OBJECTIVES

This proposal seeks to identify and characterize exonic splicing enhancers (ESEs) in the *Arabidopsis thaliana* genome, and to use knowledge of these enhancers to improve gene annotation. Understanding the function of all *Arabidopsis* genes (the goal of the 2010 project) will require knowledge of their protein sequence, often deduced from the intron/exon structure. This proposal will combine the distinct expertise of the PIs in bioinformatics, in the biochemical machinery of splicing, and in *Arabidopsis* genetics, to predict and then validate new exons, alternatively spliced genes, and exon splicing enhancers (ESEs) in *Arabidopsis.* Newly discovered ESEs and exons will be used to improve gene finders for *Arabidopsis* and other plants. A database of ESEs linked to the genes in which they appear will be released via the Internet.

The underlying hypothesis of this work, which is supported by extensive data in animal systems (Blencowe 2000), is that the sequence-specific recognition of splicing enhancer sequences is a major determinant of splice site selection generally, and plays a dominant role in the regulation of alternative splicing. The availability of the complete genomic sequence and extensive experimental data from *Arabidopsis thaliana* now makes possible a systematic experimental approach to discovering splicing enhancer sequences.

First, the sequence elements that are likely to act as splicing enhancers will be identified computationally. Then, the activity and tissue-specificity of these enhancers will be tested experimentally in transgenic plants. Finally, information from validated splicing enhancers will be used to improve genefinding software. This project will greatly improve our understanding of splicing signals that act at a distance from splice sites, and will specifically result in improved genefinding tools for plant genomes. Furthermore, we expect that these ESEs may be recognized in a tissue-specific manner. This specificity will also be described, and transgenes exhibiting these patterns of expression will be retained and made available to the community for use as reporters of splicing regulation.

Specific Aim 1) Candidate exonic splicing enhancer sequences will be identified computationally.

A database containing the complete Arabidopsis genome and all its predicted proteins has been developed at TIGR, and is currently being refined and curated further. This database is structured so that genes with experimental evidence can be extracted with relative ease. In addition, a considerable amount of cDNA data exists for Arabidopsis, and much more is expected in the next few years. We propose a combination of methods to mine this database of experimentally supported gene structures for potential ESE motifs. These methods follow from the assumption that ESE motifs will be over-represented in exons relative to introns, and will show some conservation. A method that corrects for amino acid frequency and codon bias will be applied. The distribution of spacing between potential motifs and splice sites will be studied, as will the pattern of silent nucleotide conservation among pairs of homologous *Arabidopsis* genes.

Specific Aim 2) The ability of candidate motifs to function as splicing enhancers will be tested experimentally.

First, enhancer-dependent intron-exon-intron units will be generated, and their enhancer-dependence will be verified experimentally. Specific sequence elements will then be tested for enhancer activity and defined by directed mutation. For this purpose, enhancer-dependent exons will be placed in a test vector together with their flanking introns. Incorporation will be assessed by RT-PCR, and by production of the marker protein ß-glucuronidase (GUS). To control for many sources of potential error, we plan to use a pairs of vectors such that expression of the GUS marker from one vector will require exon skipping while expression from the other vector will require exon inclusion. Examination of GUS expression in transgenic plants, backed up by direct analysis of RNA products, should result in a comparative description of the specificity of action of ESE motifs.

Specific Aim 3) Incorporate information on ESEs into genefinding software.

The experiments conducted for Specific Aim 2 will yield new information about sequences that function as splicing enhancers. We will incorporate these new ESEs into two systems already in use for Arabidopsis genome annotation, GeneSplicer (Pertea et al. 2001) and GlimmerM (Salzberg et al. 1999). These systems are part of the TIGR automated annotation pipeline, and their output is also displayed in our graphical annotation editor, which is used for manual curation of the genome.

RESULTS FROM PRIOR NSF SUPPORT (Salzberg)

Steven Salzberg is currently supported by NSF grants KDI-9980088, "Intelligent Computational Genomic Analysis," and IIS-9902923, "Interpolated Markov Models for DNA Sequence Analysis," both of which are in their second year out of 3. These awards jointly support the PI and several students, postdocs, and scientific staff, including: (1) Mihaela Pertea, 4th-year Ph.D. student at Johns Hopkins, (2) Mihai Pop, Ph.D., Bioinformatics Scientist at TIGR, (3) Maria Ermolaeva, Ph.D., Staff Scientist at TIGR, and Natalia Volfovsky, Ph.D., Postdoctoral Scientist at TIGR. Co-PIs on the grants are Simon Kasif at Boston University (KDI-9980088) and Arthur Delcher (IIS-990293) at Loyola College of Maryland.

Both of these project are developing new computational systems for analysis of DNA sequences. Grant IIS-9902923 is focused particularly on further developing and extending Interpolated Markov Models (IMMs), a technique first introduced to the realm of DNA sequence analysis by the PI. IMMs are a generalization of fixed-order Markov chains in which subsequences of different lengths are combined to compute a probability. They are the basis of the Glimmer system for microbial gene identification. In addition to improving the Glimmer system, this project plans to extend the framework so that it can work on organisms other than bacteria. These other organisms include the malaria parasite, *P. falciparum*, the model plant *A. thaliana*, and other higher organisms.

Both projects are investigating new algorithms for comparison of sequence data at the whole-genome level. Previous work focused on gene-level analysis, but with the rapid rise in the number of completely sequenced genomes, new technology is essential. The research focuses on two problems: alignment of whole genomes or chromosomes, and identification of repeated sequences. Work is under way to adapt efficient algorithms based on suffix trees to solve both of these problems. By looking for conserved regions in alignments between genomes of related organisms, one can help identify genes and regulatory sequences associated with them.

Thus far under these two awards, a major new release of the Glimmer system (version 2.0) has appeared along with a paper (Delcher et al., 1999) describing its improvements. The new release has already been distributed free of charge to Glimmer's subscribers, which now include over 400 major universities, government labs (including many NIH and DOE labs), and nonprofit institutions. Glimmer 2.0 finds over 99% of all genes in a bacteria or archaeal genome fully automatically, and continues to be the state-of-the-art gene finder for such organisms (Fraser et al., 2000). More information on Glimmer is available at its website, http://www.tigr.org/softlab/glimmer/glimmer.html. During the past two years, Glimmer has been used as the principal gene finder for several major microbial genome sequencing projects. A partial list includes the human pathogens Vibrio cholerae (Heidelberg et al., 2000), Neisseria meningitidis (Tettelin et al., 2000) Chlamydia trachomatis MoPn and C. pneumoniae AR39 (Read et al., 2000); the radioresistant bacterium Deinococcus radiodurans (White et al., 1999), and the extremophile Thermotoga maritima (Nelson et al., 1999).

The push to develop eukaryotic versions of IMM-based gene finders led to the malaria gene finder, GlimmerM (Salzberg et al., 1999; Pertea et al., 2000; Salzberg, 1999). More recently, we have developed new versions of this system for Arabidopsis, rice, and *Theileria parva*. The malaria, rice, and Arabidopsis versions are available through the GlimmerM web server; in addition, the code is freely available for download to researchers at nonprofit institutions. Active efforts are under way to improve GlimmerM and to train it on additional organisms. GlimmerM

has already been used as one of the primary de novo gene finders for the annotation of the Arabidopsis genome (The Arabidopsis Genome Initiative, 2000; Theologis et al., 2000).

The research on whole-genome alignment and repeat analysis has led to a new system for whole genome alignment, MUMmer (Delcher et al., 1999b), which has already been released and distributed for free very widely. The suffix tree algorithms at the core of MUMmer are now being used to find all repeats in large genomic sequences, and have been used to develop and release new repeat databases for Arabidopsis and rice, with other organisms intended for the future. Using this system, we recently discovered that large-scale chromosomal inversions appear to be a ubiquitous event in bacterial genomes, and an important force in evolutionary change (Eisen et al., 2000).

We are also dedicating effort to development of new statistical methods for analysis of genomic sequences, focused recently on two problems: identification of transcription terminators (Ermolaeva et al., 2000), and identification of operons (Ermolaeva et al., 2001). Both these methods have been applied to a large number of complete bacterial genomes and new and databases and software have been released as a result (http://www.tigr.org/softlab/transterm.html).

