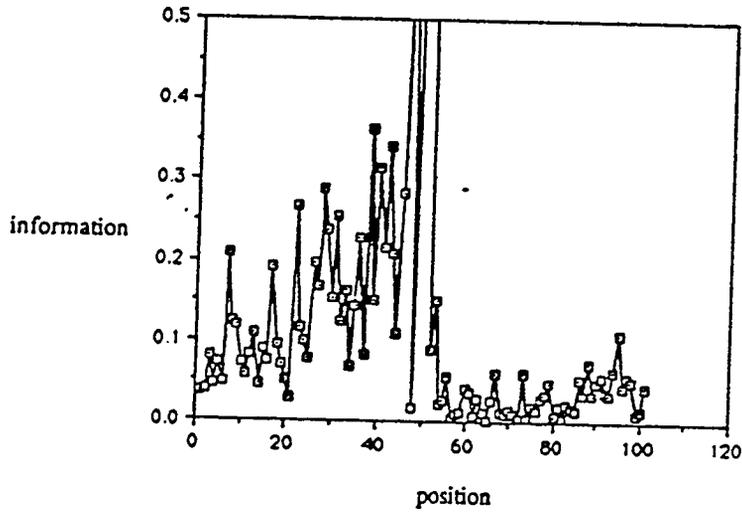
information from 5' splice sites  
- small introns



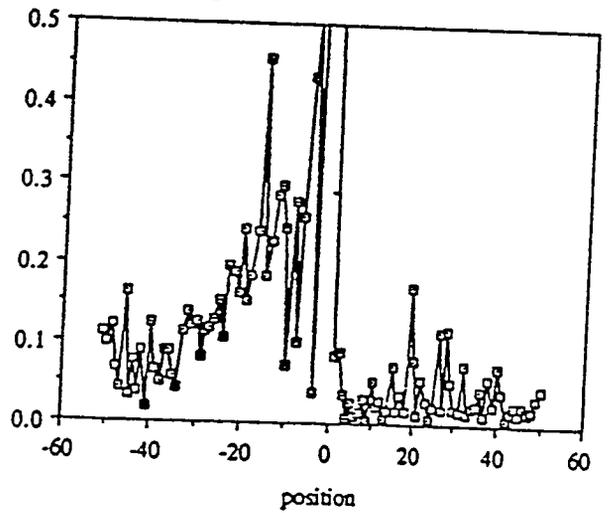
information around 5' splice sites  
- large introns

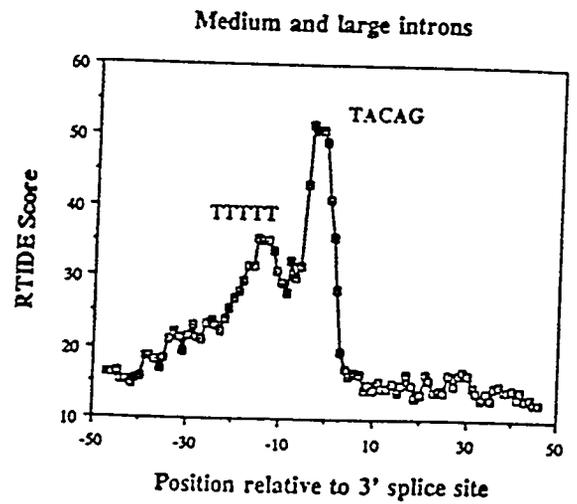
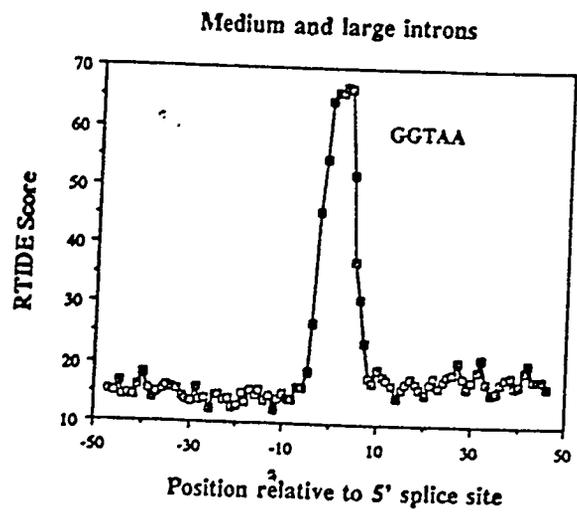
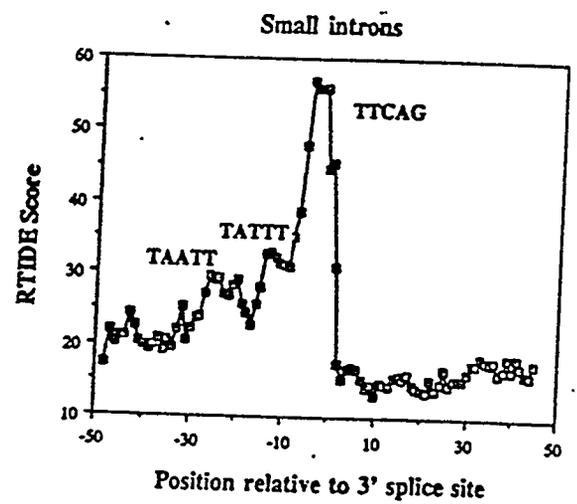
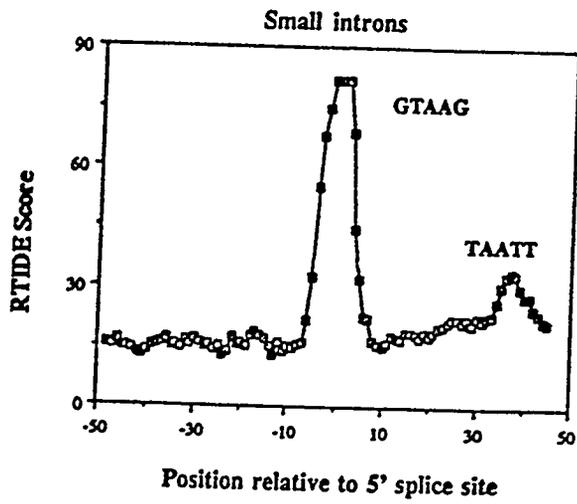
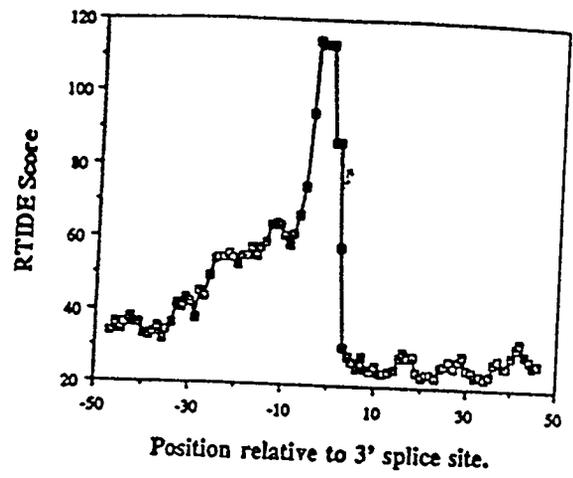
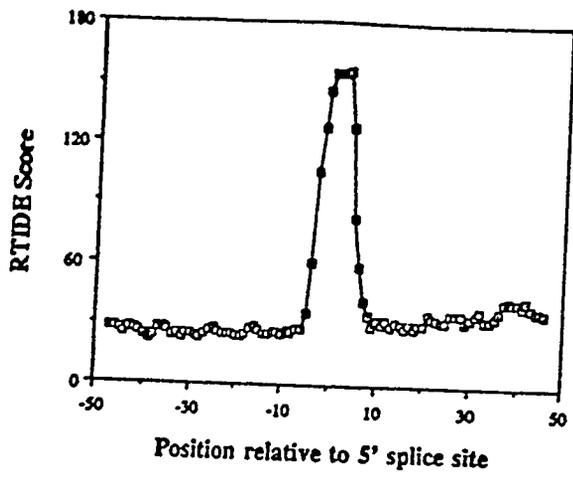


