

ISMB 2006

**Tutorial on Post-transcriptional Gene Regulation: At the Interplay of
Genomes, Networks and Evolution**

DIRK HOLSTE

*Institute of Molecular Pathology
Dr-Bohr-Gasse 7
A-1030, Vienna
Austria*

holste@alum.mit.edu

AND

UWE OHLER

*Institute for Genome Sciences and Policy
101 Science Dr, Box 3382
Duke University
Durham, NC 27708*

uwe.ohler@duke.edu

Synopsis

The availability of complete genome sequences from yeast to metazoans gave rise to the paradigm that the increase in organism complexity is not adequately reflected in terms of the number of protein-coding genes, but instead is thought to be achieved by a growth in the complexity of gene regulatory networks. Messenger RNA transcript to genome alignments show that a marked number of vertebrate genes produces more than a single transcript and reveal that alternative splicing (AS) of precursors of mRNA constitutes an essential mode of in the network regulating the expression of genes. In addition to AS, the ongoing identification of regulatory, non-coding genes reveals another mode in this network, directing the development and physiology of cell types and tissues, and molecular pathology causing human disease. In this tutorial, we firstly overview a body of computational and experimental methods used to identify, annotate, classify, and characterize transcript diversity on the gene and whole-genome level, and secondly overview what is currently known about the evolution of splicing signals and other post-transcriptional mechanisms, including small RNAs.

CONTENTS

INTRODUCTION (2)

Eukaryotic gene structure and levels of gene regulation (2)
RNA splicing reaction and alternative splicing (3)

RELEVANT BITS OF COMPUTATIONAL METHODOLOGY (4)

RNA secondary structure prediction (4)
Sequence alignments and machine learning (5)

CONCEPTS OF EXPERIMENTAL METHODOLOGY (7)

Gene level: mutagenesis, RT-PCR, northern blotting, minigenes, RNAi, exon-specific activators, splicing modification by antisense oligomers (7)
Genome-wide: gene expression and Splicing-sensitive arrays, in situ hybridization, genome-wide RNAi screens, small-molecule compounds (12)

RNA SPLICING AND ALTERNATIVE SPLICING (14)

Biology and signals (15)
Alternative pre-mRNA splicing (17)
Sequence alignments and databases (19)
Transcript-based identification of alternative splicing (20)
Available resources (x)

OVERVIEW OF OTHER MECHANISMS: MOVING BEYOND ALTERNATIVE SPLICING (26)

UTRs and RNA-binding proteins (26)
MicroRNAs: genes and targets (27)
Networks of post-transcriptional gene-regulation (28)

CONCLUSION AND OPEN ISSUES (30)

ACKNOWLEDGEMENTS (31)

WEB SITES (31)

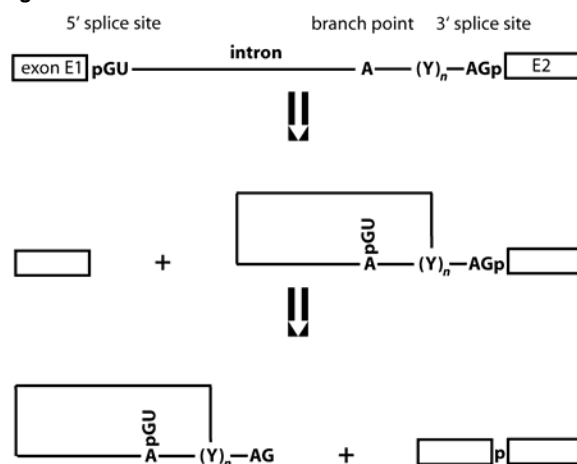
REFERENCES (32)

Introduction

EUKARYOTIC GENE STRUCTURE AND LEVELS OF GENE REGULATION

Gene expression is ubiquitous and involved with almost every process in the cell, ranging from the fertilization of germ cells or the differentiation of somatic progenitor cells, across the cell cycle, to stimuli-response pathways or apoptosis. In order to control the expression of genes under such diverse contexts, it is exerted on different several cellular levels. In higher metazoan genomes, protein-coding genes most often come in several “pieces” called exons, which contain all protein coding information, separated from each other by intervening regions of non-coding information called introns. The very 5' exonic region from the transcription start site (TSS) to the start codon is called the 5'-untranslated region (5'-UTR), and the stop codon is in turn followed by a 3'-UTR at the end of the gene. Genetic regulation of gene expression involves a series of complex biochemical mechanisms, and one can distinguish three different, albeit interconnected levels: (i) control of transcription, by utilization of *cis*-regulatory DNA elements, including promoters, enhancers, silencers, or locus control elements; (ii) splicing of precursors of mature messenger RNAs (pre-mRNAs) and subsequent mRNA nuclear-to-cytoplasmic transport, by utilization of splicing-specific signals and transport factors; and (iii) post-transcriptional control of mRNAs, by affecting the translational efficiency, the sub-cellular localization, or the stability of mRNAs. While at each level a series of distinct biochemical machineries is involved in controlling gene expression, these circuits can show a high complexity and as well as clear signs of interconnectedness (1).

Figure 1



RNA Splicing transesterification steps:

1. Two reaction intermediates: the detached 5'-direction exon and the lariat structure of the intron/3'-direction exon fragment.
2. Ligation of both exons and release of the lariat structure.

At the first level, the regulation of gene expression mainly directs whether a gene is transcribed and to what extent. In order to start transcription, the core promoter region in proximity to the transcription start site is essential and contains sufficient information of transcription factor binding sites for the recruitment of the RNA polymerase II protein complex for RNA synthesis to direct the correct expression of a gene. In addition to the core promoter, many genes require multiple *cis*-regulatory elements for precise gene expression, for instance, during different stages of development, differentiation into different cell types, or response to external stimuli. At the next level, RNA splicing regulates the generation of one or multiple distinct mature mRNA isoforms from a single gene locus and can also affect the expression of isoforms in dependence of the presence of pre-termination nonsense codons via nonsense-mediated

mRNA decay (NMD). From the nascent transcript, introns are spliced out from the pre-mRNA to produce mature mRNA that can be translated into a protein. During or after the ligation of exons, protein complexes are positioned at exon-junctions, and transport factors bind to the processed transcript to guide the transport from the nucleus to the cytoplasm for mRNA translation. At the last level, the main determinants of regulation involve the stability of the mRNA, the translation efficiency, and the cytoplasmic, nuclear, mitochondrial, or extra-cellular localization. The degradation of mature transcripts

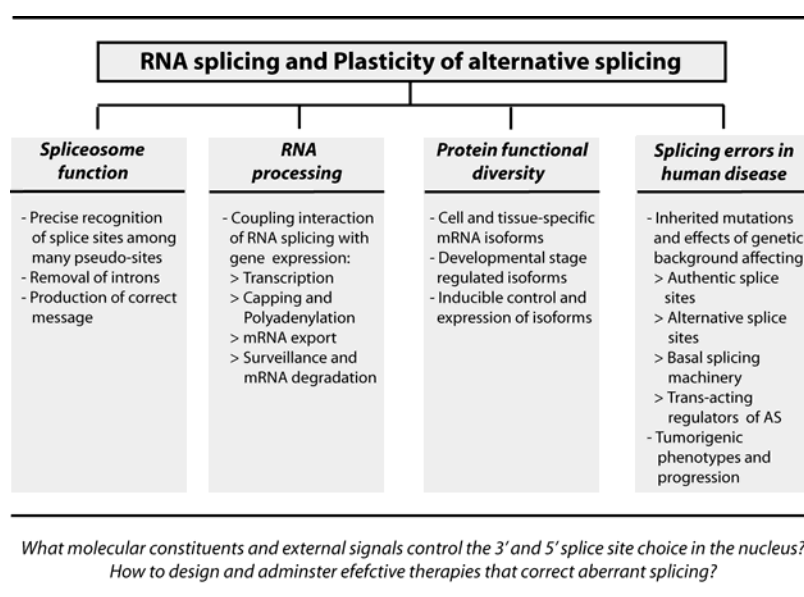
can also be affected by the presence of target sites for binding small non-coding RNAs (ncRNAs), such as microRNAs, whereas the rate of translational initiation for protein synthesis as well as the sub-cellular localization is often modulated by the structure of 5'-UTR sequences or the function of *cis*-regulatory elements for RNA-binding proteins. While the regulation by means of transcription factors is predominantly encoded in the DNA primary sequence, 5'- and 3'-UTRs can function either due to the RNA primary sequence or the formation of RNA secondary structure.

RNA SPLICING REACTION AND ALTERNATIVE SPLICING

Most protein-coding genes are transcribed as precursors of messenger RNAs (pre-mRNAs), and the splicing and excision of introns from precursors to mature mRNAs constitutes a critical mode for the regulation of gene expression at the level of pre-mRNA processing (2-4). RNA splicing is executed in the nucleus by the spliceosome, a large ribonucleoprotein complex that involves five small nuclear RNAs functioning as ribonucleoproteins (snRNPs) and potentially hundreds of proteins (5-7), the core components of which are highly conserved across metazoan genomes (4,8).