An unexpected result of this research was the development of a new combinatoric method for closing gaps in a genomic sequence. For a typical project, 99% or more of the genome is completely sequenced at the end of the first 'random shotgun' sequencing phase. The sequence at that point exists in many separate pieces, however, and the order and orientation of these pieces is usually unknown. Closing the gaps between the pieces is a difficult and time-consuming task. Working closely the genome sequencing teams at TIGR, the PI invented a new method, POMP (Tettelin et al., 1999), for closing gaps that saved substantial time and effort and has now been adopted for most large genome projects at TIGR.

Results From Prior NSF Support - Mount.

This is a new proposal, and P.I. Mount last had NSF support in the form of a Presidential Young Investigator Award in effect between July of 1987 and June of 1993. However, prior research that documents past experience on the part of the P.I. with relevance to this proposal is nevertheless summarized here.

Bioinformatics: Genes for splicing factors.

Kumar and Mount (in preparation) examined the distribution and conservation of the SR splicing factor protein family in the *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster* and human genomes and expressed sequence databases. These species appear to have at least 1, 2, 6, 20, 6 and 9 SR proteins, respectively. Thus, *Arabidopsis* has twice as many distinct SR proteins as does any other species. Specific SR protein subfamilies appear to be evolutionarily ancient, and widely distributed. For example, worms, flies, plants and humans all have one or two members of a subfamily that includes the human ASF/SF2 protein.

Bioinformatics: Analysis of *Drosophila* introns in the databases.

The P.I. has firsthand knowledge of several methods for identifying signals in DNA sequence, including insight into their advantages, disadvantages and underlying theory. In a collaboration between the P.I. and a number of computer scientists, a database of 209 *Drosophila* introns was extracted from GenBank and examined by a number of methods for features that might serve as signals for messenger RNA splicing (Mount et al., 1992). A tight distribution of intron sizes was observed. Although more than half are less than 80 nucleotides in length, most of these have lengths in the range of 59-67 nucleotides, and the shortest is 51 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. Of greatest relevance to the present proposal, potential *Drosophila* branch site signals were examined by a number of methods. It was found that the *Drosophila* branchpoint consensus

matrix resembles CTAAT (in which branch formation occurs at the underlined A), and differs from the corresponding mammalian signal in the absence of G at the position immediately preceding the branchpoint (Fig. 1).

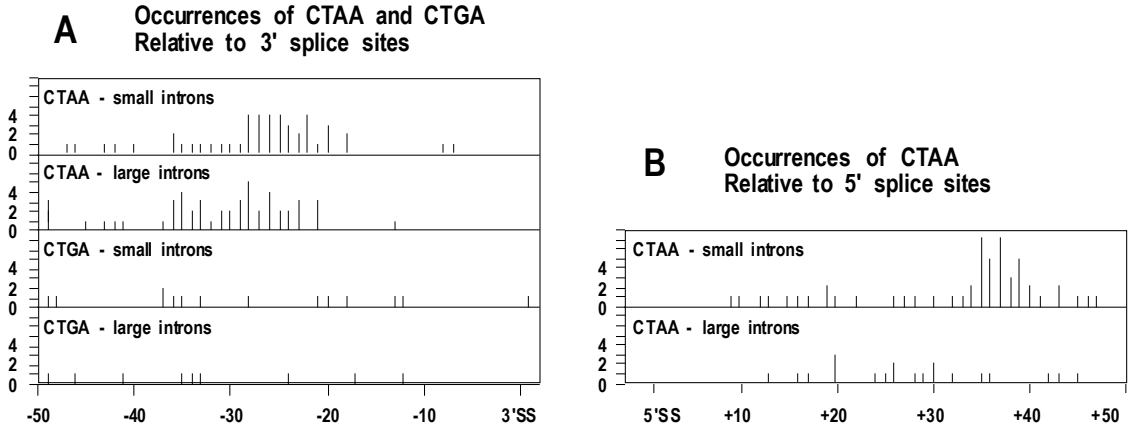

Fig. 1. Distribution of words of length 4 related to the vertebrate branch site consensus in large and small *Drosophila* introns.
All occurrences of the indicated tetranucleotide in all introns of the indicated database are plotted relative to the splice site (from Mount et al., 1992).

Experimental: A mutational analysis of splicing signals in *white-apricot*

The *Drosophila melanogaster white-apricot* allele carries a 5.1 kb *copia* retrotransposon insertion in the small (74 nt) second intron of *white.* The mutant allele produces a variety of aberrantly processed RNA products, and roughly 20-fold less *white* mRNA. This residual expression of *white-apricot* is affected by extragenic mutations at suppressor and enhancer loci that encode proteins involved in the processing of messenger RNA precursors. In order to improve our understanding of how this *copia* insertion interferes with the production of *white* mRNA, we examined the expression of a number of mutant transgenes derived from *white-apricot*. We found that 1,124 nucleotides at the 3' end of *copia*, including an inefficiently used 3' splice site and the *copia* polyadenylation site, are sufficient to reproduce the *white-apricot* phenotype. Inactivation of the *copia* 3' splice site increases the production of *white* mRNA, but a mutation that makes the same site 100% efficient has no effect on *white* mRNA levels. The *white* intron branch site is disrupted by the *copia* insertion. Restoration of this site to consensus results in efficient removal of the intron and copia sequences by splicing and complete reversion of the eye color phenotype. These results implicate multiple RNA processing sites in the genesis of the *white-apricot* phenotype. We are currently preparing a manuscript on these results.

Experimental: Genetic analysis of two genes involved in pre-mRNA splicing.

Much of the recent research in the P.I.'s laboratory has been an NIH-funded project on the genetic analysis of two pre-mRNA splicing factor genes, *U1-70K* and *B52*.

We cloned and characterized the *Drosophila* gene for a major U1 snRNP protein known as U1 70K (Mancebo *et al*. 1990). A mutation in this gene, obtained by examining single *P* element insertion lines available from the Berkeley Drosophila Genome Project gene disruption project (Spradling et al. 1999) is an embryonic lethal, showing that U1 70K is an essential gene. Of particular interest is our observation that a U1 70K gene lacking the arginine-rich domain can restore viability and fertility to this embryo. To address the possibility that this surprising result could be due to the fact that the original allele was not null, we have used EMS mutagenesis to isolate a nonsense mutation and male recombination (or "hybrid element insertion," Preston et al. 1996) to isolate a number of deletions. Recent studies with these new null alleles confirms that the arginine-rich domain of U1-70K protein is not required for viability in flies, and these results are now being prepared for publication. One can deduce from these results that interactions involving this domain cannot be essential for splicing or for accurate splice site selection.

Although the SR protein B52 is an essential gene encoding a protein that can function in splicing, it has not been possible to observe splicing defects in B52 null animals (Peng and Mount, 1995; Ring and Lis, 1994). Recently, we have generated clones of cells lacking *B52* genetic function at various times in development using the FLP - FRT somatic recombination system. Clones of *B52*-deficient cells induced early in development (before 48 hours) are inviable. Clones induced later than 48 hours cause striking eye defects including the loss or duplication of sensory bristles and a loss of photoreceptor cells. However, the initial stages of ommatidial development appear to be normal, as indicated by the expression of the photoreceptor cell-specific markers *elav* and *glass* in eye discs during the third larval instar, and the cone cell marker *cut* during pupal stages. These results suggest a specific role for *B52* (as opposed to other SR protein genes) late in eye development. We are currently preparing a manuscript on these results.

Summary: The P.I. has a combination of experience in bioinformatics and pre-mRNA splicing that places him in a unique position to carry out this proposal.

BACKGROUND AND SIGNIFICANCE

Accurate gene annotation in the absence of reliable experimental evidence remains difficult to achieve. During the course of the *Arabidopsis* genome sequencing project, investigators at TIGR evaluated the accuracy of the leading gene finders for Arabidopsis. To evaluate accuracy, we collected a set of 1131 genes from chromosome II of Arabidopsis (Lin et al., 1998) which had either full-length cDNAs or full-length homology to genes from other organisms. This dataset was assumed to represent "truth" for the purposes of comparison. This study (unpublished) showed that Genscan+, a version of Genscan (Burge and Karlin, 1997) trained specifically for Arabidopsis, was the most accurate *de novo* gene finder; it also showed that even Genscan+ predicted correct intron/exon structures for less than 50% of the genes. Similar degrees of success have been achieved in other species (e.g. *Drosophila*; Reese et al., 2000). Clearly, an improved understanding of the rules that govern splice site selection will lead to improvements in our ability to find genes in all organisms.