information from 3' splice sites  
- small introns

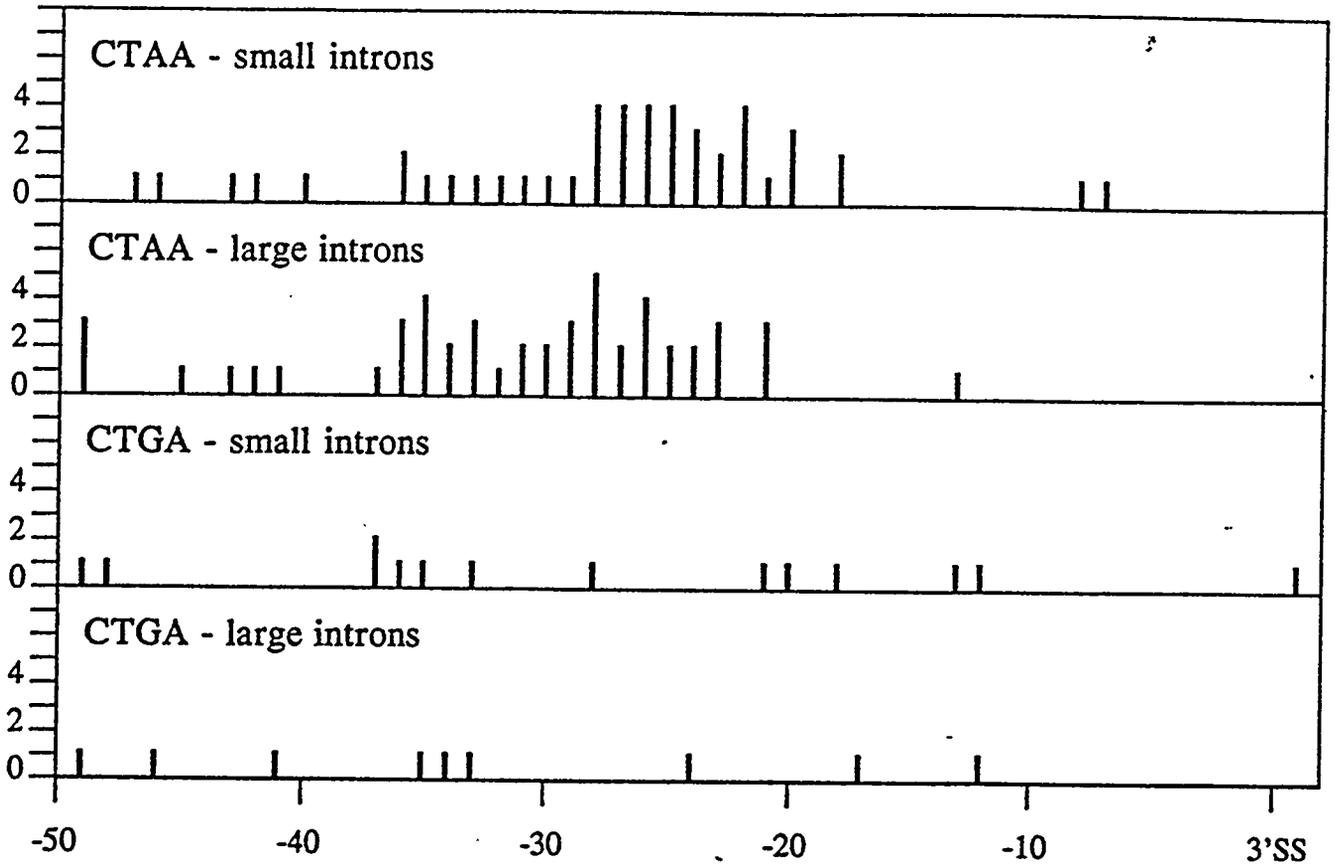


information around 3' splice sites  
- large introns

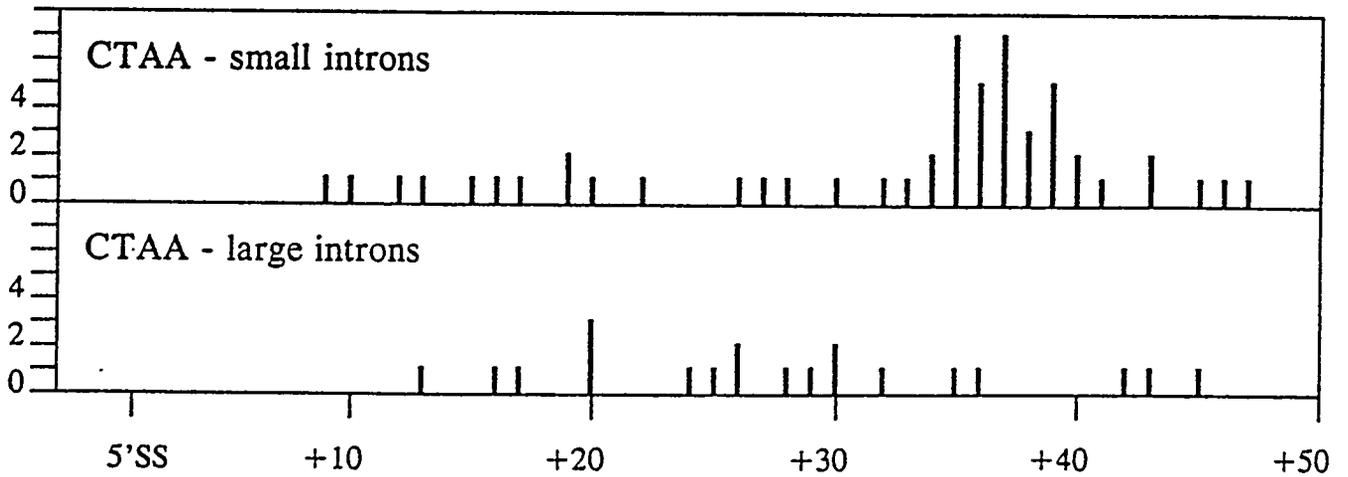




**A** Occurrences of CTAA and CTGA relative to 3' splice sites



**B** Occurrences of CTAA relative to 5' splice sites



5' splice site sequences:

*Drosophila* (all introns)

	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
A	33	34	37	52	9	0	0	60	71	9	11	39	27
C	24	21	29	15	8	0	0	1	9	2	14	13	21
G	14	23	15	11	71	100	0	35	9	82	6	19	20
T	29	22	19	21	12	0	100	4	11	6	68	29	32

consensus:                    M   A   G   G   T   R   A   G   T   W

Total (all organisms, from Senapathy et al., dominated by mammals)

	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
A			32	60	9	0	0	59	71	7	16		
C			37	13	5	0	0	3	9	6	16		
G			18	12	79	100	0	35	11	82	18		
T			13	15	7	0	100	3	9	6	50		

consensus:                    M   A   G   G   T   R   A   G   T

*Drosophila* (short introns)

	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
A	36	28	33	53	7	0	0	61	69	7	10	36	26
C	22	21	29	14	7	0	1	2	11	3	14	14	16
G	15	29	17	12	74	100	0	34	10	86	5	14	24
T	27	21	21	21	13	0	99	4	9	4	71	36	34

consensus:                            A   G   G   T   R   A   G   T   W

*Drosophila* (long introns)

	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8
A	30	39	41	51	12	0	0	60	72	11	12	42	27
C	26	21	29	16	9	1	0	0	8	3	15	12	28
G	12	17	13	10	69	99	0	36	7	78	8	24	15
T	31	22	16	22	10	0	100	4	13	8	65	21	29

consensus:                    M   A   G   G   T   R   A   G   T   A

### 3' splice site sequences:

#### Drosophila (all introns)

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	21	21	22	20	19	19	24	19	10	11	28	5	99	0	33	17	18
C	21	23	16	24	24	37	28	36	28	20	23	68	0	0	15	21	32
G	8	10	9	9	10	6	11	6	5	4	23	0	0	100	34	19	25
T	49	45	53	47	47	39	37	40	57	64	25	27	0	0	18	43	25
	T	T	T	T	T	Y	Y	Y	T	T		C	<u>A</u>	<u>G</u>	R	T	

#### Total (all organisms, from Senapathy et al., dominated by mammals)

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	11	11	10	8	11	10	11	11	7	8	25	3	100	0	27		
C	29	33	30	30	32	34	37	38	39	36	26	75	0	0	14		
G	14	12	10	10	9	11	10	9	7	6	26	1	0	100	49		
T	46	44	50	52	48	45	42	43	47	51	23	21	0	0	10		
	T	Y	Y	Y	Y	Y	Y	Y	Y	Y		C	<u>A</u>	<u>G</u>	G		