Figure 2 | adopted from Grabowski (2004)

Splicing of pre-mRNA occurs in a two-step reaction for a given pair of exons (Figure 1). During the first step, the 5' splice site



(5'ss) of the upstream exon is cleaved and the lariat intron is linked to the intron branch site, upstream of the 3'ss of the downstream exon. During the second step, the mRNA intermediate is cleaved at the 3'ss, resulting in the ligation of the pair of exons and the release of the lariat (9-11). The processing of precursors to mature mRNAs is frequently not constitutive but variable, and additional splice sites are used as alternatives, giving rise to multiple alternatively spliced (AS) mRNA products and consequently different protein isoforms (2,12-16). In addition

to the precise recognition of canonical splicing signals (the 3'ss, branch site, and 5'ss) and additional splicing *cis*-regulatory elements to remove the introns and produce the correct message, the spliceosome thus has to (i) regulate gene expression and to produce isoforms in a functional manner specific to different cell or tissue types, as well as to different stages of cell differentiation or development, and (ii) integrate RNA splicing with other components of RNA processing, such as polyadenylation (17-20), which also undergoes alternative usage. A picture emerges in which the control of gene expression is thought of as a complex network of interactions at the level of transcription, as well as at the levels of RNA processing, export and surveillance (1,21).

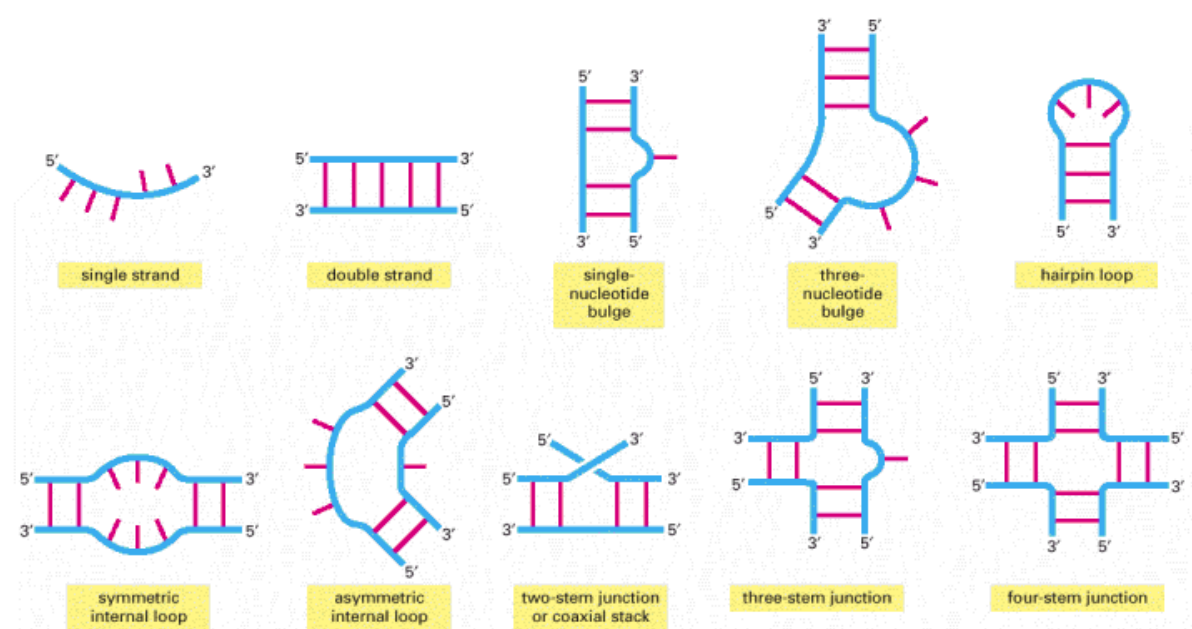
Consequently, alternative pre-mRNA splicing is thought to be biologically important for the diversification of the repertoire of expressed proteins, in different cell types or tissues, or at different

stages of development and differentiation, and has also implications for transport efficiency and cytoplasmatic mRNA stability (Figure 2). It is known that the alternative choice of splice sites is influenced by a number of factors, such as exon and flanking intron size, splice site affinity or pre-mRNA secondary structure, and that splicing *cis*-regulatory elements, functioning in context as exonic/intronic splicing enhancers and silencers, can further influence the splice site choice, by recruiting positive and negative *trans*-acting splicing factors (22-28).

Relevant Bits of Computational Methodology

Before addressing specifics of post-transcriptional gene regulation, we briefly review some computational concepts relevant to this topic. Here, we concentrate on the prediction of RNA secondary structures and on sequence analysis/classification algorithms.

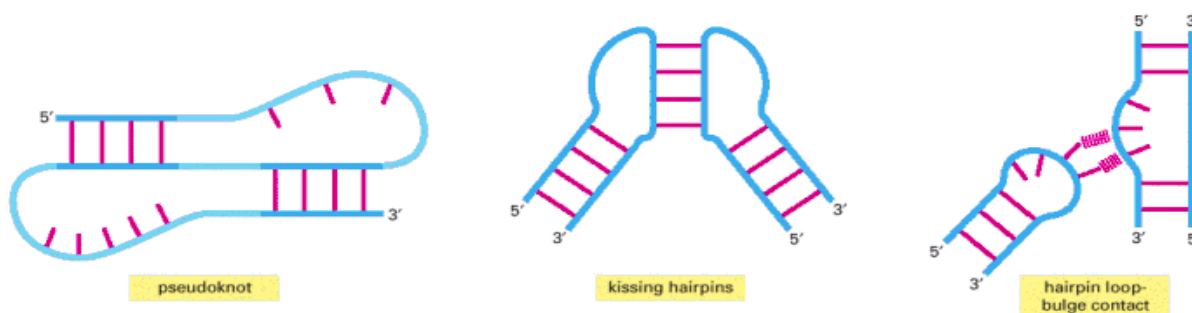
Figure 3 | Examples of common elements of RNA secondary structure (from D. Mount, *Bioinformatics*, 2nd ed)



RNA SECONDARY STRUCTURE PREDICTION

Post-transcriptional gene regulation takes place on the level of RNA, and involves functional regions within the message itself (the mRNA encoding a specific protein) as well as trans-factors, some of which are again functional RNA molecules. By interactions between complementary base pairs, a single stranded RNA molecule can form secondary structures, either involving different nucleotides in the same molecule, as well as duplexes between two distinct molecules. The different possible interactions include Watson-Crick base pairing of G-C and A-U, in analogy to the complementary nucleotides in DNA.

Figure 3 (continued)



In addition, G can interact with U to form weaker pairs. We will here focus on the case of secondary structure involving a single RNA. Some examples of common secondary structure elements are given in Figure 3.

Energy minimization A popular paradigm of RNA secondary structure prediction follows the principle of lowest free energy — a molecule will fold into the most favorable, i.e., stable secondary structure. Free energy is lowered by forming base pairs, and algorithms based on Dynamic Programming can determine the structure with the most base pairs, conditional to simple restrictions like minimum sizes of substructures like bulges or stems (29). These algorithms are among the oldest successful examples of computational biology and predate, e.g., accurate gene finding by more than a decade. Simply optimizing the number of base pairs does however not adequately reflect the underlying biology; however, a large body of research focuses on the experimental determination of parameters contributing to the free energy. These include parameters such as stacking of base pairs (the energy of a base pair in context of the preceding one), opening bulges at particular locations, size of the bulges, and so forth (30). Dynamic Programming algorithms can easily be modified to take these detailed parameters into account, and are the foundation of the popular MFOLD and RNA structure software (31,32).

Probabilistic interpretation The free energy can be related to a probability via the statistical thermodynamics formalism (Boltzmann factor, partition function). Similar to the difference between Viterbi and Forward/Backward algorithm for hidden Markov models, the best structure can be set in relation to the total free energy summed over all possible structures (33). Through appropriate summations, one can then, for instance, evaluate the probability that a particular base pair occurs in *any* structure. This allows for more detailed investigations: instead of predicting a complete secondary structure, one can predict reliable substructures that occur with high likelihood. Some of these methods are for instance implemented in the popular Vienna RNA package (34), or in the Sfold program set on principles of Bayesian statistics (35).