Pre-mRNA splicing

The boundaries between exons and introns are referred to as splice sites. The splicing reaction itself consists of two consecutive transesterification reactions involving an intermediate in which the intron at the 5' splice site (or donor site) is joined to a site within the intron known as the branch site. Splicing is carried out by the spliceosome, a large macromolecular machine which forms around each intron. The assembly of the spliceosome proceeds through a series of distinct intermediates by steps that involve the recognition of splicing signals in pre-mRNA through interactions with specific splicing factors. These factors include proteins and small nuclear ribonucleoprotein particles (snRNPs). Early in the process of spliceosome assembly, specific sites on the mRNA precursor are bound by highly conserved components of the spliceosome. In particular, the U1 snRNP associates with the 5' splice site, and U2 auxiliary protein (U2AF) and splicing factor 1 (SF1) associate with sequences upstream of the 3' splice site, including the branch site. Although there are a number of distinct steps required for the formation of a functional spliceosome, the recognition of known splicing signals occurs early and it is likely that the assembly of these early factors determines the outcome of splicing in most cases. Components of the spliceosome are in general very highly conserved among all eukaryotes, including *Arabidopsis* (see, for example, Mount and Salz 2000).

The two steps of splicing define three sites at which phosphoryl transfer reactions take place: the 5' splice site, the branch site, and the 3' splice site. Most introns show some similarity to each other at these three sites, and the nucleotide sequences at these sites contribute to the recognition of intron boundaries. For example, it is almost always the case that intron can be represented by a sequence beginning GT (GU in the RNA) and ending AG.

Splice site selection

In addition to the conserved GU dinucleotide, 5' splice sites are generally similar at the last three exon nucleotides and the first 7 intron nucleotides adjacent to the site of splicing. This

similarity can be represented by a consensus sequence. Simpson and Filipowicz (1996) provide the consensus sequence AG|GUAAGU and a detailed matrix (not shown) for plant 5' splice sites. Although this short description of the consensus is identical to the 5' splice site consensus sequence for animals, minor differences in the frequencies of specific nucleotides at specific position exist, and are important for accurate identification of splice sites in a particular species (e.g. Pertea and Salzberg 2001). Interaction between the U1 snRNP and the pre-mRNA is mediated by basepairing between the 5' end of the U1 snRNA and the 5' splice site, and the sequence of U1 is also highly conserved. In fact, the most conserved region lies near the 5' end of the RNA. U1 snRNAs from all species known (including plants, fungi, and a variety of animals) have the sequence ACUUACCUG at, or very near to, the 5' end.

Although recognition of 5' splice sites during the initial stages of splicing is primarily accomplished by the U1 snRNP, the U1 snRNP does not act in isolation. Not all sites bound by the U1 snRNP are ultimately used as 5' splice sites. Factors bound to other sites on the pre-mRNA, some of which are discussed below, can promote binding between U1 snRNP and the 5' splice site, or can facilitate progression along the pathway towards splicing. Furthermore, the 5' splice site is later "examined" by additional factors in the course of splicing. One such factor is U6 snRNA, which displaces U1 prior to the first catalytic step, and remains associated with the 5' splice site throughout the remainder of the splicing reaction. In general, initial selection of the 5' splice site by the U1 snRNP is influenced by other factors, and must be followed by appropriate interactions between the 5' splice site and other components of the spliceosome.

The 3' splice site usually occurs immediately 3' of the trinucleotide CAG or UAG (and occasionally AAG, but never, or almost never, GAG). In addition, there is some conservation of the first nucleotide of the exon adjacent to the 3' splice site, bringing the number of conserved 3' splice site nucleotides to four. In plants, but not animals, position -4 also shows some conservation, and a G-to-A mutation at this position is responsible for the *cop1-1* mutation (Simpson et al. 1998).

How is it possible for the 3' splice site to be specified by so few nucleotides? The answer to this question lies in the sequence immediately upstream of the 3' splice site, including the branch site, which is recognized together with the 3' splice site. In the introns of multicellular animals, a pyrimidine tract in this region is bound by the large subunit of U2AF, while the small subunit of U2AF binds to the 3' splice site itself. U2AF is a dimeric protein which then recruits the U2 snRNP to the branchpoint. In addition, the branch site itself can be recognized by SF1. In the case of plants, the pyrimidine tract is not always present, but U2AF is nevertheless conserved and is presumably essential for splicing. As with the U1 snRNP, recruitment of U2AF by other factors plays an important role in the selection of splice sites.

Splicing enhancers

Sequences have been identified in several mammalian genes that reside at variable distances from splice sites yet are required for splicing to occur, either *in vivo* or *in vitro*. Although such splicing enhancers have been identified in both exons and introns, exonic splicing enhancers are generally better characterized, and are probably more common. Such exonic splicing enhancers (ESEs) activate nearby splice sites (both 5' and 3' splice sites) and promote the inclusion (vs. skipping) of exons in which they reside. Initially, ESEs were recognized as purine rich motifs containing repeated GAR (GAA or GAG) trinucleotides. However, many other sequences have now been shown to have enhancer activity (see Tacke and Manley 1999 for review). A small number of well-defined enhancer-dependent splicing events (notably the immunoglobulin M and *Drosophila* doublesex introns) have been used by researchers in the field to define and characterize numerous splicing enhancers (Tacke and Manley, 1999). The female-specific *Drosophila doublesex* intron is enhancer-dependent, and requires an enhancer in the exon downstream of the 3' splice site in order to function in HeLa cell nuclear extracts. In this mammalian system, it is often observed that an enhancer identified in one assay (e.g. IgM) is functional in another (e.g. *doublesex*).

Many exonic splicing enhancers are bound and activated by one or more of several related splicing factors known as SR proteins. SR proteins contain either one or two RNA-binding domains and "RS" domains that are characterized by numerous arginine-serine dipeptide repeats.

SR proteins are not only essential for splicing, but also for each of the first three recognizable steps of spliceosome assembly. In vitro, any one of the several SR proteins can restore splicing to a splicing extract lacking SR proteins. Thus, the essential functions of individual SR proteins in splicing are at least partially redundant. However, there is considerable specificity to the activation of splicing by SR proteins through exonic splicing enhancers. Individual SR proteins differ with respect to the sequence-specificity of their RNA-binding domains, and with respect to their ability to recognize and activate different exonic splicing enhancer sequences (e.g. Liu et al 1998, Schaal and Maniatis 1999b).

The relationship between sequence-specific binding by SR proteins and the activation of splicing by exonic splicing enhancers is complex and incompletely understood. Both restoration of splicing and activation of some, but not all, enhancer-dependent splicing events by an SR protein lacking the RS domain has been reported (Zhu and Krainer 2000). Conversely, recruitment of the RS domain of SR proteins to an RNA by means of an unrelated RNA-binding domain has been also reported to promote enhancer-dependent splicing (Gravely and Maniatis 1998). Of greatest significance to this proposal are the observations that SR proteins show tissue-specific patterns of expression, and that different SR proteins work through different sequences (Liu et al. 1998). These facts support a model in which both constitutive splicing and enhancer-dependent splicing are dependent upon SR proteins bound to RNA. Although only a dozen or so splicing events have been shown to be enhancer-dependent, the existence of exonic splicing enhancers (ESEs) within constitutively spliced exons (Schaal and Maniatis 1999a) suggests the possibility that ESEs are ubiquitous, redundant, and required for all splicing events. It is estimated that as many as 15-20% of random sequences 20 nt. long contain a splicing enhancer (Blencowe 2000). Thus, it appears likely that many sequences may act as splicing enhancers. What is clear is that the motifs recognized by SR proteins are short and degenerate. Examples of these motifs (from Tacke and Manley 1999) are SRSASGA (where R=A or G; S=C or G) for ASF/SF2 and UGCNGYY (where Y=C or U and N is any base).

Exonic splicing enhancers (ESEs) in plant pre-mRNA splicing.

Early research on plant pre-mRNA splicing emphasized the role of AU-rich or U-rich sequences within introns (see Simpson and Filipowicz 1996; Brown and Simpson 1998; Schuler 1998). It is clear that U-rich sequence elements play important roles in intron definition, and that exon-skipping is a less common outcome of mutagenesis than is true in animals, where introns tend to be larger and exon definition is probably a more common mode of splice site selection (Berget 1995). On the other hand, a number of recent results have shown a role for exon sequences in the selection of plant splice sites (Egoavil et al. 1997; Simpson et al. 1998; Simpson et al. 1999; McCullough et al. 1996).