#### Drosophila (short introns)

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	25	28	24	23	16	20	23	19	7	13	23	4	100	0	36	19	21
C	17	22	11	19	22	32	32	42	32	16	27	64	0	0	16	17	35
G	5	9	9	10	7	7	10	6	5	3	19	0	0	100	30	19	23
T	53	40	55	48	55	41	35	34	57	68	31	33	0	0	19	46	21
	T	T	T	T	T	Y	Y	Y	Y	T		C	<u>A</u>	<u>G</u>	R	T	

#### Drosophila (long introns)

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3
A	16	15	19	15	22	17	26	18	13	8	33	6	99	0	30	14	13
C	27	24	20	32	27	43	24	29	23	26	19	73	1	0	14	26	30
G	12	10	9	7	13	4	11	6	6	6	29	0	0	100	38	19	28
T	45	50	51	46	38	36	39	47	57	60	19	20	0	0	18	41	30
	T	T	T	Y	Y	Y	T	Y	Y	T		C	<u>A</u>	<u>G</u>	R	T	

Mammalian examples (Actual numbers, from Nelson and Green, 1989):

	A	C	G	T	BP			
A	3	10	0	8	10	29	1	2
C	8	9	20	6	4	1	15	11
G	5	7	3	0	13	0	7	2
T	15	5	8	17	4	1	8	16
Consensus:	T	N	C	T	R	<u>A</u>	C	Y
Yeast sequence:	T	A	C	T	A	<u>A</u>	C	T

Examples from Drosophila:

ftz:	A	G	C	T	A	<u>A</u>	C	C	Rio
white:	T	C	T	T	A	<u>A</u>	T	A	Guo <i>et al.</i>
Myosin HC exon 19:	T	T	T	T	A	A	T	C	Bernstein*
Myosin HC exon 19:	A	A	C	T	A	A	T	T	Bernstein*
Myosin HC exon 6:	T	C	C	T	A	A	T	G	Bernstein*

Drosophila branchpoint matrix as determined by CONSENSUS (percentages):

A	42	39	0	24	73	88	16	9
C	18	0	71	25	17	8	12	37
G	8	21	27	2	9	0	0	19
T	31	40	1	49	0	4	72	35
Consensus:		W	C	T	A	<u>A</u>	T	Y
Information:	0.05	0.33	1.55	0.14	0.85	1.13	0.47	0.27

\*Hodges and Bernstein, unpublished.

S02

### Automated Prediction of Priming Sites for STS Sequences

C. A. Fields, B. Rappaport, C. A. Soderlund, V. Church,\* C. E. Hildebrand,\* and R. K. Moyzis\*

Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

\*Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

An automated system for screening candidate STS sequences, and predicting good PCR priming sites on such sequences, is being developed to support STS mapping projects at the LANL Center for Human Genome Studies. The software, which is currently being developed and tested, performs the following functions for each candidate STS sequence.

1. Sequences and their complements are compared with consensus sequences of known human repetitive elements, using *Ifasta*. Sequences with significant matches to a repeat consensus are rejected for use as STSs.
2. The remaining sequences and their complements are compared with all sequences in the primate section of GenBank<sup>®</sup> using *fasta*. Sequence regions having significant similarity to one or more sequences in GenBank are marked.
3. The unmarked regions of each sequence are analyzed for C+G content, ability to form secondary-structure hairpins, and com-

plementarity to each other. Pairs of noncomplementary sequences regions with similar C+G content and low secondary-structure probability are identified as potential priming sites.

4. All sequences are scanned for long open-reading frames. Any long open-reading frames are translated, and the predicted amino-acid sequences are compared with the sequences in the PIR database using *fasta*. This procedure provides a useful first check for STS sequences that overlap coding exons.

The parameters controlling the sensitivities of the database searches, significance of matches, C+G content, secondary-structure probability, and length of open-reading frames to consider significant are set by the user. Three output files are generated for each candidate STS sequence, which contain: i) a summary of matches to sequences in GenBank, and the predicted best primer sites, ii) matches between translations of long orfs and protein sequences in the PIR, if any, and iii) an archival log of all operations performed.

S03

### Approximate Pattern Matching and Biological Applications

E. L. Lawler and W. I. Chang

Computer Science Division, University of California, Berkeley, CA 94702

The proposed *sequence tagged sites* (STS) genomic map database of the Human Genome Project, as well as several mapping strategies, require the approximate matching of DNA sequences. Algorithms based on dynamic

programming calculate most if not all entries of an  $m$  by  $n$  table, where  $m$ ,  $n$  are respectively the lengths of the pattern and text sequences. We have done careful theoretical and empirical comparisons of these methods; apart

## Integration of Automated Sequence Analysis into Mapping and Sequencing Projects

C. A. Fields, C. A. Soderlund, and P. Shanmugam

Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

The goals of the *gm* automated sequence analysis project include both the development of tools to make sequence analysis faster and more informative, and the discovery of new features of sequences that can be used to facilitate the analysis of large DNAs that may contain many genes. During the last year we have added a number of functions to *gm*, developed several new interactive analysis tools, and used the system to analyze sequences, primarily from the nematode *C. elegans*, obtained from GenBank® or from collaborators at other laboratories. Some of the results of this work are as follows.

- Functions for displaying cDNA data, restriction maps, and STSs have been added to the *gm* interface. These allow *gm* to be used for a variety of experimental design tasks. *gm* can also now predict structures of genes given partial cDNA data (see accompanying abstract by Soderlund et al.).
- *gm* has been used extensively in the analysis of the 54 kb *unc-22* cosmid from *C. elegans*, which was sequenced by Guy Benian and colleagues at Emory University. This cosmid contains at least 5 different genes; two of these were predicted by *gm* independently of their experimental discovery. Working with this cosmid has been very valuable for understanding the behavior of sequence analysis methods on sequences containing multiple genes.
- The *gm* graphic interface is being used for the display and manipulation of exon maps in the *C. elegans* Community System, a distributed database of genomic information on *Caenorhabditis* being developed by B.Schatz and S. Ward at the University of Arizona. The interface displays exon maps, restriction maps, the DNA sequence, and predicted protein sequences in a single window; it thus provides access to both physical and functional information at high resolution in the Community System.
- A variety of new tools for calculating information contents of sets of aligned sequences have been developed. These are being used to examine the information contents of both single-base positions and base correlations in splice sites of *C. elegans* and *Drosophila*. Significant differences exist in the information contents of introns from short versus long introns in both organisms.
- New tools for sensitive base-composition analysis of sequence regions have also been developed. These are currently being used to better characterize both exon and intron sequences in *C. elegans* and humans.

As additional tools are developed and tested, they will be distributed to the community as part of the *gm* software package.

of large data sets generated with public funds should no longer be acceptable for genome grants. Centralized repository databases play many important roles for biological researchers, yet lack the immediacy of access to "fresh" data that is desired for collaborative research and community support. Recent advances in networking and databases permit a 3rd option: cooperating individual databases that function as a larger "distributed" database yet maintain complete local autonomy.

Some modern commercial relational databases provide true server/client access, which opens the possibility of allowing collaborators to have direct access to each other's data, under complete control of the owner. All details of access to remote database(s) can be hidden from end users with suitable front-ends. Issues of data security are in general handled trivially by the database package (by allowing read-only access to specified tables/views to external collaborators). Data control issues are more complex (e.g., if we collaborate with A and B by sending them clones/probes and put their results into our database, if both A and B are external users of our database they may not want the other to be able to view "their" data until they publish.) One suggested solution is to place a six-month hold on data before it could be seen by non-collaborators.

Without arguing the merits or ethics of this approach, we have decided to demonstrate the technical feasibility of multi-collaborator and distributed genome databases. Experiments were conducted with the help of Debra Nelson of LANL that proved the ease of use of allowing controlled access to external databases over the Internet. Other experiments

P17

Contig Assembly Program (CA)

C. A. Soderlund, P. Shanmugam, and C. A. Fields  
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

The Contig Assembly program (CA) solves the following problem: given multiple clones, each of which is made up of multiple fragments,

were done at LLNL to verify that the 6-month "data hold" concept was feasible, at the cost of adding owner and timestamp fields to any tables that might contain "proprietary" data.

We view this means of data sharing as an important tool for communities of tightly-coupled collaborating researchers willing to work with external databases at a fairly low level. We recognize that objections may be raised by sites that have databases from different vendors, and note that Sybase front-end tools are relatively inexpensive for genome researchers. In addition, the LLNL C-language interface library was designed to be portable to other relational databases and would make it easy to "import" data from a collaborator's "foreign" database, or to do cross-database pseudo-join queries.