What is the performance? Evaluations of RNA prediction algorithms showed that about 70% of base pairs are currently correctly predicted by energy minimization algorithms (36). The false predictions are certainly partly due to some inaccuracy concerning the energy parameters, which are furthermore strongly influenced by experimental conditions such as the pH value of the medium; however, there is also increasing evidence that some molecules do not necessarily always fold into the most stable structure, for instance if it is very different from a larger ensemble of similar but slightly suboptimal structures. Furthermore, due to computational constraints, the prediction of secondary structure is typically restricted to sequences up to several hundred nucleotides in length. This means that many algorithms use a sliding window to predict structure within a larger sequence, and the predicted secondary structure is often influenced by the size and particular location of such windows. To explore the possible variability, most available algorithm allow for the enumeration of suboptimal structures, both for the probabilistic setting

as well as the “classic” energy minimization (37). Finally, due to the complexity of the problem, most algorithms only address secondary structures made up of strictly nested components (i.e., those that can be modeled by context-free grammars). To predict more complex structures such as the infamous “pseudo-knots” (two interleaved hairpins), algorithms with higher computational costs have been developed and appropriate formal grammars been given to formalize the space of structures that they represent (38). Application of these algorithms on larger datasets is often practically infeasible.

All the above methods refer to prediction of secondary structure in a single RNA molecule. Different models are used to represent specific families of RNAs, based on collections of related sequences, both cis-elements as well as RNA gene families. We will briefly touch this issue in the last section of this tutorial (miRNA genes), but as this topic is not so relevant to AS, we refer the interested reader to an recent review (39) and book (40), or to last year’s ISMB 2005 tutorial on RNA (by Peter Clote).

SEQUENCE ANALYSIS AND MACHINE LEARNING

The second algorithmic topic we will discuss concerns sequence analysis, and in particular classification techniques. This is thought to only give the briefest snapshot to provide the context for the particular applications on AS discussed below. Specifically, we refer to a few representative excellent textbooks that have been published over the last decade (40-43).

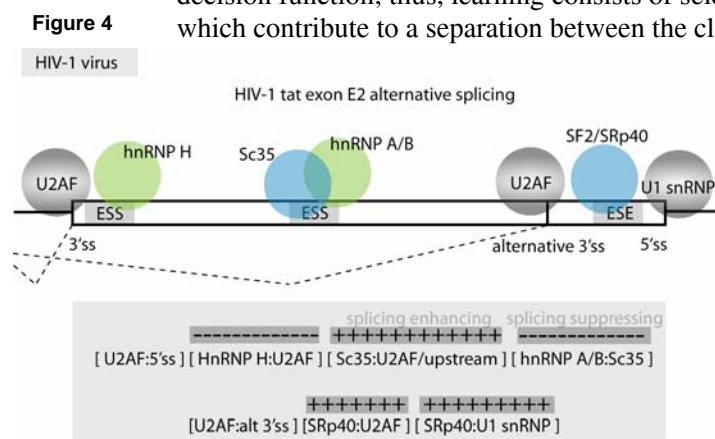
Dynamic Programming as the principle to obtain the optimal solution of a pair wise alignment problem is well known. To align ESTs and cDNAs to genomic sequence, a number of tools have been published which have adapted this framework to the case of a “spliced alignment”: Given a contiguous sequence (the genomic DNA), provide an alignment of a second sequence to it, whereby the second sequence can be broken into “pieces”, i.e., where long contiguous gaps are introduced which correspond to the introns that have been spliced out. Standard gap opening/extension penalties are not appropriate in this context; rather, gap penalties should be based on known intron length distributions, and gaps should preferentially appear at positions, which correspond to (canonical) splice sites. Furthermore, one of the sequences (the genomic one) is usually of excellent quality, whereas an EST sequenced in a single pass is often quite error-prone, especially towards the beginning and the end of the read. In the context of millions and millions of available EST sequences, some systems also use shortcuts to avoid the often-prohibitive quadratic runtime complexity. A number of tools are available which provide solutions, ranging from the largely heuristic but widely used SIM4 (44) to very recent ones which use the raw quality scores from the EST sequencing reaction (45). Below, we will place these algorithms in the context of a computational pipeline to infer AS on a genome wide scale.

Classification algorithms can be regarded as founded on one of two principles, depending on whether they aim at modeling a feature distribution of individual classes (generative approaches) or a boundary function between the classes (discriminative approaches). This is not a strict classification scheme and many overlaps exist, but it is helpful in our discussion here to point out differences between the two examples we consider.

Hidden Markov models (HMMs) provide a probabilistic approach to a large number of problems in computational biology, and have been applied successfully to diverse topics, ranging from gene finding to protein domain modeling (40). A discrete HMM contains a set of states, which emit symbols from an alphabet (here, the 4 nucleotides) according to a probability distribution. The states are connected by transitions to which probabilities are assigned. A state in such an

HMM has an associated probability of observing each residue, and the transitions determine the possible order of the states. A number of Dynamic Programming algorithms for HMM training and application are well known. The forward algorithm calculates the total probability that a sequence can be generated by a model, and can be applied to classification problems with several HMMs representing different classes. The Viterbi algorithm yields the parse of a sequence with the highest likelihood, thus assigning the symbols to model states, which may represent different functional categories such as exons and introns. Pair HMMs (pHMMs) are extensions of HMMs, originally described to perform a local or global alignments of two sequences (40). In general, the states of the model now contain probability distributions for an alignment of two residues, and by using several different states, a pHMM can be used to model different patterns of conservation. For example, pHMM systems to identify protein coding genes (46,47) include different states corresponding to pairs of aligned coding and non-coding nucleotides as well as splice sites. The standard HMM algorithms have been generalized and described in more detail for pHMMs (47,48) or, more generally, phylogenetic HMMs (49,50). PHMM algorithms are often applied on pre-aligned sequences to reduce the runtime, but in principle, the pHMM Viterbi algorithm computes the optimal alignment while obtaining the optimal parse of the alignment into different functional classes as “side effect”, based on the sequence of states used to generate the optimal alignment. As relevant examples, an early pHMM implementation focused on the problem of spliced alignment (51), and a sampling algorithm for generalized HMMs (which are used as standard in ab initio gene finding) was described to obtain predictions of complete alternative gene structures (52).

Support vector machines (SVMs) or in general related approaches from statistical learning theory, aim to learn a decision function separating between two classes. The important feature is that positive and negative training examples (the “support vectors”) build the basis for the decision function; thus, learning consists of selecting a subset of the training set which contribute to a separation between the classes, subject to the constraint that



Human immunodeficiency virus type 1 (HIV-1) alternative splicing pathway of tat exon E2: HIV-1 utilizes the cellular splicing machinery to generate more than 40 different mRNA isoforms, via multiple alternative 3'ss and 5'ss exons that exhibit weak splice sites, non-consensus branch-points, and short polypyrimidine tracts; E2 ESS elements are necessary for HIV-1 splicing inhibition, binding hnRNPs A/B

the decision function leads to the margin with maximum possible width between the support vectors. Further regularization terms may limit the number of chosen samples to prevent over-fitting and poor performance on unseen data. Similarity of data is calculated via the dot product of two samples, and classification of a test sample is performed, by comparing it to all support vectors. In general, the classifier does not compare the samples in the input space; instead, one employs a so-called kernel

function, which corresponds to a dot product in a different “feature” space (often with higher dimension), which allows one to learn an appropriate separation function. SVMs have gained immensely in popularity, as they are built on a rigorous mathematical framework which formalizes (and generalizes) older classification algorithms, like the k-nearest-neighbor algorithm

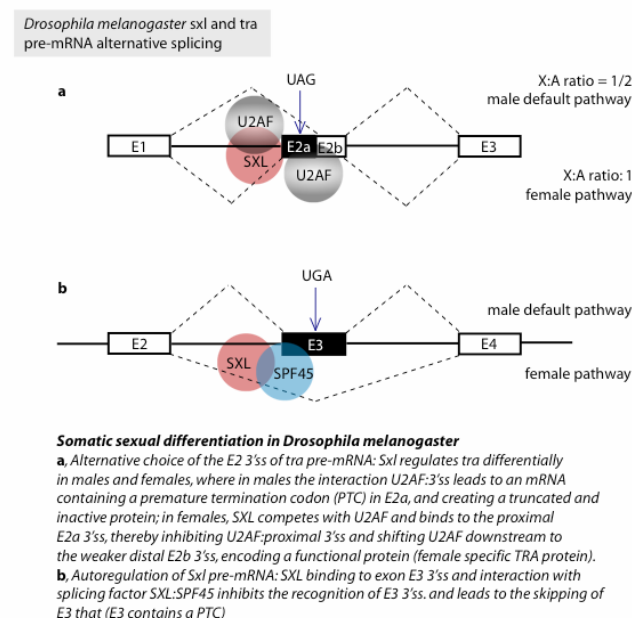
(53). Originally formulated for the two-class classification problem, variants can address multiple classes or cases in which only positive samples of one class are known. Many application examples use one of several popular “black-box” kernels (linear, polynomial, Gaussian) with reasonable success. To achieve good performance however, kernels need to be developed that reflect the data structure of a specific application area and the classification task. For instance, several string kernels (quantifying sequence similarity) have been proposed for applications on biosequences, with properties such as invariance to shifted positions within the input data or higher-order dependencies of the residues (54). Recent work has focused on how to integrate different kernel functions that work on different data in one classifier (55). In the context of AS, several groups have recently used SVMs for the classification of exons into skipped or constitutive, and we will discuss some of these methods below.