There are compelling reasons to believe that ESEs play an important role in plant splicing. SR proteins, the mediators of ESE activity in vertebrates, are highly conserved in plants. This pattern of conservation includes reactivity with the monoclonal antibody mAb104 (Lopato et al 1996) and extends to function. A mixture of *Arabidopsis* SR proteins (Lopato et al. 1996), and atRSZp22 in particular (Lopato et al 1999a) can complement SR-deficient mammalian splicing extracts. Furthermore, plant SR proteins can influence splice site choice in mammalian nuclear extracts (Lazar et al 1995), and regulate alternative splicing in vivo (Lazar and Goodman 2000; Lopato et al. 1999). Our own phylogenetic analysis of SR protein genes in complete genomes (*Saccharomyces, Schizosaccharomyces, Caenhorabditis, Drosophila, Arabidopsis* and *Homo*; Kumar and Mount, in prepararion) shows that Arabidopsis has more SR protein genes than any other known organism by a factor of two (20 genes in *Arabidopsis* vs. 9 in humans). Some of these genes encode highly similar pairs. In at least one case, that of the two SF2/ASF homologues atSRp30 and atSRp34/SR1, the two genes are expressed in distinct patterns during development (Lopato et al. 1999b), suggesting functional differentiation.

In a preliminary analysis, we searched our database of 5249 exons from 1131 confirmed genes in chromosome II of Arabidopsis (see above) to identify all over-represented short oligomers. The most common 9-mer in this set matches the consensus [AG]AAGAAGA[AG], which is a near-perfect match to the known ESEs in Drosophila. The bulk of these 9mers occur

40-80bp from either the 3' or the 5' end of the exon, a fact that is consistent with their function as ESEs.

No splicing event in *Arabidopsis* has yet been shown to be enhancer-dependent. However, exon sequences have been shown to contribute to splice site selection in the manner expected of an ESE (Egoavil et al. 1997), and the absence of documented enhancer-dependence can be readily attributed to the fact that detailed mutational analysis is necessary to show enhancer-dependence. There is also ascertainment bias. Very often the effects of mutations in exons are attributed to missense or nonsense when they may also effect splicing. For example, Liu et al. (2001) have shown that a nonsense mutation causing the skipping of BRCA1 exon 18 affects splicing in vitro, and that a miss-sense mutation at the same position also causes exon skipping. These same authors present a statistical analysis of 50 mutations that cause exon skipping in vivo (based on a list compiled by Valentine, 1998) that supports the possibility that they are mutations in ESEs. It is quite possible that effects of many mutations in plant ESEs on splicing have been missed merely because RNA was never examined. Indeed, several recent reviews emphasize the possibility that (human) mutations in ESEs may be much more common than is recognized (Blencowe 2000, Mount 2000, Maquat 2001).

Computational splice site identification and genefinding

Our group has developed a computational gene finder, GlimmerM, which has been applied to *Arabidopsis*, *Plasmodium falciparum* (the malaria parasite), rice, and the parasite *Theileria parva* (Salzberg et al., 1999). GlimmerM uses interpolated Markov models (IMMs) to identify coding regions. IMMs form the basis of the Glimmer system for finding genes in bacteria, archaea, and viruses. Glimmer correctly identifies approximately 98-99% of the genes in bacteria without any human intervention, and with a very limited number of false positives (Salzberg et al., 1998; Delcher et al., 1999). For eukaryotes, the accuracy of all gene finders is considerably lower, but the GlimmerM system has been relatively quite successful. In an evaluation on *P. falciparum* chromosome 2 using genes confirmed by homology, GlimmerM found 98 out of 113 (87%) of the genes exactly; i.e., the exon/intron structure was exactly correct.

In Arabidopsis, where genes tend to have more exons, the accuracy of GlimmerM is considerably lower. In a recent evaluation comparing GlimmerM, Genscan+, and Genemark.HMM, we measured each system's performance on our benchmark database of 1131 confirmed genes from Arabidopsis chromosome II. The results of this evaluation are shown in Table 1.

As the table (next page) shows, GlimmerM performed somewhat better than the other systems at putting together a complete gene model, getting 558/1131 (49%) genes exactly correct. The table also shows how many predicted exons were exactly correct, and how many predictions got at least one of the exon borders (i.e., either the 3' or the 5' edge) correct. GlimmerM found 3754/5249 (72%) of the exons correctly, but (as expected) a smaller percentage of complete gene models are correct.

Table 1: Gene finder performance on A. thaliana

Gene Finder	True Exons	Predicted Exons	One Border	Exact Exons	Exact Genes
GlimmerM	5249	4997	3815	3754	558
Genscan+	5249	4967	4100	3715	495
Genemark.HMM	5249	5088	3753	3276	337

We have developed a different algorithmic technique for splice site identification (Pertea and Salzberg, 2001), using Markov chains and decision trees. This technique is incorporated into GlimmerM and is available as a standalone system. Within GlimmerM, it is run as a pre-processing step to identify all candidate splice sites. Improvements in the accuracy of this

algorithm, called GeneSplicer, have a dramatic impact on the overall accuracy of gene finding. GeneSplicer is described in much more detail by Pertea and Salzberg (2001).

Briefly, GeneSplicer computes a score for each splice site based on two criteria. First, it constructs a maximal dependence decomposition tree (Burge, 1998) separately for the acceptor and donor sites. These MDD trees are constructed from a set *D* of *N* aligned DNA sequences of length *k*, extracted from a set of donor (respectively acceptor) sites. For each of the k positions, let the most frequent base at that position be the consensus base. The variable C_i will be 1 if the nucleotide at position *i* matches the consensus at position *i*, 0 otherwise. Next, compute the X^2 statistics between the variables C_i and Y_j, for each *(i,j)* pair with $i \neq j$. If strong dependencies are detected between non-adjacent positions, then proceed as in (Burge, 1998) by partitioning the data into two subsets based on those positions. Recursive application of this method builds a small tree. At the leaf nodes of this tree, we create a Markov chain model of the sequences around the donor (acceptor) site using 0^{th} or 1^{st}-order probabilities; these Markov chains are used to score potential splice sites.

The research proposed below has strong potential to improve both GeneSplicer and GlimmerM for Arabidopsis and for other organisms. Because we plan to make both our data and our code freely available to other researchers, any improvements will also be available to the general scientific community.

EXPERIMENTAL PLANS

Specific Aim 1) Identification of candidate exonic splicing enhancer sequences computationally.

All known ESE motifs are short (less than 10 nt.) and highly degenerate, meaning that many variants appear to function. Some examples are given above. In combination with the fact that the location of the ESE relative to splice sites is variable, this means that they cannot be identified by simply tabulating nucleotide frequencies aligned with the splice site. Fortunately, the availability of abundant sequence information means that statistical analysis can be used to detect rare and variable signals. Just as the first dozen or so gene sequences allowed the identification of splice site consensus features that are now the standard tool of existing gene finders, the current, much larger, database of genomic and cDNA sequence can be mined for the second-order sequence features that the cellular machinery must use for splice site selection.

The genome of *Arabidopsis* has now been sequenced and annotated (The *Arabidopsis* Genome Initiative 2000), and we (TIGR) maintain a database of annotated genes including sequence from all laboratories internally and on our external website (see http://www.tigr.org/tdb/ath1/htmls/ath1.html). A considerable amount of cDNA data exists for *Arabidopsis*, including a large data set recently provided by Ceres, Inc. (see press release at www.tigr.org/new/press_release_ceres.shtml). We propose to use a combination of methods to mine this database for potential ESE motifs.

1a) Correction for codon bias.

If ESEs function in exons and act to define exons, then sequences that act as ESEs should be over-represented in exons. As described above, our preliminary study using only chromosome II genes already found that the most over-represented 9mer in exons does indeed correspond to the known ESE motif from vertebrates. Our first and most obvious step will be to extract from the Arabidopsis database a set of genes confirmed by either full-length cDNAs or full-length protein sequence homology, and to compile a list of the most over-represented short oligomers in this set. Selection for the encoded amino acid sequence makes it difficult (or impossible) to identify ESEs as conserved blocks in the same way that transcriptional enhancers can be recognized as conserved blocks of sequence upstream of the start site of genes. However, it is possible to statistically correct for both amino acid composition and codon bias (Antezana and Kreitman 1999). In brief, the frequencies of sequence motifs in the actual sequence of a gene are compared to the frequencies expected summing ove all sequences with the same amino acid sequence and codon bias. Antezana and Kreitman observed that there are strong reading frame-independent forces that affect codon preferences. Thus rather than using simple frequencies, we will use

relative frequencies for single bases, dinucleotides, trinucleotides, and longer oligomers (up to octamers). Such a statistical analysis will be greatly facilitated by the large amounts of genomic sequence avilable in *Arabidopsis.*

1b) Use of homologous gene families.