We conclude that multi-collaborator and distributed genome databases are technically feasible, necessary, and worth the overhead in data storage and database administration.

Given Internet access and Sybase front-end tools any collaborator can easily share our data in a tightly-controlled fashion that does not put an undue burden on the collaborator. Collaborators who wish to maintain their own Sybase database(s) can grant similar privileges to us, under their complete control. Work in progress with Johns Hopkins will allow GDB to access our physical mapping data using these methods.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

determine if there exists one or more combinations of the fragments, such that the clones may be arranged in a contiguous line. CA

works by determining overlaps, and is intended for use with total digest data. List all possible solutions.

Problems arise due to: 1) matching fragments do not always have exactly the same length, 2) unique fragments may have the same length, and 3) fragments may be missing. Due to these problems, CA often cannot find a solution even when one exists. To solve this problem, CA is used interactively, as follows: 1) CA is run with pairs of clones until a good overlap is found. 2) A new clone can be displayed to see

the relation of its fragments to the existing contig. 3) If the new clone overlaps the existing clone, but has one of the problems stated above, the clone can be edited to correct the problem and then added to the contig.

Our current work concentrates on: 1) finding the "best" solution regardless of errors, 2) using STS data to resolve inconsistencies, and 3) allowing weights to influence what partial solutions are pursued.

P18

### New Features in gm, Version 2.0

C. A. Soderlund, P. Shanmugam, and C. A. Fields  
Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001

Version 1.0 of the gm automated sequence analysis system was released in January, 1990, and has now been installed at several dozen laboratories worldwide. The gm system predicts the exon-intron organization of genes from genomic DNA sequence data, and displays the resulting exon maps and predicted amino-acid sequences via a graphic user interface (Fields and Soderlund, *CABIOS* 6 (1990) 263-270). It is designed to support incremental, exploratory analysis of new sequences as they are obtained, and is intended for use as a laboratory tool. We have spent the last year testing gm with sequences of known genes, using gm to analyze new sequences obtained by collaborators, and designing and implementing new functions. The new release, gm version 2.0, includes the following new features.

- Partial 3'-end cDNA data can be used to initiate the exon maps. This allows maps consistent with a partial cDNA to be examined either alone, or together with all other possible maps. By predicting only exons that consistently extend a known cDNA, gm can be used to efficiently design PCR primers for amplification of sequences from a total cDNA library.

- Predicted exon maps are now ranked in order of increasing protein-coding capacity. The user can choose to view only the highest-ranked nonoverlapping maps, each of which has the maximal coding capacity for the region that it spans. This procedure generates the maps most likely to produce hits in protein database searches, while greatly decreasing the total number of maps that the user must examine.
- Functions for displaying microrestriction maps, STS locations, cDNAs, and repetitive DNA elements at the same scale as the predicted exon maps have been added to the graphic interface. This allows predicted genes to be quickly aligned with physical maps, and facilitates the selection of restriction fragments to be used in probes of Northern blots.
- The graphic interface supports multiple system runs with different parameter settings. This facilitates use of gm as an exploratory analysis tool. It also allows gm to serve as an interface for displaying exon maps of

## SOFTWARE FOR THE C. ELEGANS GENOME PROJECT

R. Durbin<sup>1</sup>, S. Dear<sup>1</sup>, T. Gleeson<sup>1</sup>, P. Green<sup>3</sup>, L. Hillier<sup>3</sup>, C. Lee<sup>1</sup>, R. Staden<sup>1</sup> and J. Thierry-Mieg<sup>2</sup> <sup>1</sup>MRC Laboratory of Molecular Biology, Cambridge, U.K.; <sup>2</sup>CRBM du CNRS, Montpellier, France; <sup>3</sup>Department of Genetics, Washington University School of Medicine, St Louis, MO.

One of the aims of the *C. elegans* sequencing project is to develop an integrated computer package to organize and automate the collection and analysis of data. We have been working on two main programs, one for sequence assembly from fluorescent sequencing machine data, and the second a database for storing and forming links between all sorts of genomic stations) using the X11 windows system.

The sequence assembly program, dap, has been developed from Rodger Staden's sap program. Sap has already been used for numerous large scale sequencing projects, including one of 150kb done as a single shotgun (A. Davison, personal communication). In addition to handling conventional radioactive sequence readings, dap can manage raw trace and sequence data from both ABI 373A and Pharmacia ALF fluorescent sequencing machines. It automatically assembles readings and all editing is performed using a new mouse operated contig editor that displays aligned sequences and their traces together on the screen. Dap also provides numerous tools for assessment of progress and further developments are underway. We have assembled one 45kb cosmid, are part way through several more, and are finding that the improvements embodied in dap are a major advance for cosmid scale sequencing. Smaller stand alone programs exist for oligo selection (osp) and for display and editing of the raw sequence together with its traces (ted). Osp can also suggest PCR primers.

The database program (acedb) is a flexible mouse driven system that handles the sequence data, genetic and physical maps, the *C. elegans* bibliography, and strain and raw genetic data (gene and allele lists and mapping data). Objects are stored in an extendable structure, so that arbitrarily large amounts of information can be stored in them, including annotations (both structured with keywords, and unstructured with arbitrary comments), and cross-references to other objects. There is a general search facility, and displays can be output in Postscript for laserprinting, or plain text for transfer to other programs. We are currently working to integrate functions that perform genetic map and sequence calculations. The likelihood based map calculations will allow a user to either assemble a map automatically or work on it interactively.

## PERFORMANCE OF gm, VERSION 2.0

C. A. Fields, C. A. Soderlund, and P. Shanmugam, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001, USA.

The gm system predicts the exon-intron organizations of genes from genomic DNA sequence data, and displays the resulting exon maps and predicted amino-acid sequences via a graphic user interface (Fields and Soderlund, *CABIOS* 6 (1990) 263-270). It is designed to support incremental, exploratory analysis of new sequences as they are obtained, and is intended for use as a laboratory tool. gm has been distributed to several dozen user sites, which have provided feedback on the utility of the system for analyzing DNA from a number of organisms. We have spent the last year testing gm, version 1 on sequences of known genes from *C. elegans*, using gm to analyze new sequences obtained by collaborators, and designing and implementing new functions (see abstract by Soderlund, et al.). We are currently evaluating the performance of the new release, gm version 2.0, using primarily nematode and human DNA sequences.

The 54 kb *wnc-22* cosmid of *C. elegans* (G. Benian et al., *Nature* 342 (1989) 45-50) was used for an initial large-scale test of gm v. 1. The locations of several new *wnc-22* exons, and of four additional genes, were predicted using gm. The existence of several of the predicted exons, and two of the four predicted genes has been independently confirmed by G. Benian and colleagues at Emory University by cDNA sequencing; the existence of an additional predicted gene has been confirmed by D. Baillie and colleagues at Simon Fraser.

gm has also proven useful for comparative sequence analysis, by allowing sets of sequences to be analyzed with the same procedure and stringencies. The number of predicted exon maps generated by gm at a fixed analysis stringency provides a measure of the complexity of a sequence. At least in *C. elegans*, sequences containing regulatory genes are more complex by this measure than sequences containing structural proteins; hence regulatory genes are substantially harder to predict accurately from sequence data. Intron lengths and base compositions also differ significantly between regulatory genes and structural protein genes in *C. elegans*. The results of these analyses have been used in the design of improved compositional analysis functions that are incorporated in gm version 2.

The gm version 2.0 graphic interface is being used for the display and manipulation of exon maps in the *C. elegans* Community System, a distributed database of genomic information on *Caenorhabditis* developed by B. Schatz and colleagues at the University of Arizona. The interface displays known exon maps, restriction maps, the DNA sequence, and predicted protein sequences in a single window; it thus provides access to both physical and functional information at high resolution in the Community System. We expect that the interface will be also useful for the display of archival exon map information in other genomic information systems.

## AUTOMATED PREDICTION OF PRIMING SITES FOR STS SEQUENCES

B. Rappaport, C. A. Soderlund, and C. A. Fields, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001, USA.

An automated system for screening candidate STS sequences, and predicting good PCR priming sites on such sequences has been developed. The software performs the following functions for each candidate STS sequence.