Concepts of Experimental Methodology

GENE LEVEL: *MUTAGENESIS, RT-PCR, NORTHERN BLOTTING, MINIGENES, RNAi, OLIGONUCLEOTIDES*

Significant understanding about gene regulation has come from studies utilizing site-directed mutagenesis of protein-coding genes or viral gene expression systems, as well as from disease-associated misregulation of AS (56). For instance, molecular dissections of individual genes have illuminated the role of AS in the *Drosophila melanogaster* sex determination pathway (3,57), the combinatorial control of the neuron-enriched *c-src* N1 exon (14,58), or the induction of AS in ion channels (14,59). Other indications for the role of splicing regulatory elements were derived from disease-associated studies, whereby chance disruption pointed to the detection of sites for functional motifs, and often to evolutionary conservation around sites. Since correct splicing is mandatory for the

Figure 5



viability of any cell, aberrant AS can have causal effects and influence on progression of human disease (60-64). It has been estimated that at least 10-15% of point mutations which give rise to genetic defects are due to changes in splice site sequences (65). A variety of standard experimental techniques typically found in molecular biology labs are available for the specific investigation of, e.g., the affinity of splice sites, the localization of putative splicing *cis*-regulatory elements, depletion of *trans*-acting splicing factors, or the determination of different isoforms. Some of these techniques are introduced below.

Site-directed mutagenesis Mutagenesis is a technology that seeks to change the base sequence and test its effect on gene or DNA function. It can be conducted in vivo or in vitro, and the mutagenesis can be site-directed in a

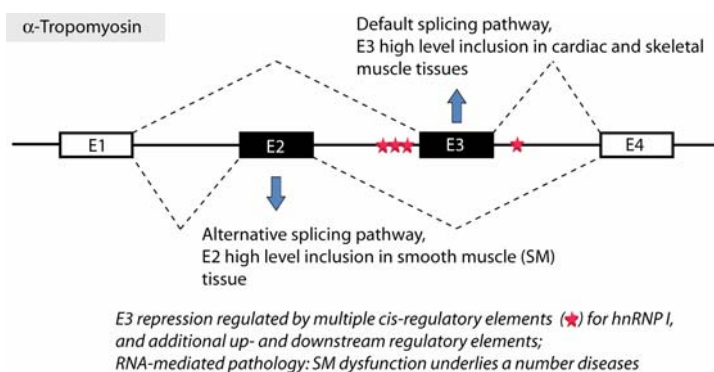
- (i) A popular general approach involves cloning the gene or cDNA into a vector, which permits recovery of single-stranded recombinant DNA. A mutagenic oligonucleotide primer is then designed whose sequence is perfectly complementary to the gene sequence in the region to be mutated, but with a single difference: at the intended mutation site it bears a base that is complementary to the desired mutant nucleotide. The mutagenic oligonucleotide primes new DNA synthesis to create a complementary full-length sequence containing the desired mutation.
- (ii) Site-directed mutagenesis by PCR enables base substitutions, deletions and insertions. In addition to producing specific predetermined mutations in a target DNA, it permits addition of a desired sequence or chemical group.

5' mutagenesis: a new sequence is added to the 5'-end of a PCR product, by designing primers that have the desired specific sequence for the 3'-part of the primer, while the 5'-part of the primer contains the novel sequence. The extra 5' sequence does not participate in the first annealing step of the PCR reaction as only the 3' part of the primer is target-specific, but it subsequently becomes incorporated into the amplified product. Alternatives for the extra 5' sequence include (a) suitable restriction sites that can facilitate subsequent cell-based DNA cloning, and (b) functional components (e.g. a promoter or reporter).

Mismatched primer mutagenesis: the primer is designed to be only partially complementary to the target site, but in such a way that it will still bind specifically. Mutations can also be introduced at any point within a chosen sequence using mismatched primers. Two mutagenic reactions are designed in which the two separate PCR products have partially overlapping sequences containing the mutation. The denatured products are combined to generate a larger product with the mutation in a more central location. Homologous recombination can then be used to substitute the genomic sequence by the exogenous DNA fragment carrying the mutation.

Northern blotting The temporal and/or spatial location of RNA expression can be determined by “running” an RNA blot (or a “Northern”). To this end, (i) total RNA is isolated from the nucleus or cytoplasm, or from whole tissues/organs, placed side-by-side on a gel, and separated by their sizes through electrophoresis (smaller RNAs will move faster through the gel); (ii) separated RNAs are transferred to nitrocellulose paper or nylon membrane filter, and the RNA-containing filter is incubated in a solution with radioactively-labeled single-stranded probe complementary to the mRNA of interest. After binding (if present) and washing off of any unbound DNA, labeled samples can be detected by autoradiography (exposure to X-ray film).

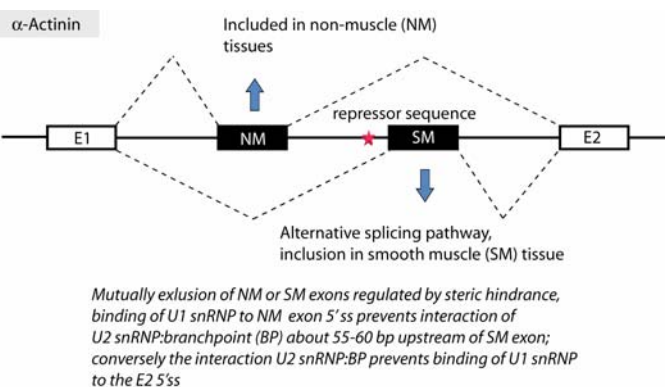
Figure 6



Reverse-transcriptase polymerase chain-reaction (RT-PCR) PCR, a method of in vitro DNA cloning, is used to determine whether a gene of interest is transcribing mRNA in a particular cell- or

tissue type. In combination with RT-PCR, it enables the conversion of mRNA into DNA, and to copy a certain DNA sequence of interest. (i) Purified mRNAs are converted into cDNA by using the enzymes RT and RNase H; (ii) A specific cDNA is targeted for amplification, by adding the first primer complementary to the mRNA fragment of interest to the cDNA population (designed to hybridize to one end of the target sequence); then, the second strand is complemented by using DNA *Taq* polymerase; (iii) The target dsDNA is denatured and the second primer (designed to hybridize to the opposite end) is added. The product is amplified by repeatedly cycling through the steps of DNA denaturation, primer hybridization, and extension of new strands.

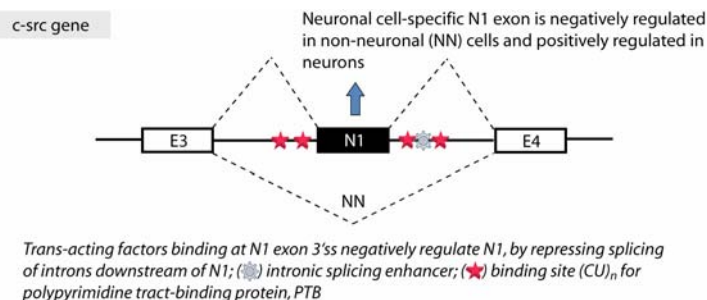
Minigene system A minigene contains (i) a transcriptional enhancer/promoter for ubiquitous expression; (ii) an upstream exon and 5' ss; (iii) a cloned genomic fragment from a gene of interest that includes the alternative exon(s) and flanking genomic regions, practically



several hundred nucleotides up- and downstream of the exon; (iv) a downstream exon and 3' ss; and (v) a *cis*-elements for 3'-end formation. The genomic fragment is amplified by PCR from genomic DNA by using oligonucleotides with restriction enzyme sites that match the restriction sites in the recipient plasmid vector. It is then transfected into cells suitable for the study of the gene, e.g., HeLa or HEK293 cell lines (mostly for non-constitutive splicing), keeping in mind that the same

minigene system can undergo different splicing behavior in different cell lines. Transient expression of minigenes is used in *in vivo* assays to study the intrinsic features of a gene that direct exon usage, *cis*-regulatory elements for the control of constitutive and/or alternative exons or cell-specific usage of alternative exons, as well as *trans*-acting factors that interact with these elements.

Green fluorescent proteins (GFPs) as a reporter To know whether a particular promoter was active, it is desirable to be able to determine its activity without having to measure mRNA levels: the Green Fluorescent Protein (GFP) provides such a way. GFP was the



first of many fluorescent proteins to be used as a reporter protein; when excited with a blue light, it will fluorescence green. GFP is very stable, can function when added to either end of a protein of interest, and does not fade easily. Similar to the usage as a reporter for promoter activity, GFPs can be used to report the usage of specific splice variants, when coupled with splice-dependent

creation of a premature termination codon and downstream non-sense mediate decay pathway.