If ESEs function to control splicing, then they should be conserved; base changes that affect ESE function will be less likely to be fixed during evolution than base changes that do not. To examine conservation, we will look at groups of homologous genes that share their exon-intron structure. In fact, 65% of *Arabidopsis* genes belong to a gene family containing two or members (using a BLASTP value $E < 10^{-20}$), so the majority of the protein-coding genes can be used in this analysis (AGI, 2000). As part of our comprehensive annotation of Arabidopsis, we (TIGR) are in the process of compiling a carefully curated set of gene families. Each of these families contains all the genes within the genome that are paralogous based on sequence comparisons. We will extract a subset of these genes representing close paralogs in order to calculate the likelihood of silent versus nonsilent mutations. These close paralogs must be similar enough that any positions that have undergone mutation would have undergone, on average, less than one nucleotide change. (If a position has mutated from A to G and then from G to C, for example, it will be indistinguishable from a single mutation from A to C.) Based on these paralogs, the frequency of observed silent nucleotide substitution will be examined in codons, and compared to an expected number that incorporates both the overall rate of substitutions of that exact type (e.g. CUG to CUA) and the overall rate of synonymous substitution in the gene pair being examined. Once again, it is expected that the large data set will allow a relatively subtle signal to be apparent.

A significant new resource that will appear during the course of 2001-2002 is genomic sequence data for *Brassica oleracea*, a close relative of *Arabidopsis*. Current plans are to begin a 1x coverage shotgun sequencing project (participated in jointly by TIGR and Cold Spring Harbor Laboratory) with the express purpose of using this sequence to identify small genes in *Arabidopsis*. Small genes are often missed by BLAST and by *de novo* gene finders, and the sequence of a close relative will provide a powerful tool for finding these genes. Likewise, any other short conserved sequences may also be discovered by inter-genome comparisons. We will align the *B. oleracea* data with *A. thaliana* to identify short conserved stretches of DNA. In addition to novel short genes, we will also use this data to search for conservation at silent positions that might represent ESEs.

1c) Spatial distribution of candidate motifs.

Because ESEs are expected to be over-represented at the appropriate distance from splice sites (Graveley et al. 1998), the spatial relationship between candidate ESEs and splice sites (the distribution of their distances from splice sites) will be analyzed, and our attention will be directed towards motifs whose distribution relative to splice sites is consistent with a role as ESEs.

In order to avoid biasing this analysis, we will consider the spatial distributions of all short oligomers from single nucleotides up to 12-mers (or longer, if the data will support it). We will calculate the locations of over-represented N-mers and compare these distributions to the expected distributions based on prior probabilities. Oligomers whose distributions deviate from the expected will be examined further and experimentally tested as part of Specific Aim 2.

It is interesting in this regard that frequencies of single nucleotides differ at one end of the exon from the other. In particular, the frequency of G is higher adjacent to the 3' splice site (25.1% vs. 21.6% in a 50 nt. window), while the frequency of C is higher in a 50 nt. window adjacent to the 5' splice site (24.0% vs. 19.9%) (Simpson & Filipowicz 1996). This difference could conceivably reflect ESEs.

1d) Alternative splicing.

The current annotation of *Arabidopsis* reflects only one gene for each locus; alternative splicing has not yet been addressed. Therefore, we expect that some splice sites, and some ESEs,

will appear in locations inconsistent with existing gene models. Once a set of candidate ESEs has been identified, we will scan the *Arabidopsis* genome again, looking specifically for these ESEs within potential exons. Many of these ESEs might be associated with alternatively spliced forms. We will develop new software to identify such putative alternatively spliced genes using the ESEs as a key. We will search the predicted gene products of these predictions in order to identify any other organisms in which the alternative form has already been sequenced and validated. The *Brassica oleracea* sequence described above will provide another, very powerful resource for identifying alternatively spliced genes. Any conserved sequences not corresponding to annotated exons may in fact represent an alternative splice form. By adjusting the parameters to GeneSplicer, we can generate additional putative splice sites and compare those to alignments between *Arabidopsis* and *Brassica* in order to build more evidence for alternative exon-intron structures. Depending on how many predictions we generate, we will submit some or most of these predictions to the validation steps in Specific Aim 2. Thus in addition to providing evidence for splicing enhancers, this work may at the same time provide new evidence for alternative splicing in *Arabidopsis*.

Specific Aim 2) The role of exonic splicing enhancers will be experimentally tested in vivo.

An in vivo assay will be essential for the confirmation and functional analysis of *Arabidopsis* ESE sequences. Therefore, we have devised an in vivo splicing assay to functionally test the candidate ESEs identified above. The concept of the experiment is to test ESEs for the ability to confer splicing, as reported by a GUS transgene in stably-transformed *Arabidopsis* plants (described below). The ESE sequences will be assayed either in their natural exon context within the reporter system or inserted into a weakly-spliced exon within a common intron-exon-intron construct. Using this transgene reporter system, we propose to examine expression patterns of up to 1000 distinct ESE sequences, and to contribute the transgenic seeds to the *Arabidopsis* stock centers. Because we expect the recognition of ESEs to be highly regulated in the plant, the resulting collection of transgenic lines should provide a useful resource for studying many aspects of plant growth and development on the basis of regulated mRNA splicing.

Although it would be useful to study the function of plant ESEs in vitro (particularly with respect to analyzing ESE-binding factors), there is currently no in vitro splicing system available for plants. Even if one were available, efficient recognition of ESEs might require multiple proteins (Li et al. 2000) or unknown proteins. Heterologous (animal) in vitro systems (Lopato 1996; 1999) could be used, but are likely to give misleading results. With our in vivo system, such problems will be avoided.

The first step in setting up the in vivo splicing assay will be to identify an intron-exon-intron unit that is enhancer-dependent, i.e. an intron-exon-intron unit in which the exon cannot be efficiently spliced unless it contains an ESE. Enhancer-dependence is a property of the splice sites themselves and of the surrounding introns. ESEs can activate splicing of upstream introns or downstream introns; when the exon is small, then the ESE will activate splicing both upstream and downstream. Thus, for a small enhancer-dependent exon, the insertion of a candidate ESE into the exon will confer inclusion of the exon upon splicing of the intron-exon-intron unit. In the absence of the ESE, the exon will be skipped, and the exon will be excluded from the spliced product. Similar studies have been successful at identifying ESEs in other organisms. In mammalian cells, for example, a novel set of ESEs was defined using exon inclusion versus exon skipping in an assay much like the one that we propose here (Coulter et al. 1997.). In addition, fully enhancer-dependent exons have been defined in mammalian systems, and have been used extensively to select splicing enhancers, both in vivo (Coulter et al. 1997) and in vitro (Liu et al. 1998; Schaal and Maniatis 1999b)

2a) The ESE assay system.

Exon inclusion vs. skipping will be assayed in plants using an introduced intron-exon-intron unit. Expression of the well-studied histochemical marker protein ß-glucuronidase (GUS) will be dependent upon the outcome of a splicing event. The reporter gene will consist of the following order of elements: an enhanced CaMV 35S promoter, the intron-exon-intron unit, the coding

sequences of the GUS reporter gene (either in or out of frame with the exon) and the 3' nos terminator sequence. We will construct and use a pair of vectors such that expression of the GUS marker from one vector will require exon inclusion while expression from the other vector will require exon skipping. These vectors will differ by a single nucleotide at the 3' splice site of the downstream exon. In vector I, inclusion of an enhancer-dependent exon is required in order for the GUS coding region of the mRNA to have an in frame AUG translational start site. The AU nucleotides of the AUG codon are provided at the 3' end of the middle, enhancer-dependent, exon as shown below. When the exon is included, translation initiation will result in GUS expression. When the exon is skipped, the correct translation initiation site is absent, and GUS protein will not be synthesized. The GUS sequence contains three out-of-frame AUGs upstream of the next downstream in-frame AUG, which is residue 103 in the full-length GUS protein). In vector II, an AUG will be created by the exon-skipping splicing event itself. In this case, if the exon is included, the correct translation site will be absent, and GUS protein will not be synthesized.