1. Sequences and their complements are compared with consensus sequences of known human repetitive elements, using *fasta*. *Masta* returns all significant matches of a query sequence to a database; hence it will identify multiple fragments of repeats such as *Alu* in a candidate STS. Sequences with significant matches to a repeat consensus are rejected for use as STS.
  2. The remaining sequences and their complements are compared with all sequences in the primate section of GenBank using *fasta*. Sequence regions having significant similarity to one or more sequences in GenBank are identified using a sliding-window averaging procedure with adjustable window width and similarity score cutoff. This procedure yields a summed similarity histogram for the STS sequence.
  3. The regions of each sequence that are not similar to any known sequences at the levels set in step 2 are analyzed for length, separation, C+G content, dinucleotide content, ability to form secondary-structure hairpins, and complementarity to each other. Pairs of noncomplementary sequences regions with similar composition and low secondary-structure probability are identified as potential priming sites.
  4. All sequences are scanned for long open-reading frames. Any long open-reading frames are translated, and the predicted amino-acid sequences are compared with the sequences in the PIR database using *fasta*. This procedure provides a useful first check for STS sequences that overlap coding exons.
- The parameters controlling the sensitivities of the database searches, significance of matches, compositional analysis, secondary-structure probability, and length of open-reading frames to consider significant are set by the user. Three output files are generated for each candidate STS sequence, which contain: i) the summary histogram showing similarity to sequences in GenBank, and the predicted best primer sites, ii) matches between translations of long orfs and protein sequences in the PIR, if any, and iii) an archival log of all operations performed.

Sequence data from random genomic clones generated for physical mapping of Chromosomes 5, 7, and 16 (Center for Human Genome Studies, Los Alamos National Laboratory) are being used to test the software, and develop additional heuristics for use at step 3. Virtually all sequences tested that do not contain repetitive elements nonetheless contain 20-nucleotide or larger regions that match sequences in GenBank at *u.*: 90% or greater level, confirming the utility of GenBank scans in primer prediction. One sequence has thus far been identified as probably coding for a previously uncharacterized human gene.

## AN STS-RAPD AND RFLP MAP IN *Arabidopsis thaliana* USING RECOMBINANT INBREDS.

Robert Reiter, Kenneth Feldmann, Andrew Paterson, Antoni Rafalski, Scott Tingey, John Williams and Pablo A. Scolnik. Du Pont Central Research and Development, Wilmington, DE 19880-0402.

Due to its small genome size and rapid generation time *Arabidopsis thaliana* can be used as a model system to clone plant genes by genetic approaches. To facilitate this work we are constructing a high resolution recombination map. Two RFLP maps of *Arabidopsis* have been constructed (1,2). However, these maps are based on F2 populations, which cannot be propagated indefinitely. Recombinant inbreds (RI) lines are generated by crossing two highly inbred progenitors and maintaining the progeny of the F2 under a strict inbreeding regime. As the number of generations increases, RI lines approach full homozygosity. Thus, RI lines constitute a permanent collection. We developed an *Arabidopsis* RI collection by inbreeding the progeny of a W100 X WS (Wassilewskij) cross. W100 is a marker line containing nine phenotypic markers (3). A total of 175 lines have been inbred to F8. We are mapping three classes of probes into this population: i. Random amplified polymorphic DNA (RAPD, ref. 4); ii. RFLPs (1,2), and iii. known *Arabidopsis* genes. RAPD reactions with 10-mers yield an average of 5 bands, of which about 20% are polymorphic. To date, we have mapped 200 RAPD markers. Including RFLPs the map will contain in excess of 300 markers.

1. Chang et al. (1988) Proc. Natl. Acad. Sci. USA 85:6856-6860.
2. Nam, H-G (1989) The Plant Cell 1:699-705
3. Koornneef et al. (1987) A.I.S. 23:46-50.
4. Williams et al. (1990) Nuc. Acid. Res. 18:6531-6535.

#### IDENTIFYING AND MAPPING GENES ON HUMAN CHROMOSOME 3

D.J. Smith<sup>1</sup>, W. Golembieski<sup>1</sup>, V. Shridhar<sup>1</sup>, O.J. Miller<sup>1</sup>, W. Liu<sup>1</sup>, S. Smith<sup>1</sup>, K. Merchant<sup>1</sup>, F. Recchia<sup>1</sup>, A. Kamati<sup>1</sup>, and H. Drabkin<sup>2</sup>. <sup>1</sup>Molecular Biology and Genetics, Wayne State University, Detroit, MI; <sup>2</sup>University of Colorado Health Sciences Center, Denver, CO

Our laboratory is focusing its efforts on the isolation of chromosome 3-specific genes. Our ultimate goal is the cloning of the genes responsible for small cell lung and renal cell carcinoma and Von Hippel Lindau (VHL) disease. In addition to physical mapping studies which should enable us to narrow the regions that must contain these genes we have also concentrated on identifying large numbers of genes from the larger regions known to contain these disease loci. We have isolated over 6,500 chromosome 3-specific recombinants, representing 210,000 Kb of chromosome 3 DNA, prepared DNA from each clone and made filters containing these clones for rapid screening and cosmid walking. We isolated unique sequence probes from 616 of the recombinants and then localized them by hybridization to a somatic cell hybrid deletion mapping panel. We isolated a total of 25 clones from human chromosomal band 3p21.1 (thought to contain the renal cell carcinoma tumor suppressor) and over 100 recombinants from 3p25 (containing the VHL gene). We performed cosmid walking within 3p21.1 to isolate several contigs of overlapping cosmids. We isolated overlapping cosmids to link aminoacylase-1 to D3S2 (a region of over 150 Kb). There are 6 HTF islands and at least 3-4 genes within this region. Chromosome walking from several other recombinants within 3p21.1 reveals that this region contains large numbers of HTF islands and a very high density of genes. Similar findings were also observed with clones from 3p25. DNA sequences from human chromosomal bands 3p21.1 and 3p25 are well represented in our resource of isolated chromosome 3-specific recombinants. We localized sufficient recombinants to these bands to isolate all the DNA sequences contained within them in overlapping YAC clones. All the genes isolated from within these bands will then be tested individually as candidate genes for renal and small cell lung carcinoma or VHL disease.

#### NEW FUNCTIONS IN gm, VERSION 2.0

C. A. Soderlund, P. Shanmugam, and C. A. Fields, Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001, USA.

Version 1.0 of the gm automated sequence analysis system (Fields and Soderlund, *CABIOS* 6 (1990) 263-270) was released in January, 1990, and has now been installed at several dozen laboratories worldwide. The new release, gm version 2.0, includes the following new functions.

- Partial 3'-end cDNA data can be used to initiate construction of exon maps. This allows maps consistent with a partial cDNA to be examined either alone, or together with all other possible maps. By predicting only exons that consistently extend a known cDNA, gm can be used to efficiently design PCR primers for amplification of sequences from a total cDNA library.
- Repetitive elements are required to be contained in either introns or intergenic regions. Ifasta or a similar similarity searching program can be used to generate a file of the positions of repetitive elements. This file can be used as input into gm to screen for introns and intergenic regions.
- Two- and three-dimensional Markov matrix analyses are used to measure single, di- and tri-nucleotide compositions of introns and exons. The tests applied to these matrix elements to identify exons and introns can be varied for use with sequences from different organisms.
- Predicted exon maps are now ranked in order of increasing protein-coding capacity. The user can choose to view only the highest-ranked nonoverlapping maps, each of which has the maximal coding capacity for the region that it spans. This procedure generates the maps most likely to produce hits in protein database searches, while greatly decreasing the total number of maps the user must examine.
- Functions for displaying microrestriction maps, STS locations, cDNAs, and repetitive DNA elements at the same scale as the predicted exon maps have been added to the graphic interface. This allows predicted genes to be quickly aligned with physical maps, and facilitates the selection of restriction fragments to be used in probes of Northern blots.
- The graphic interface supports multiple system runs with different parameter settings. This facilitates use of gm as an exploratory analysis tool. It also allows gm to serve as an interface for displaying exon maps of known genes together with cDNA and physical mapping data. The interactive menu tool can also be accessed directly from the graphic interface. menu now includes a function that builds exon maps from known exon coordinates.

Our current effort is focussed on developing functions to allow use of arbitrary cDNA data to initiate exon map construction, and on developing and testing methods for dealing with sequencing errors.