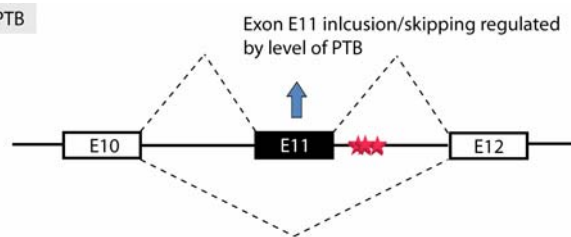
RNA-mediated interference (RNAi) RNAi is a mechanisms that works in protozoa and almost all higher eukaryotes, by which dsRNA induces the degradation of mRNAs sharing sequence homology

with the dsRNA. To this end, (i) dsRNA is introduced into cells or endogenously generated, upon which it is processed by the dsRNA-specific RNase Dicer to form effector molecules, termed small-interfering RNAs (siRNA), of 21-23 nucleotides in length; (ii) effector molecules are loaded into RNAi-induced silencing complex (RISC), where one siRNA strand is selected as the “guiding” strand, and the other is degraded; (iii) the guide strand is used by RISC to direct repeated rounds of target mRNA recognition, cleavage, and release. Note that microRNAs (miRNAs) are non-coding genes that exploit the RNAi pathway, by causing translational repression or cleavage (see below). As any protein can be “knocked-down” in principle by introducing dsRNA corresponding to the mRNA of interest, RNAi has been used in many experimental applications and is playing a prominent role as a tool in the analysis of gene function. In order to use RNAi to study the function of proteins encoded by AS, the dsRNA has to be made alternative exon-specific: mRNA isoforms containing the targeted exon are degraded, while isoforms lacking the targeted exon remain unaffected. Note that exon-specific RNAi is unlikely to work in organisms that show the phenomenon of transitive RNAi (such as worms and plants). Here, RNA-dependent RNA polymerase (RdRPs) can prime the synthesis of dsRNA complementary to the mRNA, and hence induce degradation of exons in addition to the target exon. Apparently, human and fly genomes seem not to encode genes homologous to the family of RdRPs, and exon-specific RNAi has been demonstrated to work properly.

Exon-specific synthetic activators ESEs are exonic splicing *cis*-regulatory elements which serve as binding sites for splicing factors, such as arginine/serine-rich (SR) proteins. The lack of ESE signals for accurate and efficient splicing in addition to the canonical splice signals

Figure 9

hnRNP I/PTB



Autoregulatory functional loop: steady state level of E11 skipping low (<1%), skipping of E11 causes frame-shift and creation of premature termination codon (PTC) in E12 (>55 bp of authentic stop codon), substituting the substrate to NMD pathway

can cause exon skipping. The main function of SR proteins is based on two features: (i) a targeting domain that recognizes specific RNA motifs, and (ii) a recruitment domain that mediates interactions with the splicing machinery. Based upon (i) and (ii), synthetic compounds can be designed which emulate ESE function and promote exon inclusion. Cartegni & Krainer (66) proposed the

utilization of ESSENCE compounds (exon-specific splicing enhancement by small chimeric effectors), which tethered RS-repeats to the exon by Watson-Crick base pairing of synthetic peptide nucleic acid (PNA) oligonucleotides. PNAs have a higher binding strength to DNA than DNA itself, and are resistant against enzyme degradation, and were able to reconstitute exon inclusion in *BRCA1* and *SMN2* minigene systems.

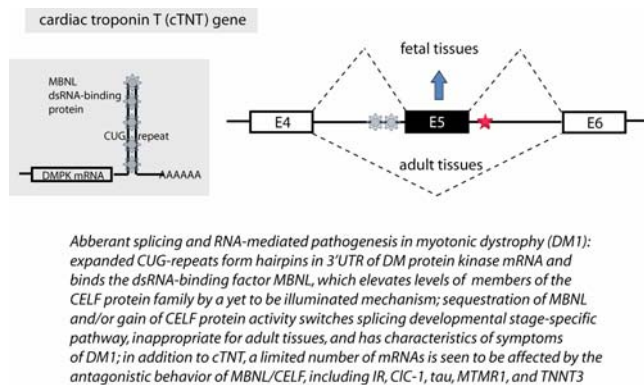
Splicing modification by antisense oligomers Antisense oligonucleotides for specific downregulation of gene expression work by Watson-Crick base pairing with a target mRNA, followed by destruction of the target by RNase H, which destroys the RNA in DNA/RNA duplexes. To utilize oligonucleotides for the modification of RNA splicing, they have to be insensitive to RNase H and compete efficiently with splicing factors for access to their target. Several synthetically designed oligonucleotides fit these requirements (e.g., PNAs or otherwise chemically modified oligomers). Such antisense oligonucleotides can be used to (i) silence mutations which cause the creation of cryptic splice sites, by basically blocking access of the spliceosome to cryptic splice sites; (ii) induce skipping of an authentic constitutive exon; or (iii) force selection of an alternative against a

competing splice site. This was successfully proved for several genes, including *Bcl-x*, *CFTR*, or *tau*, and also systemically delivered to upregulate gene expression in mouse tissues.

Single-molecule profiling Profiling complex AS patterns via the polymerase colony (polony) platform provides a digital assay for quantitating complex populations of mRNA isoforms. It permits the parallel amplification of individual mRNAs up to several orders of magnitude, such that each template gives rise to an individual colony. To this end, one strand of each amplification product is matrix-attached and serves as a template for probe hybridization and/or single-base extensions. Detection of multiple gene-specific sequences is achieved by combinations of spectrally distinct fluorophores and/or repeated cycles of probing. As each polony arises from a single molecule, it provides means for quantifying individual splicing variants in one or more pools of interest, coupled with a mini-sequencing strategy yielding single-nucleotide resolution (67).

GENOME-WIDE: GENE EXPRESSION AND SPLICING-SENSITIVE ARRAYS, GENOME-WIDE RNAi SCREENS

Figure 10 The above methods work efficiently for one gene at a time, but are hardly useful for high-throughput screens (HT) or genome-wide investigations for systematic functional analysis.



For instance, general and splicing-sensitive expression arrays have been used for the analysis of tissue-specific expression or splice site variation (21,26,68-72); moreover, a cross-linking/immunoprecipitation strategy has been introduced for identification of RNAs bound by a given splicing factor (26), and large-scale screens for small-molecule compounds that specifically interfere with splicing *cis*-regulatory elements have been performed (68). Clearly, these newly developed methods set a direction toward increasingly parallel experimental analysis of splicing regulation, and the combination of different HT assays

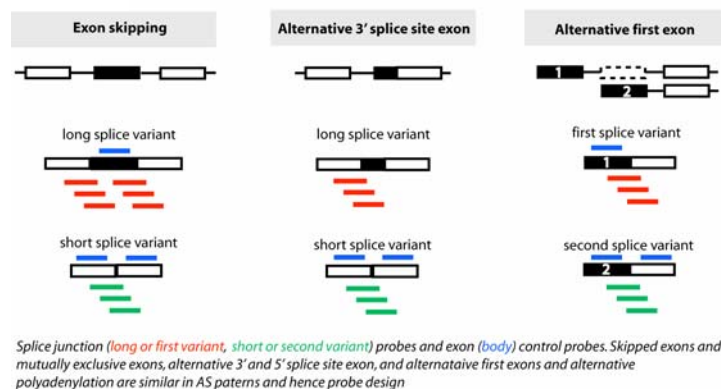
may even provide deeper insight to dissect combinatorial mechanisms.

Gene expression arrays DNA microarray technology has created means to monitor the transcription of thousands of genes simultaneously, combining Northern and PCR with robotics. One can distinguish two basic types of arrays, using either competitive binding or “absolute” measurement of mRNA abundances. In the former, “target” mRNAs are prepared, by isolating mRNAs from two cells or tissues being compared to one another and adding two different types of fluorescent dyes (green and red). “Probes” are made, by taking a number of cDNAs made from mRNA and, after cloning, denaturing and PCR amplification, the products are adhered to glass microscope slides. The probes and targets are hybridized together, and fluorescent intensities are obtained from laser confocal fluorescent microscopy, followed by image processing. Depending on the individual mRNA abundance, the signal obtained is a relative measurement, either predominantly green or red, or in case of roughly equal proportions yellow. Non-competitive hybridizations use designed oligonucleotides (often between 25 to 70 nucleotides long) to interrogate if a particular cDNA (and hence, the mRNA of interest) is present in differential amounts in a cell or tissue type (provided, e.g., by Affymetrix, Agilent, or Nimblegen). Using cDNAs from different mRNA isoforms, different splicing variants can roughly be measured. In addition, tiling arrays interrogate the expression of all non-random parts of whole-genomes and have also the potential to measure different splice variants.

Splicing-sensitive arrays Microarrays are designed to measure the overall level of mRNA from a gene, and thus provide only limited information about pre-mRNA splicing and alternative splicing. Standard microarrays can be adopted for the detection and measurement of alternative splicing. To this end, it requires a gene model or reference exon structure with known exon-intron boundaries and splice junctions. On the one hand, such a gene model will be built from prior knowledge of spliced alignments of cDNA sequences/expressed sequence tags (ESTs) or computational models (see below) and the array experiment is then conducted similar to a statistical case-control study, in which certain splice sites carry a non-zero prior probability of being alternative splice sites. On the other hand, with no prior information available, all splice sites may be treated as candidates for alternative usage.