Vector 1 (active ESE makes GUS; no ESE activity means no GUS activity):

____A	TC ESE AT	GTTACGTCCTGTAG
Exon 1	Exon 2	Exon 3

Exon skipping (E1-E3): AGUUACGUCCUGUAG (no AUG, no expression of GUS)
Exon inclusion (E1-E2-E3): AUG UUA CGU CCU GUA G (expression of GUS)

Vector 2 (active ESE means no GUS activity; no ESE makes GUS activity):

____A	TC ESE AT	TGTTACGTCCTGTAG
Exon 1	Exon 2	Exon 3

Exon skipping (E1-E3): AUG UUA CGU CCU GUA G (expression of GUS)
Exon inclusion (E1-E2-E3): AUUGUUACGUCCUGUAG (no AUG, no expression of GUS)

The 5' exon (E1 in the schematic) used for these constructs will be derived from an *Arabidopsis* gene (F23H14.6), in which the initiating ATG is naturally split by an intron. Its transcription will be under the control of the well-characterized cauliflower mosaic virus (CaMV) 35S promoter, which confers essentially constitutive expression (Odell et al. 1985). The 3' exon (E3 in the schematic above) will be derived from GUS, starting with the last nucleotide (vector 1) or two nucleotides (vector 2) of the initiator AUG codon.

Between E1 and E3 will be an intron-exon-intron combination. A number of sources for the internal exon (E2 in the schematic) will be tried. Initially, we will use an exon based on *Arabidopsis COP1* exon 6, which is skipped in the *cop1-1* mutant (Simpson et al. 1998). The sequence of the wild-type *COP1* exon is:

```
tttattcgcag | TCGTCTAAG ... GCCAACATTGTATCAAG | gtgaattttctat
```

and the cop1-1 exon-skipping mutation changes position -4 of the 3' splice site: gcag| to acag|. We will incorporate this mutation in the first intron of the intron-exon-intron unit. In addition, we will alter the 3' end of the exon to introduce the required ATG:

```
...acag | TC GCT TAA {RE site 1}...{RE site 2} atttcgccAT | gtgaat...
```

Phase I - Validation of the in vivo splicing assay. Generation and verification of an enhancer-dependent exon.

Initially, we will identify an enhancer-dependent intron-exon-intron unit using a variety of exon sequences including our best candidate ESEs. As detailed below, we believe that GAAGAA repeats are very likely to function as a splicing enhancer. Other sequences recognized by mammalian SR proteins may also act as enhancers. For example, the sequence

CUUCGAUCAACGCCACGCCA is bound by both 9G8 and the *Arabidopsis* SR protein atRSZp22 (Lopato et al. 1999). In order to functionally identify an enhancer dependent exon, three candidate exons will be explored:

1) Exon 6 of the *Arabidopsis COP1* gene. This exon is an internal exon sensitive to splice site mutations that cause exon skipping (Simpson et al. 1998) as diagrammed above.

2) The alternatively-spliced exon from the tobacco *N* resistance gene (Dinesh-Kumar and Baker, 2000).

3) A completely synthetic exon with splice sites of average strength, but no predicted ESEs.

During this phase, experimental tests will require a rapid feedback. In order to quickly assay enhancer-dependence of these exons, routine particle bombardment will be performed on detached *Arabidopsis* leaves as described in Leister et al. (1996). The intron-exon-intron units (containing the candidate exons) will be synthesized by overlap extension or megaprimer PCR (Ling and Robinson, 1996). The PCR product will be ligated into corresponding restriction sites located between the CaMV 35S promoter and the GUS coding sequence in a modified version of plasmid pRITAI, which has been provided by John Bowman (R. Khodash and J. Bowman, unpublished). All constructs will be verified by DNA sequencing. The splicing pattern of the reporter gene will be assessed by RT-PCR, and GUS reporter activity will be visualized by light microscopy.

Phase II - Characterization of ESEs.

Having obtained one or more enhancer-dependent exons (intron-exon-intron units), we will be in a position to test candidate ESE's for activity. These will include motifs that were both defined computationally, as described in Specific Aims 1a through 1c, and alternatively-spliced exons identified as outlined in Specific Aim 1d. One example is the motif GAAGAAGAA, which acts as an ESE in vertebrates, is very characteristic of exons (vs. introns) in *Arabidopsis*, and is bound by a vertebrate protein (ASF/SF2) with clear counterparts (atSRp30 and atSRp34/SR1) in *Arabidopsis*. We expect that when this 9 nt. sequence is cloned into the test exon described in the previous section, it will promote exon inclusion. What we then hope to learn is what sequences within the nonamer are essential for its function as an enhancer. In vertebrates, $(AAG)_8$ is very effective, $(GGA)_8$ less so, and $(AAAGGG)_4$ ineffective (Tanaka et al. 1994). Will the same be true in *Arabidopsis*? Will variants of this motif show different patterns of tissue-specific activity? One reason to think they might is that the two SF2/ASF homologs, atSRp30 and atSRp34/SR1, do show tissue-specific patterns of expression (Lopato et al. 1999). How long must the motif be to act as an enhancer? Dissection of the caldesmon gene exon 5 enhancer showed that distinct purine-rich enhancers have different effects on two competing nearby 5' splice sites (Elrick et al 1998). Would flanking sequences or the arrangement of two 9 nt. motifs in space affect the pattern of activity in our assay?

We estimate that it will be possible to characterize 1,000 exon sequences (in 1,000 pairs, or 2,000 total clones) in the period of this grant. This number should be sufficient to answer questions like those posed above for each of 20 or so general types of motifs, in addition to simple validation of candidate motifs suggested by the bioinformatics. In addition, we will test segments from exons in alternatively spliced genes to see if they are sufficient to confer alternative splicing activity, and in order to map the sequences necessary for proper regulation of the alternative splicing.

To facilitate rapid construction of the 2000 different clones, the reporter constructs will be designed with two unique restriction sites in the internal exon. To introduce the 2000 different ESE versions into the exon, each specific version will be synthesized as a pair of complementary oligonucleotides that will create the intron-exon-intron unit (with sticky ends) upon annealing. Each of the annealed intron-exon-intron fragments will be directionally ligated into corresponding sites in the binary vector's reporter construct, and will be verified by DNA sequencing before transfer to *Agrobacterium*. A uniform exon length of 66 nucleotides will be maintained, so that overlapping oligonucleotides of 30 bp. can be used together with 36 constant nucleotides (an average of 18 at each end of the intron). We (the Mount lab) have used this technique for the construction of mutations in the past, and find that it works extremely well.

For *Agrobacterium*-mediated transformation of *Arabidopsis*, we will use the binary vector pMLBart (K. Richardson, personal communication). This is a modified version of pART7 (Gleave, 1992), which confers BASTA resistance to plants. (BASTA is ammonium glufosinate, and is commercially available as Finale (AgrEvo, Montvale, NJ). BASTA can be used to effectively select young plants growing in soil) NotI restriction sites flanking the reporter gene in pRITAI will be used to subclone the reporter genes from pRITAI into the NotI site of pMLBart. *Agrobacterim*-mediated transformation of *Arabidopsis* will be performed using the vacuum-infiltration method (Bechtold et al. 1993) or the "dip" method, as routinely performed in the Chang lab. Seeds from the *Agrobacterium*-treated plants will be sown directly on soil and the transformants selected by treatment with BASTA.

In the first generation of BASTA-resistant plants, GUS reporter activity will be examined in leaves of young plants as well as in flowers, in order to identify the pattern of expression that is typical of that construct. The subsequent generation of seeds, produced by self-fertilization, will be collected and grown for the analysis of tissue-specific reporter expression. Because temperature variation should be expected to directly affect the strength of RNA-RNA interactions that are critical to the removal of introns by the spliceosome (Staley and Guthrie 1998), we expect that some splicing events will have to be temperature-compensated, and we will look at effects of temperature as well.