As more worm genes are sequenced, considerable variability in the characteristics of both exons and introns at the DNA level is becoming apparent. Features of interest include:

**Length.** Lengths of both exons and introns range from a few tens of bases to around 10 kb. Short exons are very common. While the peak of the intron size distribution is still at 50-52 nucleotides, introns with lengths greater than 1 kb are now fairly common.

**Base composition.** Exons are richer in C and G than introns, but the average difference in composition between exons and introns is decreasing as more sequences become available. Long introns often have islands of relatively high C+G content. The dinucleotides AA and TT are, however, much more common than CC or GG in introns.

**Codon usage.** Codon preferences differ significantly between gene families. There is a general trend toward less asymmetrical codon preferences in weakly-expressed gene families; however, there are also strong codon preference residuals in some gene families, e.g. *gfp-1* and *lin-12*.

**Splices.** 5' splice-site sequences differ significantly between long and short introns. This variation, first observed in *C. elegans*, has now been noted in *Drosophila* and plants, and may be ubiquitous. 3' splice site sequences appear not to vary with intron length.

**Complexity.** Regulatory genes appear to contain larger numbers of potential splice sites, and to have less-asymmetric base compositions between exons and introns, than highly-expressed structural-protein genes. The efficiency of correct splice-site selection may be affected by these features.

The variability in these features significantly increases the difficulty of gene identification by sequence analysis. We have used *C. elegans* sequences extensively in testing our gm automated DNA sequence analysis software. A number of new introns have been incorporated into gm v. 2.0 to provide greater accuracy of predictions in the face of compositional and site variations between genes. These include more accurate site-identification methods that combine consensus matrix and compositional methods and flexible multinucleotide compositional measures. Functions have also been added that allow automated use of partial cDNA data for genomic sequence analysis. The performance of gm v. 2.0 is currently being evaluated on a large set of worm genes.

### Expression of *fem-1*.

A. M. Spence, Dept. of Medical Genetics, University of Toronto, Canada.

Loss-of-function mutations in the *fem-1* gene transform animals of both chromosomal sexes into fertile females, indicating that *fem-1* function is essential for male development in both soma and germline<sup>(1)</sup>. The DNA sequence of the gene predicts that its product is an intracellular protein containing six tandemly arranged 33 amino acid cdc10/SWI6 motifs near its N-terminus<sup>(2)</sup>. Northern analysis shows that the 2.4 kb *fem-1* mRNA is expressed throughout hermaphrodite development at approximately constant levels. Although *fem-1* acts in hermaphrodites only to influence the sexual fate of the germline, its mRNA is present in mutant hermaphrodites lacking a germline [*glp-1(e2144)* or *SS104(bn2)*] suggesting that in contrast to *fem-3*<sup>(3)</sup>, the gene is expressed in hermaphrodite somatic tissues. Reduced levels of *fem-1* transcripts in animals lacking germ cells suggest that the germ line is however a major site of *fem-1* expression. In adult males, a sub-population of transcripts slightly larger than those seen in hermaphrodites is evident. Whether the change in size is due to regulation of polyadenylation as has recently been observed in *fem-3(gf)* mutants<sup>(4)</sup> is under investigation.

(1) Doniach, T. and J. Hodgkin. (1984). *Dev. Biol.* 106: 223.

(2) Spence, A.M., A. Coulson, and J. Hodgkin. (1990). *Cell* 60: 981.

(3) Rosenquist, T. and J. Kimble. (1988). *Genes and Dev.* 2: 606.

(4) Ahringer, J. and J. Kimble. (1991). *Nature* 349: 346.

THE EXPRESSED SEQUENCE TAG DATABASE: A RESOURCE FOR GENOMICS AND DEVELOPMENTAL BIOLOGY. C. A. Fields, M. D. Adams, M. Dubnick, A. R. Kerlavage, and J. C. Venter. Section of Receptor Biochemistry and Molecular Biology, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD 20892 USA.

The cDNA sequencing and mapping projects are expected to generate well over 10,000 sequences, and several thousand map positions, of expressed sequence tags (ESTs) per year for the next several years. ESTs from humans, *Drosophila*, and *Caenorhabditis* are already being obtained. ESTs correspond to clones that can be used directly as probes for gene activity; physically mapped ESTs thus allow gene expression data to be attached directly to positions on physical maps.

We are developing a database of EST sequences, map positions, and expression data, together with a number of tools for automatically comparing EST sequences with various databases and picking PCR primers for mapping or follow-up sequencing. The database is being implemented in SYBASE, with most of the processing tools implemented as Unix scripts for existing programs. This strategy allows a working system to be assembled relatively quickly, both for immediate use and to serve as an evolving functional specification.

We anticipate that the EST database will be one of a number of loosely-linked, locally-curated sequence and mapping databases. Several collaborating groups will be remotely retrieving data from and depositing data into the database; we will also be distributing sequence data to the public databases on a regular basis. The availability of large numbers of EST sequences will considerably simplify the analysis of anonymous genomic sequences, and will significantly increase the size and diversity of the amino-acid sequence databases.

## COORDINATING AND OPTIMIZING cDNA MAPPING AND SEQUENCING EFFORTS

Charles R. Cantor, University of California, Berkeley, CA

It is now quite widely accepted that mapping and sequencing most detectable human cDNAs is an attractive and important activity within the scope of the Human Genome Project. HUGO, the international Human Genome Organization, is attempting to coordinate various national efforts in human cDNA characterization. At a minimum, HUGO can serve as a clearing house to keep track of numerous cDNA activities and as a forum for communication among interested participants. A number of technical issues involved in human cDNA mapping and sequencing need to be resolved before the most efficient strategies for these activities become apparent. These mostly concern how to optimize the rate of obtaining information in any large scale mapping effort. Some of the issues involved will be illustrated. After the cDNA cream has been skimmed from the genome will it still be worthwhile to sequence the remainder? It is debatable, but there are strong arguments in favor of it.

**ARTIFICIALLY INTELLIGENT TOOLS FOR GENE RECOGNITION AND ASSEMBLY FROM DNA SEQUENCE DATA.** E. Uberbacher<sup>1,2,3</sup>, R. Einstein<sup>2</sup>, X. Guan<sup>2</sup>, R. Mann<sup>2</sup>, and R. Mural<sup>1</sup>. Biology<sup>1</sup> and Engineering Physics and Mathematics<sup>2</sup> Divisions, Oak Ridge National Laboratory and UT-ORGSBMS<sup>3</sup>, Oak Ridge, TN 37831.

We are building a hybrid neural network and rule-based inference system to examine and characterize regions of anonymous DNA sequence based on a new approach to feature identification using a multiple sensor-neural network formalism. As an example, we present a module which correctly recognizes 92% of the coding exons of 100 or more bases with very little noise. This, along with additional such modules to recognize splice junctions and other features, represents powerful tools which can be used in a stand-alone manner and in the integrated system. These tools have recently been combined with a rule-based interpreter and user interface to allow analysis of DNA sequence over the Internet (Gene Recognition and Analysis Internet Link; GRAIL). Users E-mail DNA sequences to the system and have the analysis returned automatically by E-mail. The current analysis includes potential exon positions, with strand assignment and preferred reading frame determination, and parallel database comparison. The integrated system we are constructing uses the blackboard CLIPS rule-based expert system shell to automatically integrate recognized features into hypothetical gene models. This system consists of a series of hierarchically arranged "blackboard panels" on which connection of potential features and construction of gene fragments is accomplished by a set of logically independent knowledge modules. A tentative coding message will be automatically compared to sequence databases using highly parallel methods. (This research was sponsored by the Office of Health and Environmental Research, U. S. Department of Energy, under contract DE-AC05-84OR21400 with the Martin Marietta Energy Systems, Inc.)

#### GENE IDENTIFICATION USING gm

Carol Soderlund, Pari Shanmugan, Owen White and Chris Fields, Computing Research Laboratory, New Mexico State University

In our pursuit to make gm v2.0 work for all organisms, we studied *C. elegans*, *Drosophila*, humans, monocots, and dicots. In summary: (1) Our results have supported other studies which show that vertebrate and invertebrate exons may be characterized by 6-mers; however, the distribution of information among 6-mers differs between different organisms. (2) The false-positive and false-negative rates for identifying splice sites can be minimized in invertebrates by evaluating the nucleotides on both sides of a potential splice site. These two evaluation schemes have been implemented in gm v2.0, along with the following additional functions: (1) the ability to extend a cDNA to generate a complete exon map, and (2) the use of repeat sequences to disqualify potential exons.

gm V2.0 can be used for different organisms by using the composition table for the specific organism. With the release of the software, we will supply the composition tables for exons and introns for the five organisms previously stated. For greater flexibility, composition tables can be built based on some other criteria than organism-specific; for instance, composition tables may be built for a functional class of genes, this would result in gm evaluating exons and introns based on functionality instead of organism. Additionally, gm v2.0 has been designed so that the composition tests for exons, introns, and functional sites may easily be changed. The initial results of evaluating worm and human cosmid sequences with gm v2.0 will be presented.