The design of a splicing-sensitive microarray depends on the specific questions asked by the experiment, such as the sensitivity and specificity of interrogated splice types, and pertinent parameters for probe design, such as oligonucleotide size, melting temperature, secondary structure, low information content, or uniqueness/cross-hybridization, normalization of data, rule- or score-based decision of splice choices. Figure 11 shows an example for the design of probes for splice junctions for three different splice types: exon skipping, alternative 3' splice site exon, and alternative first exon (i.e. alternative promoter). Similar design principles have been used to measure, e.g., for several mutant yeast strains (lacking mRNA processing factors) the affect upon correct splicing, exon skipping patterns across a panel of about 50 human tissues and 20 cell lines, across a panel of about

Figure 11



ten mouse tissues, or alternative splice patterns across the National Cancer Institute NCI60 cancer cell lines. Apart from oligonucleotide-based platforms, there are alternatives available, such as RNA-mediated annealing, selection, and ligation (RASL), which are based on fiber-optics arrays.

In situ hybridization While Northern or microarrays can give an approximate time and spatial resolution of gene

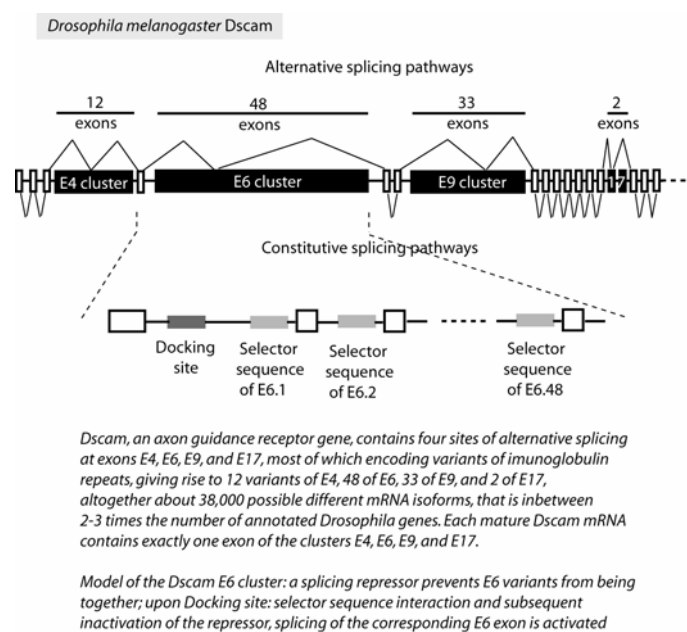
expression patterns, in situ hybridization (or sequence-specific staining) provides a detailed map of such patterns. To this end, (i) tissues, organs or embryos used are fixed to preserve their structure and prevent RNA degradation, and after sectioning are placed on a glass slide; (ii) a radioactively labeled antisense mRNA probe, made from a cloned gene in which the gene is reversed with respect to a promoter within a vector, is hybridized with the mRNA in the entire tissue or organ; (iii) the transcribed mRNA is complementary to the “sense” mRNA and used as a probe that will recognize the sense mRNA in the cell; (iv) unbound probes are washed off, the glass slide is prepared for autoradiography, and dark-field microscopy or bright-field imaging can be used to visualize cells that have accumulated a specific type of mRNA. Using dye-labeled probes (whole-mount in situ), entire organs or embryos can be stained.

Alternative splicing sequence enriched tags (ASSETs) Experimental identification of AS events can be carried out, by constructing an alternatively spliced library (ASL). To this end, (i) derive two

FL-cDNA libraries from distinct samples of cytoplasmatic RNA; (ii) prepare ssDNA from cDNA inserts of each library to obtain both sense and antisense strands; (iii) hybridize the ssDNA to form hybrids containing ssDNA from two the different libraries, where alternative exons form loops within dsDNA regions; (iv) remove unhybridized ssDNA, and enrich the remaining hybrids with single-stranded loop structures, by annealing to randomized oligonucleotides and capturing by magnetic beads; (v) ligate recovered DNA fragments, amplify by PCR and clone them into ASL, and sequence the ASL to obtain ASSETs (21).

RNAi screens of splicing regulators RNA-mediated interference works by silencing the expression of a sequence-specific targeted message of mRNA through effectors molecules (siRNAs), which cause cleavage of the corresponding mRNA. To identify the targets and splicing patterns governed by specific *trans*-acting splicing regulators, the combination of (i) systematically “knocking down” of known or predicted regulators via RNAi, and (ii) splicing-sensitive microarrays or the amplification of alternative exon-specific PCR products can be exploited. As a proof of concept in cultured cells of the model system *Drosophila melanogaster*, the lab of Donald Rio (69) studied the silencing of four known splicing factors and monitored several thousand splice patterns annotate in *Drosophila*, while the lab of Brenton Graveley (70) silenced more than 70% of annotated splicing factors in *Drosophila* and monitored the changed splice patterns in the genes *dAdar*, *para*, and *Dscam*. Alternatively, the

Figure 12



investigate phenotypic consequences of proteins encoded by alternative splicing.

Cross-linking and immunoprecipitation (CLIP)

Ultraviolet (UV) cross-linking and immunoprecipitation (CLIP) provides a way for in vivo identification of mRNAs that bind known *trans*-acting splicing factors. To this end, cells or tissue is UV-irradiated to form covalent bonds between protein and RNA in direct contact (photo-crosslinking). Cells are lysed, RNA partial digested into tags with RNase, protein-RNA complexes are immunoprecipitated with antiserum, and RNA is radioactively labeled for visualization by autoradiography. Complexes are

separated using electrophoresis (SDS-PAGE) and transferred to nitrocellulose, which remains protein-RNA complexes but not free RNA. Finally, proteins are removed with Proteinase K and RNA is cloned with the use of linker ligation and RT-PCR. Sequencing of CLIP tags and subsequent genomic sequence comparisons reveals the approximate location of the binding sites. This approach was developed and applied to Nova-regulated mRNAs in *M. musculus* brain by Bob Darnell's group (Rockefeller U, New York).

Small-molecule compounds Pharmacological drugs can be used for interference with spliceosome assembly and hence for modulation of splicing. Small-molecules may alter, e.g., post-transcriptional

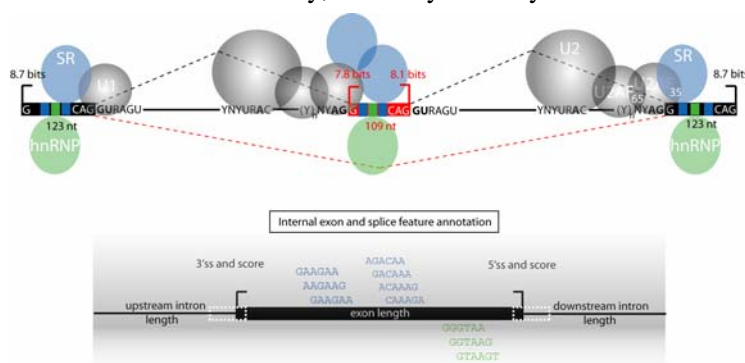
modifications or the interaction of arginine-serine (SR-rich) RNA-binding proteins with constitutive or alternative splicing *trans*-acting factors. For instance, the screening of several thousands small-molecules for their potential to inhibit phosphorylation of ASF/SF2 by topoisomerase I for splicing repression of reporter constructs, and for disruption of spliceosome assembly, pointed to indole derivatives. Positive compounds were tested for modulation of HIV-1 alternative splicing and several drugs specifically prevented HIV-1 production, opening new approaches to treatment (68).

RNA Splicing and Alternative Splicing

BIOLOGY AND SIGNALS

Similar to the genetic code, there exists an “RNA splicing code”. A typical mammalian gene spans tens of thousands of nucleotides, with on average nine exons/eight introns and the protein-coding region typically spanning about 1,500 nucleotides. In addition to the precise recognition of exon-intron boundaries among many additional possible pseudo-splice sites, the removal of introns and the production of the correct message, the spliceosome also has to control tissue- and developmental stage-specific mRNA isoforms and integrate it with other steps in RNA processing, including capping, cleavage, and polyadenylation. An interconnected network machinery of RNA splicing, functional *cis*-regulatory elements and *trans*-acting factors control these decisions (Figure 13). The spliceosome is highly conserved from yeast to man, with increasingly more complex eukaryotes adding more components to the regulatory network; e.g., yeast does not have any SR proteins, while flies and mammals do.

Figure 13



Noteworthy, baker's yeast only uses constitutive pre-mRNA splicing pathways, while flies and mammals exhibit AS. Consequently, a starting point in the understanding of AS is the analysis of signals achieving constitutive splicing in baker's yeast, and additional signals in higher eukaryotes.