Our analysis of ESEs will include:

- about 500-800 **candidate sequences** selected based on oligonucleotide bias by the bioinformatics project. This number includes variants of approximately 12-20 specific motifs as described above. Because mammalian SR protein binding sites are typically 8-10 long, this number should allow a moderately thorough analysis of the sensitivity of each motif to mutation. Alteration of every base of a 10-mer to every other base requires 30 experiments.
- 200-500 exons, or fragments of **actual exons**, that are of interest, including those that have been shown to undergo alternative splicing (e.g. Zhang et al. 1999).
- **Expression pattern** documentation. For every sequence tested, we will examine GUS activity in transgenic seedlings, leaves, roots, stems, siliques and flowers. This should result in a comparative description of the regional specificity of action of ESE motifs. We will examine expression after exposure to low and high temperatures (15° or 37°, respectively). We anticipate that the transgenic lines will not only exhibit tissue and stage-specific GUS expression, but show regulated GUS patterns in response to a broad array of signals. However, a full analysis of the effects of numerous plant hormones and environmental conditions is beyond the scope of this proposal.

This project should facilitate the analyses of gene function in a novel manner. We envision that researchers will able to locate ESE motifs in their genes of interest, alerting them to the possibility of alternative splicing as a regulatory mechanism. If the ESE motifs are similar to any among the 1000 we have assayed, such researchers will be able to obtain seeds of the reporter lines for detailed analyses of expression patterns. Of course, we recognize that factors other than ESE activation will play a role in splicing regulation. Nevertheless, there is considerable variation in SR protein abundance among tissues, and the *Drosophila doublesex* gene serves as a well-characterized example of a regulated ESE. These lines will serve as a resource for rapid identification of regulation through ESEs and contribute to the identification of ESE regulatory factors using biochemical or genetic approaches. This will lead to an improved understanding of numerous genes involved in different aspects of plant growth and development.

Eventually, the observed regulated splicing patterns will be compared with the expression patterns of the twenty known SR proteins in *Arabidopsis*. Most of these splicing factors are represented on the *Arabidopsis* Functional Genomics Consortium microarrays (http://afgc.stanford.edu), so information on their expression patterns (at least at the RNA level) should be available soon for a wide range of conditions. Once the ESEs are well-defined by this project, full genetic characterization of SR proteins and other possible splicing factors will be of interest to the Mount lab for future funding periods.

Specific Aim 3) Incorporate ESEs into genefinding software

The experiments conducted for Specific Aim 2 will yield new information about sequences that function as splicing enhancers. We will incorporate these new ESEs into two systems already in use for *Arabidopsis* genome annotation, GeneSplicer and GlimmerM. These systems are part of the TIGR automated annotation pipeline, and their output is also displayed in our graphical annotation editor, which is used for manual curation of the genome.

GeneSplicer uses a combination of decision trees (MDD trees, see Burge and Karlin (1997)) and Markov chains to assign a score to potential donor and acceptor sites. The scoring algorithm computes a probability P(A|M) which represents the probability that a sequence contains an acceptor site given the model M, where M is a Markov chain. (A similar computation applies to donor sites.) This probability is then divided by the probability that the sequence is not an acceptor, P(N|M), which is computed using a set of AG dinucleotides that are located in intergenic regions. (Of course, a small number of this might be 'missed' acceptor sites, but the data set is extremely large, containing all AG dinucleotides in the genome other than acceptor sites, and the statistics of a small number of false negatives should have an insignificant affect on the Markov model.) To add ESEs to the models, we will use a large set of exons from the entire *Arabidopsis* genome that have been validated either through protein or cDNA sequence homology, collected under Specific Aim 1. From this data, we can compute the likelihood that any of a set of motifs matching validated ESEs are present within an exon. We plan to use a simple hidden Markov model (HMM) design to allow for the ESE motif to be located a variable distance from donor/acceptor site; using insert/delete states gives an HMM this particular advantage over the straight Markov chains used in GeneSplicer.

TIGR is also one of the centers sequencing and annotating the rice genome. We propose to search the ESEs against predicted rice genes to identify potential splicing enhancers in that organism. Although experimental validation is not proposed here, we will annotate and release via our website any putative ESEs that we identify, with the goal of improving rice annotation and of facilitating further research into splicing enhancers in other plants. We note that ESEs are highly conserved among animals.

GeneSplicer is already incorporated in GlimmerM, so that once ESEs are integrated into the GeneSplicer algorithm, we will proceed directly to add ESEs to GlimmerM. In addition to their use in identifying splice sites, we will explore modifying the IMM scoring algorithm in GlimmerM. This algorithm assigns a score to coding regions that might benefit from a parameter reflecting the best match to an ESE motif within each exon. We will use our set of confirmed *Arabidopsis* genes to test any changes. The improved software will be made freely available to anyone through our web server, as the current version of the system is now. Furthermore, nonprofit institutions will have free access to the system for download and local installation.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, SE., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., et al. (2000). The genome sequence of *Drosophila melanogaster*. Science *287*,:2185-2195.

The *Arabidopsis* Genome Initiative (100+ authors) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815.

An, Y.-Q., McDowell, J.M., Huang, S., Mckinney, E.C., Chambliss, S., and Meagher, R.B. (1996). Strong, constitutive expression of the *Arabidopsis ACT2/ACT8* actin subclass in vegetative tissues. Plant J. *10*, 107-121.

Antezana, M.A., and Kreitman, M. (1999). The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J. Mol. Evol. *49*, 36-43.

Bechtold, N., Ellis, J., and Pelletier, G. (1993). *In planta Agrobacterium* mediated gene transfer by infiltration of adult *Arabidopsis thaliana* plants. C. R. Acad. Sci. Paris *316*, 1194-1199.

Blencowe, B.J. (2000). Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. *25*, 106-110. Erratum: Trends Biochem. Sci. 25(228), 2000.

Brown, J.W.S., and Simpson, C.G. (1998). Splice site selection in plant pre-mRNA splicing. Annu. Rev. Plant Physiol. *49*, 77-95.

Burge, C. B. (1998). Modeling dependencies in pre-mRNA splicing signals, In Computational Methods in Molecular Biology, S.L. Salzberg, D. Searls, and S. Kasif, eds. (Amsterdam: Elsevier Science) pp. 127-163.

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

Cáceres, J.F., and Krainer, A.R. (1997). Mammalian pre-mRNA splicing factors, In Eukaryotic mRNA Processing, A.R. Krainer, ed. (New York: IRL Press). 174-212.

Cleave, A.P. (1992). A versatile binary vector system with a T-DNA organisational structure conducive to efficient integration of cloned DNA into the plant genome. Plant Mol. Biol. *20*, 1203-1207.

Coulter, L.R., Landree, M.A., and Cooper, T.A. (1997). Identification of a new class of exonic splicing enhancers by in vivo selection. Mol. Cell. Biol.*17*, 2143-2150.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999**a**). Improved microbial gene identification with GLIMMER. Nucleic Acids Res. *27*, 4636-4641.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. (1999**b**). Alignment of whole genomes. Nucleic Acids Res. *27*, 2369-2376.

Dinesh-Kumar, S.P., and Baker, B.J. (2000). Alternatively spliced *N* resistance gene transcripts: their possible role in tobacco mosaic virus resistance. Proc. Natl. Acad. Sci. USA *97*, 1908-1913.

Egoavil, C., Marton, H.A., Baynton, C.E., McCullough, A.J., and Schuler, M.A. (1997). Structural analysis of elements contributing to 5' splice site selection in plant pre-mRNA transcripts. Plant J. *12*, 971-980.

Eisen, J.A., Heidelberg, J.F., White, O., and Salzberg, S.L. (2000). Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biology *1*, 1-9.

Elrick, L.L., Humphrey, M.B., Cooper, T.A. and Berget, S.M. (1998) A short sequence between two purine-rich enhancers determines 5' splice site specificity. Mol. Cell. Biol. *18*, 343-352.

Ermolaeva, M.D., Khalak, H., White, O., Smith, H.O., and Salzberg, S.L. (2000). Prediction of transcription terminators in bacterial genomes. J. Molec. Biol. *301*, 27-33.

Ermolaeva, M.D., White, O., and Salzberg, S.L. (2001). Prediction of operons in microbial genomes. Nucleic Acids Res., in press.

European Union Chromosome 3 *Arabidopsis* Genome Sequencing Consortium, The Institute for Genomic Research, and Kazusa DNA Research Institute. (2000). Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. Nature *408*, 820-823.

Fraser, C.M., Eisen, J.A., and Salzberg, S.L. (2000). Microbial genome sequencing. Nature *406*, 799-803.