ABSTRACT FORM

**INFORMATICS SUPPORT FOR LARGE-SCALE SEQUENCING**

Chris Fields, Institute of Neurological Disorders and Stroke, NIH  
Bethesda, MD 20892

The principal goals of the current phase of the Human Genome Project are map construction and gene identification. Two types of large-scale DNA sequencing efforts are currently contributing to realizing these goals: expressed-sequence tag (EST) and sequence-tagged site (STS) sequencing to obtain large numbers of short unique sequences for mapping, and in the case of ESTs, identifying genes by sequence similarity, and cosmid contig sequencing, largely in support of positional cloning efforts. Both types of sequencing project rely heavily on automated sequence analysis. Six human cosmids and over 3000 human ESTs have now been sequenced in this laboratory. These projects provide a testbed for evaluating available analysis methods, and for developing custom software to automate sequence data analysis and database maintenance.

Name Chris Fields

Institution/Address Institute of Neurological Disorders & Stroke  
NIH, Bethesda, MD. 20892

Phone (301) 496-8800 FAX (301) 480-8588

gm v2.0 alpha  
11/5/91

Copyright (c) Computing Research Laboratoy, New Mexico State  
University, Las Cruces, NM 88003-0001, USA.

#### Description:

gm takes as input a file of parameters, and outputs a set of candidate exon maps. The gm program may be run on a text terminal. The gmwin program runs under X-windows: it runs gm and displays the output graphically. The menu program provides a set of auxiliary routines.

#### Authors:

Cari Soderlund: gm and menu software  
Pari Shanmugan: gmwin software  
Ted Slater : Help files  
Owen White, Cari Soderlund, and Chris Fields: compositional matrices  
Chris Fields and Cari Soderlund: functional specification

#### Citation:

C.Soderlund, P.Shanmugam, O.White, C.Fields, "A tool for exploratory analysis of DNA sequence data", Proceedings of the Hawii International Conference on System Sciences, Biotechnology Computing Minitrack, January 7-10, 1992.

#### Acknowledgements:

This research was supported in part by US Department of Energy, Genome grant 89ER60865 to C.A.F. and C.A.S.  
We would like to thank Alen Dunn, Electrical Engineering, University of Sydney, for the sopen routine, which allows the user to set a search path for gm files.

#### Contents:

- I. Install
- II. Demo
- III. Files
- IV. Getting Started
- V. Tuning gm for a different organism
- VI. Adjusting parameters
- VII. Bugs
- VIII. On-line documentation

Please send all bug reports or suggestions to cari@nmsu.edu.

#### I. INSTALL

##### Execute

```
ftp haywire.nmsu.edu (userid is "anonymous")
> binary
> cd gm
> get gm2.tar.Z
> quit
uncompress gm2.tar.Z
tar xvf gm2.tar
```

The following directories will be built:

```
gm/bin
gm/demo
gm/files
gm/help
gm/et-src
gm/gm-src
gm/X-src
```

## Execute

```
cd gm
sh install
```

The executables gm, gmwin, and menu will be built and moved into the directory gm/bin.

1. It is assumed that gm is built under /usr. If it is placed elsewhere, please change "/usr" to the appropriate directory in (1) this README file and (2) gm/Help/Help.get\_start.

2. If it is desirable to place the binaries in a directory other than gm/bin, it is necessary to redirect where gm will find the Help files. The include file gm/gm-src/Cons.h has it defined as:

```
#define HelpDir "../Help/"
```

change this appropriately.

3. The default organism is human: (a) The defaults in the menu program are set up for humans, and (b) the discussion below assume that the organism of interest is humans. If the organism of interest is C.elegans, monocots, dicots, or Drosophila, make the following changes:

```
a. cd gm-src
   mv Defs.h Defs.hum
   mv {Defs.worm, Defs.mono, Defs.dicot, or Defs.fly} Defs.h
   cd ..
   sh install
```

```
b. cd Help
   edit Help.get_start
   find - setenv GMSEARCHPATH /usr/gm/files/human
   change human to {worm, mono, dicot, or fly}
   Note: a user may have all five directories in their search path.
```

This only initializes the defaults; a user may run gm on any organism, as long as the input parameters specify the matrices for that organism.

## II. DEMO

-----

Assume that gm has been installed under the /usr directory.

Edit your .cshrc file, and enter the following lines:

```
setenv GMSEARCHPATH /usr/gm/files/human:/usr/gm/files/worm:/usr/gm/files/re
set path = ($path /usr/gm/bin)
```

## Execute

```
source .cshrc
```

If running in X-windows, you must have the following line in your .Xdefaults.

```
*font: 6x13
```

1. If you have permission to run in /usr/gm/demo, cd to this directory; otherwise, copy the files from /usr/gm/demo to your directory; i.e. type "cp /usr/gm/demo/\* ." (the "." indicates your current directory).

2. If running under X-windows: type "gmwin".  
The gmwin graphics will appear.

A. Enter "deb-1.in" in the Input File slot.

This runs the gm algorithm.

Click on "deb-1-1" in the right-hand box.

An exon map will appear.

Enter "deb-1.cdna" in the cDNA File slot.

This displays the cDNA across the top.

Enter "EcoRI" in the Enzyme File slot.

This displays the restriction map across the top.

Enter "deb-1.sts" in STS File slot.

This displays the STS's across the top.

Click on the "Window 1" button. Click on "emap-2".

You should see 2 exon maps; you may view up to 4 exon maps at a time.

Click on any of the entities in the upper window.

The entities will be highlighted in the dna sequence and its position and scores will be display in the window on the right. To highlight the amino acid sequence of an exon, click on the exon and toggle the "Sequence" button.

B. Enter "humadag.emap" in the Exon Map File slot.

This displays the contents of the exon map file.

Click on "humadag-1" in the right-hand box.

An exon map will appear.

Enter "humadag.rep" in the Repeat File slot.

This displays the repeats along the bottom.

Play with clicking on different entities.

C. Click on the "quit" button to exit gmwin.

3. If running from a text terminal: type "gm deb-1.in".

To view the output files, use an editor or the more command; i.e. type "more deb-1.emap".

4. Type "menu". At the prompt, type "?". In response to the help menu prompt, type "1". You should see general documentation displayed on your screen. Type "x" to exit.

### III. FILES

1. The following files are distributed with gm:

Sequence analysis file:

	gm/files/worm	gm/files/human
	-----	-----
(1) codon usage	wormcodon	humcodon
(2) exon evaluation	wormexon.{1-6}	humexon.{1-6}
(3) intron evaluation	wormintron.{1-6}	humintron.{1-6}
(4) 5' splice site	worm5-13.log	hum5-13.log
(5) 3' splice site	worm3-13.log	hum3-13.log
(6) promoter	tataa	tataa
(7) polyA	aataaa	aataaa
(8) Initiation	wormatg-9.log	atg
(9) example input file	worm.in	hum.in
gm/files/fly	gm/files/mono	gm/files/dicot
-----	-----	-----
flycodon	monocodon	dicodon
flyexon.{1-6}	monoexon.{1-6}	diexon.{1-6}
flyintron.{1-6}	monointron.{1-6}	diintron.{1-6}
invert5-13.frq	plant5-9.frq	plant5-9.frq
invert3-13.frq	plant3-9.frq	plant3-9.frq
tataa	tataa	tataa
aataaa	aataaa	aataaa

flyatg-9.frq  
fly.in

plantatg-9.frq  
mono.in

plantatg-9.frq  
dicot.in

{1-6} implies a file for each of the 6 suffices; i.e. wormexon.1, wormexon.2, etc.