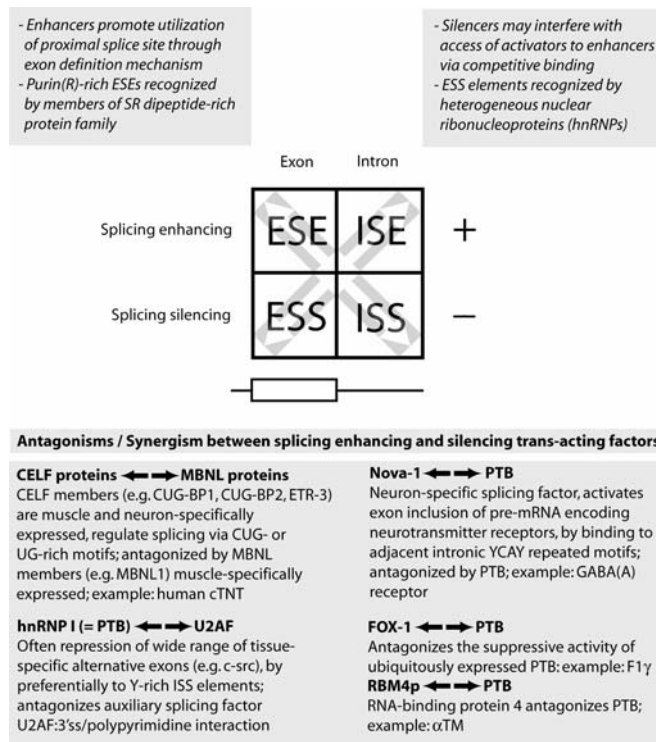
The signals that drive exon-intron recognition are located at the termini of introns. With the exception of about 1%, almost all introns are characterized by the canonical /GT terminus at the 5'ss,

and AG/ terminus at the 3'ss (U2-type introns). A small fraction of U2-introns exhibits /GC-AG/ termini, while a tiny fraction exhibits /AC-AT/ termini (U12-type introns). U2-type and U12-type introns are spliced by distinct spliceosomes in eukaryotic nuclei. A fission/fusion model proposes that the two splicing systems may have evolved in separate lineages, which were fused in a eukaryotic progenitor, converting most or all U12-type introns to U2-type (71)

During spliceosome assembly, individual splice sites are not independently recognized, but there are interactions between 3'ss and 5'ss, and splicing factors that recognize them. A model that invokes pairing between splice sites across an exon is termed “exon-definition”, whereas pairing across an intron is termed “intron-definition”. Typically, in pre-mRNAs with small introns (comparable to exon sizes), the spliceosome searches for closely spaced 5'ss-3'ss termini across an intron, in contrast to exon-definition, where the spliceosome searches for a closely spaced 3'ss-5'ss termini across an exon. When such a pair of splice site signals is found, the exon is defined by interactions U2 snRNP:3'ss and U1 snRNP:5'ss, and

Figure 12

additional general splicing factors (GSFs) and their interactions, including U2AF₆₅:branch site, U2AF₃₅:poly-(Y)_n. The average information content of canonical splicing signals (3'ss, branch site, and 5'ss) that define exon-intron junctions in *S.cerevisiae*, *C.elegans*, *D.melanogaster*, and *H.sapiens*, decreases – 31, 25, 25, and 23 bits (spanning altogether 40 nucleotides with a possible maximum of



80 bits). With increasing organismal complexity, splicing signals in mammals are therefore less conserved than in yeast, despite the fact that only a small proportion of yeast genes contain typically one intron (less than 300), whereas almost all genes in mammalian genomes contain several introns. Thus, compared to their yeast counterparts, the information content of the canonical splicing signals is generally insufficient in higher eukaryotes to ensure the correct assembly of the spliceosome. Additional signals are necessary, in particular when weak or regulated splice sites are involved. One class of such signals are exonic splicing enhancers (ESEs), which are ubiquitous in constitutive as well as alternative exons. ESEs act as *cis*-regulatory elements, similar to transcription factor binding sites, for arginine/serine-rich (SR) proteins. Members of the family of SR proteins contain several RNA recognition motifs (RRMs), are highly conserved and structurally similar, and SR-dipeptide-rich

splicing factors that are involved at multiple steps in the RNA splicing pathway. Mediated by their RS-domains, these factors recruit spliceosomal components via protein-protein interaction. The family of nuclear heterogeneous RNPs (hnRNPs) characterizes another class of splicing factors and often antagonizes members of the SR protein family. Biochemical investigations have further revealed tissue-specific factors, including members of the CELF (72,73), NOVA (14,74), PTB (75), and FOX (76) families of proteins that are in control of specific splicing events (Figure 14). For instance, FOX-1 is a RNA-binding protein recruited to the motif to the intronic motif TGCATG motif. This motif has been computationally identified downstream of the 5'ss, by searching genes specifically expressed in brain tissue, and also in genes expressed in muscle tissue.

Early indications for the role of splicing regulatory elements were obtained from disease-associated studies, whereby chance disruption pointed to the detection of sites for functional motifs, and often to evolutionary conservation around sites. Subsequently, several computational and/or experimental assays have been developed to identify ESEs and other splicing regulatory elements. In the following, we introduce a few of such assays.

Computational identification of ESEs Starting from the observation that ESEs compensate for weaker 3'ss and/or 5'ss, a computational screen (RESCUE) was developed that predicts ESEs, by statistical analyses of exons, introns, and splice site compositions. Starting from a large set of

constitutive exons, about 240 human RESCUE-ESEs (6-mers) have been predicted by discriminating hexamers that are enriched in exons flanked by weak splice sites, and all representatives of ten motifs exhibited enhancer activity in a minigene system in vivo (77). To further validate RESCUE-ESEs, a population genetics strategy was developed: mutations that create deleterious alleles will be underrepresented in common genetic variations, and therefore can be used to assess the extent of purifying selection pressure on functional sequences. By using a large-scale collection of human single-polymorphisms (SNPs), about one-fifth of RESCUE-ESEs were found to be eliminated by natural selection (77). Similar in spirit to RESCUE, oligomer frequencies of non-coding exons were contrasted against both pseudo-exons and 5' UTRs of intronless genes, carefully avoiding any potential bias resulting from codon usage. Clusters of 8-mers overrepresented in non-coding exons were selected as putative ESEs (PESEs), while underrepresented 8-mers as PESS elements. Cluster representatives were tested for function in minigene systems, and 10/12 PESS motifs and all eight PESEs generally functioned as silencers and enhancers, respectively (78).

Assay-based identification of exonic splicing silencers An in vivo splicing GFP reporter system coupled with downstream fluorescence-activated analysis (FACS) was developed to systematically search for ESS elements. This screen identified seven generally GT-rich ESS clusters (10-mers), in contrast to typically either purine- or AC-rich ESEs. A dozen of selected 10-mers displayed silencing activity in when transfected into a second cell-type, and splicing simulations were indicative of roles in pseudo-exon suppression and definition of splice sites (79).

Functional SELEX (Systematic Evolution of Ligands by Exponential Enrichment)

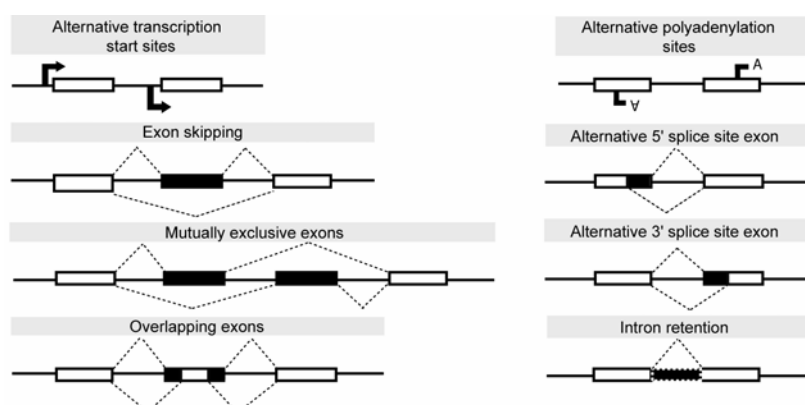
A minigene containing ESE motifs that are sufficient and necessary for the efficient splicing of its pre-mRNA is used by functional in vitro or in vivo SELEX: (i) an authentic exonic enhancer is substituted by a sequence from a randomized oligomer library; (ii) the pool of constructed minigenes is then transcribed in vitro, or transfected into cultured cells, to create a pool of pre-mRNAs; (iii) after splicing, the pools of spliced mRNAs is gel-purified and amplified by RT-PCR; (iv) this pool is then used to reconstruct the precursor and start a new cycle. The iteration (i-iv) will eventually enrich a small number of “winners”, that is ESEs with outcompeting splicing enhancer activity. For instance, the four well-known SR proteins ASF/SF2, SC35, SRp40, and SRp55 recognize the following consensus motifs, respectively: C[A/G][C/G][A/C][C/G][A/T], G[G/A][G/T][G/T][G/A]C[T/C][G/A], [T/C][G/A]C[G/A][T/G][C/A], and [T/C][C/T][A/T][C[A/T][G/C]G, and in a typical constitutive human exon, one most often finds altogether between 10-20 of these motifs.