Gallagher, S.R. (1992). GUS Protocols. (San Diego: Academic Press).

Graveley, B.R. (2000). Sorting out the complexity of SR protein functions. RNA *6*, 1197-1211.

Gravely, B.R., Hertel, K.J., and Maniatis, T. (1998). A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers. EMBO J. *17*, 6747-6756.

Gravely, B.R., Hertel, K.J., and Maniatis, T. (1999). SR proteins are 'locators' of the RNA splicing machinery. Curr. Biol. *9*, R6-R7.

Graveley, B.R., and Maniatis, T. (1998). Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. Mol. Cell *1*, 765-771.

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

Lazar, G., and Goodman, H.M. (2000). The *Arabidopsis* splicing factor SR1 is regulated by alternative splicing. Plant Mol. Biol. *42*, 571-581.

Lazar, G., Schaal, T., Maniatis, T., and Goodman, H.M. (1995). Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF. Proc. Natl. Acad. Sci. USA *92*, 7672-7676.

Leister, R.T., Ausubel, F.M., and Katagiri, F. (1996). Molecular recognition of pathogen attack occurs inside of plant cells in plant disease resistance specified by the *Arabidopsis* genes *RPS2* and *RPM1*. Proc. Natl. Acad. Sci. USA *93*, 15497-15502.

Li, X., Shambaugh, M.E., Rottman, F.M., and Bokar, J.A. (2000). SR proteins Asf/SF2 and 9G8 interact to activate enhancer-dependent intron D splicing of bovine growth hormone pre-mRNA in vitro. RNA *6*, 1847-1858.

Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.-I., Town, C.D., Fujii, C.Y., Mason, T., Bowman, C.L., Barnstead, M., et al. (1999). Sequence and Analysis of Chromosome 2 of the plant *Arabidopsis thaliana*. Nature *402*, 761-768.

Ling, M., and Robinson, B.H. (1996). Rapid construction of three-fragment recombinant DNAs by polymerase chain reaction: application for gene-targeting in Saccharomyces cerevisiae.Anal. Biochem. *242*, 155-158.

Liu, H.X., Cartegni, L., Zhang, M.Q., and Krainer, A.R. (2001). A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. Nat. Genet. *27*, 55-58.

Liu HX, Chew SL, Cartegni L, Zhang MQ, and Krainer AR (2000). Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol. Cell. Biol. *20*, 1063-1071

Liu, H.X., Zhang, M., and Krainer, A.R. (1998). Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. *12*, 1998-2012.

Lopato, S., Mayeda, A., Krainer, A.R., and Barta, A. (1996**a**). Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors. Proc. Natl. Acad. Sci. USA 9I3, 3074-3079.

Lopato, S., Waigmann, E., and Barta, A. (1996**b**). Characterization of a novel arginine/serine-rich splicing factor in *Arabidopsis*. Plant Cell *8*, 2255-2264

Lopato, S., Gattoni, R., Fabini, G., Stevenin, J., and Barta, A. (1999**a**). A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. Plant Mol. Biol. *39*, 761-773.

Lopato, S., Kalyna, M., Dorner, S., Kobayashi, R., Krainer, A.R., and Barta, A. (1999**b**). atSRp30, one of two SF2/ASF-like proteins from *Arabidopsis thaliana*, regulates splicing of specific plant genes. Genes Dev. *13*, 987-1001.

Mancebo, R., Lo, P.C., and Mount, S.M. (1990). Structure and expression of the *Drosophila melanogaster* gene for the U1 small nuclear ribonucleoprotein particle 70K protein. Mol. Cell. Biol. *10*, 2492-2502.

Mayeda, A., Screaton, G.R., Chandler, S.D., Fu, X.D., and Krainer, A.R. (1999). Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. Mol. Cell. Biol. *19*, 1853-1863.

McCullough, A.J., Baynton, C.E., and Schuler, M.A. (1996). Interactions across exons can influence splice site recognition in plant nuclei. Plant J. *8*, 2295-2307.

Mount, S.M. (2000) Genomics Sequence, Splicing and Gene Annotation. Amer. J. Hum. Genet. *67*, 788-792.

Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., and Fields, C. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. *20*, 4255-4262.

Mount, S.M., and Salz, H.K. (2000). Pre-messenger RNA processing factors in the *Drosophila* genome. J. Cell Biol. *150*, F37-F44.

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

Odell, J.T., Nagy, F., and Chua, N.-H. (1985). Identification of DNA-sequences required for activity of the cauliflower mosaic virus-35S promoter. Nature *313*, 810-812.

Peng, X., and Mount, S. M. (1995). Genetic enhancement of RNA-processing defects by a dominant mutation in B52, the *Drosophila* gene for an SR protein splicing factor. Mol. Cell. Biol. *15*, 6273-6282.

Pertea, M., Lin, X. and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res., in press.

Pertea, M., Salzberg, S.L., and Gardner, M.J. (2000). Finding genes in *Plasmodium falciparum* chromosome 3. Nature *404*, 34.

Preston, C.R., Sved, J.A., and Engels, W.R. (1996). Flanking duplications and deletions associated with P-induced male recombination in *Drosophila*. Genetics *144*, 1623-1638.

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

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. (2000). Genome annotation assessment in *Drosophila melanogaster*. Genome Res. *10*, 483-501.

Ring, H.Z., and Lis, J.T. (1994). The SR protein B52/SRp55 is essential for *Drosophila* development. Mol. Cell. Biol. *14*, 7499-7506.

Salzberg, S.L. (1999). Gene discovery in DNA sequences. IEEE Intelligent Systems *14*, 44-48.

Salzberg, S.L., Delcher, A., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Res. *26*, 544-548.

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J., and Tettelin, H. (1999). Interpolated Markov models for eukaryotic gene finding. Genomics *59*, 24-31.

Schaal, T.D., and Maniatis, T. (1999a). Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. Mol. Cell. Biol. *19*, 261-273.

Schaal, T.D., and Maniatis, T. (1999b). Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences.. Mol. Cell. Biol. *19*, 1705-1719.

Simpson, C.G., Clark, G.P., Lyon, J.M., Watters, J., McQuade, C., and Brown, J.W.S. (1999). Interactions between introns via exon definition in plant pre-mRNA splicing. Plant J. *18*, 293-302.

Simpson, G.G., and Filipowicz, W. (1996). Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. Plant Mol. Biol. *32*, 1-41.

Simpson, C.G., McQuade, C., Lyon, J., and Brown, J.W.S. (1998). Characterization of exon skipping mutants of the *COP1* gene from *Arabidopsis*. Plant J. *15*, 125-131.

Spradling, A.C., Stern, D., Beaton, A., Rhem, E.J., Laverty, T., Mozden, N., Misra, S., Rubin, G.M. (1999). The Berkeley *Drosophila* Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics *153*, 135-177.

Staley, J.P. and Guthrie, C. (1998) Mechanical devices of the spliceosome: motors, clocks, springs and things. Cell *92*, 315-326.

Stark, J.M., Cooper, T.A., and Roth, M.B. (1999). The relative strengths of SR protein-mediated associations of alternative and constitutive exons can influence alternative splicing. J. Biol. Chem. *274*, 29838-29842.

Tacke, R., and Manley, J.L. (1999). Determinants of SR protein specificity. Curr. Op. Cell. Biol. *11*, 358-362.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods & application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA *96*, 2907-12.

Tanaka, K., Watanabe, A. and Shimura, Y. (1994) Polypurine sequences within a downstream exon function as a splicing enhancer. Mol. Cell. Biol. *18*, 1347-1354.

Tettelin, H., Radune, D., Kasif, S., Khouri, H., and Salzberg, S.L. (1999). Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. Genomics *62*, 500-507.

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

Theologis, A., Ecker, J.R. , Palm, C.J., Federspiel, N.A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C.L., et al. (2000). Chromosome 1 of *Arabidopsis thaliana*. Nature *408*, 816-820.

Valentine, C.R. (1998). The association of nonsense codons with exon skipping. Mutat. Res. *411*, 87-117.

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

Zhang N, Portis AR Jr. (1999) Mechanism of light regulation of Rubisco: a specific role for the larger Rubisco activase isoform involving reductive activation by thioredoxin-f Proc Natl Acad Sci U S A 96(16):9438-43.

Zhu, J., and Krainer, A.R. (2000). Pre-mRNA splicing in the absence of an SR protein RS domain. Genes Dev. *14*, 3166-3178.