Restriction enzymes: gm/files/re

BglII BssHII ClaI EcoRI HindIII KpnI NaeI NotI NruI  
PstI PvuI PvuII SacI SacII SalI SmaI XbaI XhoI

Demo: gm/demo

c.elegan demo:

-----  
deb-1.dna : input sequence  
deb-1.in : input parameters  
deb-1.cdna : cDNA coordinates  
deb-1.coord : complete coordinates  
deb-1.sts : sts coordinates (made up for example)  
deb-1.emap : output exon maps  
deb-1.trans : output translation file

human demo:

-----  
humadag.dna : input sequence  
humadag.in : input parameters  
humadag.coord : complete gene coordinates  
humadag.rep : alu repeats in sequence  
humadag.emap : output exon maps  
humadag.trans : output translation file

## 2. Naming conventions:

- \*.dna : DNA sequence in gm format
  - a. The filename is a parameter in the input file.
  - b. A DNA file can be converted to gm format in the menu program under "File Conversion".
- \*.in : gm input parameters
  - a. Input to gm; i.e. execute "gm deb-1.in".
  - b. Input to the gmwin slot labeled "Input File".
  - c. This file is created and updated from the "gm Input File: create/update" option in menu.
- \*.emap: output exon maps
  - a. The filename is a parameter in the input file.
  - b. Output of both gm and gmwin (2a. and 2b).
  - c. Input to the gmwin slot labeled "Exon Map File".
- \*.trans: output translation file  
The filename is a parameter in the input file.
- \*.cdna: 3' cDNA coordinates
  - a. The filename is an optional parameter in the input file. gm will try to extend the cDNA upstream.
  - b. Input to the gmwin slot labeled "cDNA file".
- \*.coord: partial or complete gene coordinates
  - a. Input to many of the menu program routines; the specific routine is run on all coordinates in the file.
  - b. If this file is the complete gene, the file may be used for testing purposes as the cDNA parameter in the input file; gmwin will display the complete gene and all predicted genes.

c. If this file is the complete gene, an exon map file can be created by selecting the "Exon Map: create from coordinate file" option from the menu program; this file may be display by gmwin.

- \*.rep: repeat coordinates
  - a. Optional parameter in the input file.  
gm will exclude splice sites found in the repeat regions.
  - b. Input to the gmwin slot labeled "Repeat File".
- \*.sts: sts coordinates  
Input to the gmwin slot labeled "STS file".
- \*.cmp: complement DNA file  
This file can be created in menu under "File Conversion".
- \*.pep: Peptide format of a .trans file.  
This file can be created in menu under "File Conversion".

Samples of all these files are in gm/demo.

### 3. Extra files

The menu program creates a file called .gmMenuDefaults. This will record the files you used when using menu, so that the next time you use menu, it will use your previously specified files for the defaults. You do not have to concern yourself with this file, except to know that it is created and used by menu.

### IV. GETTING STARTED

-----

Put the following lines in the .cshrc file.

```
set path = ($path /usr/gm/bin)
setenv GMSEARCHPATH /usr/gm/files/human:/usr/gm/files/re
```

If running in X-windows, you must have the following line in your .Xdefaults.

```
*font: 6x13
```

In this example, the dna sequence file name is yyy.dna and the input file yyy.in will be created. In the following discussion, replace yyy with the file name of your sequence.

1. Type "menu".
2. To get your DNA file in gm format, pick item number 2 on the main menu, and item number 1 on the file conversion menu. Enter your DNA filename at the prompt, and a new filename for the second prompt.
3. From the main menu:
  - a. Pick item number 3 (gm Input File: create/update).
  - b. Pick item number 4 (Create gm input file from system defaults). You will be prompted for the "New gm input file", type yyy.in. A list of parameters will be displayed in menu format. Change:
    - 1) Sequence  
Enter "yyy.in" at the file name prompt.
    - 2) Output  
You will be prompted for the exon map file and the default will be yyy.emap. Hit return.  
You will be prompted for the translation file and the default will be yyy.trans. Hit return.
  - c. Pick item number x (Write file & exit).

This will give you initial values to run with. You may start up gmwin and enter yyy.in in the Input File slot, or you may run "gm yyy.in" from a text terminal.

#### V. TUNING gm FOR A DIFFERENT ORGANISM

-----

Different user's in a lab may work on different organisms. gm is compiled with the defaults set for a specific organism. If you are working on a different organism:

- (a) Copy the appropriate initial input file {worm.in, dicot.in, monocot.in, fly.in or human.in} from the appropriate directory (See Files).
- (b) From the main menu:
  - Pick item number 3 (gm Input File: create/update).
  - Pick item number 5 (Create gm input file from existing file).
  - and use the file from (a) for the old file.
  - Change filenames as described above.

If you are working on an organism that has different characteristics than human, worm, fly, or plants, please send mail to

cari@nmsu.edu

and we will create the necessary matrices for you.

#### VI. ADJUSTING PARAMETERS

-----

If the input parameters are set too low, gm may generate enormous amounts of output. gm will stop generating intermediate maps at 10,000, and will then try to complete these maps. Regardless of the cutoff, this is a lot of output to look at and will take a long time to run.

The default values have high cutoffs for the score of each entity (i.e. splice sites, exons, introns, atg, and aataaa). In our tests, these parameters created a small set of initial exon maps in a small amount of time (i.e. under 3 minutes). This does not guarantee that you will get good results and fast execution time with these parameters; your sequence may be inherently more difficult and consequently, take a long time. On the other hand, you may not get any results from these parameters. We recommend the following:

1. Start with the supplied parameters.
  - A. If gm runs too long (i.e. over an hour or until you get impatient), cntl-C out of gm (or gmwin). Enter menu, and reset the cutoffs at higher values, and run gm again.
  - B. If you do not get any exon maps, check the statistics in the bottom window of gmwin, or in the output text file; find the entities for which there is a low number. Enter menu, reset the respective cutoffs at a lower value, and run gm again.
  - C. Continue adjusting parameters until you get good results.
2. Always run initially with the option "Longest non-inclusive maps". Once you have identified the area of interest, run gm with the "All emaps" options to see the alternative spliced exon maps.
3. Partial exon maps can be attained by lowering the cutoff on

ATG and AATAAA to get partial maps. The analysis option to generate partial exon maps is currently not available.

You do not have to exit gmwin to alter the parameters; enter menu through the menu button on gmwin, or start up another window and run menu. After changing the desired parameters, enter "w" and the input file will be written. You may then re-enter the input file name in the input file slot of gmwin. In other words, you can simultaneously run gmwin and menu; and alternate between adjusting parameters and running gm.

The windowing option detects coding regions and only looks for splice sites in those regions. The current windowing routine is the same as the exon evaluation; if the exon evaluation is given a matrix of size 6, this takes a long time to run over a large sequence, and it can be cause some good splice sites to be missed. So, until we have a better way to detect regions of gene activity, the following kludge was used: a cutoff score of -99.0 turns the windowing off; i.e. the whole sequence is automatically used.

CHANGING PARAMETERS CAN CAUSE UNEXPECTED RESULTS: the best way to explain this statement is through an example. I ran gm on an input file; gm predicted most of the gene, but it missed a few 5' splice sites. I lowered the stringency on 5' splice sites and ran gm again. gm predicted "less" of the gene than it predicted in the first run. Common sense says that lowering the stringency on splice sites should cause gm to predict more candidate genes. The reason it predicted less is as follows:

By predicting more 5' splice sites, the candidate exons ended up to be much shorter. There is a parameter for the "Protein Coding Capacity"; i.e. gm will delete any predicted genes shorter than this user supplied parameter. Therefore, with more 5' splice sites, the resulting genes ended up shorter than this parameter (which was set at 300). I changed this parameter to 0, and I got all the predicted coding region that I expected (and it took much longer to run). So, the moral is: beware of the effect of changing one parameter may have on another parameter.

## VII. BUGS

Files must either be in your current directory, or in a directory specified by GMSEARCHPATH.

If given an incorrect file (wrong file format for any of the input files), gm ( or gmwin or menu) may die ungracefully, or may give garbage output. That is, there is not perfect error checking on the input files, so if your execution is faulty, check your input files.

In gmwin, the menu button does not bring up a window when running with openwindows. You may simulate this event easily by: when running gmwin, bring up another window and run menu.

Highlighting is incorrect if there are blanks at the end of any sequence line.

## VIII. ADDITIONAL DOCUMENTATION

On-line help is provided in menu. Specifically:

- 1) The pattern matching routines used by gm are supplied stand-alone under the "Interactive Tools" item. Under the "Help" item for interactive tools is a description of each routine.

2) The files formats are described under the main menu help.

Good luck and have fun,  
Cari Soderlund