ALTERNATIVE PRE-MRNA SPLICING

Based on the current known and annotated gene count, the human genome has the potential to express about 25,000 coding messages under a variety of different cellular contexts: at various stages and levels of differentiation, during the development of stages from an embryo to the human adult, while undergoing apoptosis, in response to external stimuli, or in case of human pathogenesis. Counting different proteins is more difficult than genes, but it is estimated that the human adult synthesizes more than 100,000 proteins. Many proteins are isoforms, or one of several “forms” of one protein, and have been produced by a gene that encodes their mature isoform messenger RNA by alternative pre-mRNA splicing (AS), which can be used to generate condition-specific protein variants and regulate gene-expression (4,13,80). AS is now estimated to affect the majority of actively transcribed human genes (81), and hence the computational identification and bioinformatics investigations of AS has gained in importance to cope with the diversity of mRNA isoforms (82,83). Powerful examples are given by the *Drosophila*

melanogaster homolog of the Down syndrome cell adhesion molecule (*Dscam*), as well as by the human neurexin (*NRXN*) gene family, which have the potential to produce and express hundreds to thousands of alternative mRNA isoforms (26,84,85).

Figure 13 AS events are commonly distinguished in terms of whether mRNA isoforms differ by inclusion or exclusion of an exon (the exon involved is referred to as "skipped exon" (SE)



or cassette exon), or whether isoforms differ in the usage of a 5' splice site or 3' splice site, producing alternative 5' splice site exons (A5Es) or alternative 3' splice site exons (A3Es), respectively (Figure 14). These descriptions are not necessarily exclusive, and an exon can make several alternative splice site choices. Another type of alternative splicing, termed as "intron retention" (IRE), occurs when two isoforms differ by the

presence of an unspliced intron in one transcript that is spliced in the other. Transcript sequence analyses indicate that AS pathways predominantly generate SE events in human and mouse transcriptomes, and likely generally across vertebrates. The frequency of occurrence of SE events is followed by A3E and A5E events, which in turn are followed by IRE, "overlapping exons" i.e. the simultaneous occurrence of A3E and A5E events, and mutually exclusive exons, i.e. a pair of exons with

Table 1

Database/Dataset	Web-based reference
Collection of EST-inferred splice sites	http://www.ebi.ac.uk/~thanaraj/splice.html
Collection of literature-based AS events	http://cgsigma.cshl.org/new_alt_exon_db2/
ASDB	http://cbcg.nersc.gov/asdb
ISIS (intron database)	http://isis.bit.uq.edu.au/a_splicers.html
ASMamDB	http://166.111.30.65/AsMamDB
SpliceDB (U2/U12 splice sites)	http://www.softberry.com
SpliceNest	http://splicenest.molgen.mpg.de
PALSdb	http://palsdb.ym.edu.tw
AltExtron	http://www.ebi.ac.uk/asd/altextron
AS graphs	http://www-cse.ucsd.edu/groups/bioinformatics/ESTs
ASAP	http://www.bioinformatics.ucla.edu/ASAP
ProSplicer	http://bioinfo.csie.ncu.edu.tw/ProSplicer
AS patterns in plants	http://pasdb.genomics.org.cn
ASD	http://www.ebi.ac.uk/asd
Splicing-conserved AS patterns	http://www.soe.ucsc.edu/~sugnet/psb2004/altGraphXCon.html
AS patterns across different tissues	http://genes.mit.edu/genoa
ASG, browsing splice patterns	http://statgen.ncsu.edu/asg
SpliceInfo	http://spliceinfo.mbc.nctu.edu.tw
ECgene	http://genome.ewha.ac.kr/ECgene
STACKdb	http://www.sanbi.ac.za/Dbases.html
EASED	http://easedb.bioinfo.mdc-berlin.de
Collection of alternative poly-A sites	http://physics.nyu.edu/~jy272/altA
LSAT (literature-based)	http://www.bork.embl.de/LSAT
MAASE (literature-based)	http://maase.genomics.purdue.edu
HOLLYWOOD	http://hollywood.mit.edu

one exon present in one transcript, while absent in another, and vice versa.

How is alternative splicing regulated? Similar to the genetic code, one can think of mechanisms of alternative splicing as governed by a "RNA code". The human spliceosome is estimated to consist of altogether 200-300 proteins, making it one of or the most complex cellular machinery.

In addition to understanding the assembly of the spliceosome at different spliceosome complex stages,

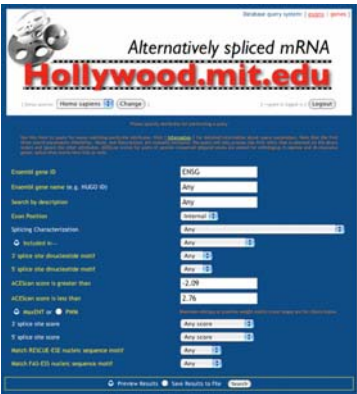
the RNA code involves the whole body of splicing regulatory elements (including splice sites, and ESE and ESS elements), general and tissue and/or developmental stage-specific splicing factors, and the network and nature of interactions between recruited factors and protein-protein interactions. While systematic analyses promise to deepen our understanding, there are still many gaps to fill before writing

down the components and interactions of the RNA code. Mechanisms producing SE events are likely the best understood AS events, and several individual functional skipped exons have been described and their regulation deciphered (see examples in this tutorial). SE events display a steadily increasing complex regulatory scheme, prototypes of which include alternative splicing of the N1 exon of the *c-src* gene, the *SMN1* and *SMN2* genes, or exon clusters of the *Dscam* and *CD44* genes. Other AS events are less well understood; in particular, IRE events are more difficult to infer and analyze because of the difficulty in distinguishing authentic IRE events from contamination by unspliced pre-mRNA sequences. Yet, individual functional IRE, A3E, A5E, and MXE events in various lineages are known.

DATABASES

As experimental systems and models for the regulation of AS have been steadily validated and refined (17), so have bioinformatics tools and computational models. With the availability of complete genome sequences and approaches of comparative genomics, the

Figure 16 | Database example (HOLLYWOOD)



identification of candidate sequence elements that can then be evaluated for their activity in controlling gene expression has become a major challenge in computational molecular biology. In order to systematically address the level of complex control achieved by utilizing AS, experimental and computational large-scale studies have started to illuminate the extent, structure and regulatory consequences of AS and its differential usages in mammalian genomes. To this end, databases for

recording types of AS have been constructed, and they can be grouped according into two approaches: based on curated searches of published research (86,87), and automated large-scale comparisons of transcript sequences, as well as hybrid approaches (88). While the first approach emphasizes the manual curation and sets the benchmark on the “specificity” of (authentic) AS events,

Table 2

Algorithm	Web-based reference
ASPIC	http://aspic.algo.disco.unimib.it/aspic-devel/
BLAST	http://www.ncbi.nlm.nih.gov
BLAT	http://genome.ucsc.edu/~kent
DDS/GAP2	http://www.tigr.org/software/alignment.shtml
Ensembl genome Browser	http://www.ensembl.org
EST_GENOME	
EXALIGN	http://blast.wustl.edu/exalign
GMAP	http://www.gene.com/share/gmap
MGAalign	http://origin.bic.nus.edu.sg/mgalign/
mRNAvsGen	http://genes.mit.edu/genoa
Sim4	http://globin.cse.psu.edu
Spidey	http://www.ncbi.nlm.nih.gov/spidey
GeneSequer	http://globin.cse.psu.edu/
TAP	http://sapiens.wustl.edu/~zkan/TIP/
UCSC Genome Browser	http://genome.ucsc.edu
WABA	http://www.cse.ucsc.edu/~kent/intronator
SPA	http://www.biozentrum.unibas.ch/personal/nimwegen/

computational approaches are made possible by the availability of large sequence acquisitions of complementary DNA (cDNA) sequences and expressed sequence tags (ESTs), derived from different tissues or cell lines, and set the benchmark more on “sensitivity” (breadth of detection) of AS events. Currently available data enable large-scale analysis of AS in human and mouse, and a few other organisms, and achieves on average a 200 to 300-fold

EST coverage per known human gene, respectively (89). Genome-wide screens for AS have been conducted, e.g. in (90-93), and related databases have been built (17,81,94) and (Table 1). While the number of AS genes and types of splice types can differ markedly between databases, due to differences in primary sequences and algorithms used to infer AS events, bioinformatics has shown its merit in, e.g.,

revealing underlying AS patterns and mechanisms in tissue-specificity, conservation and prediction of splice types in orthologous genes, or in disease-association (84,94,95,97-104).

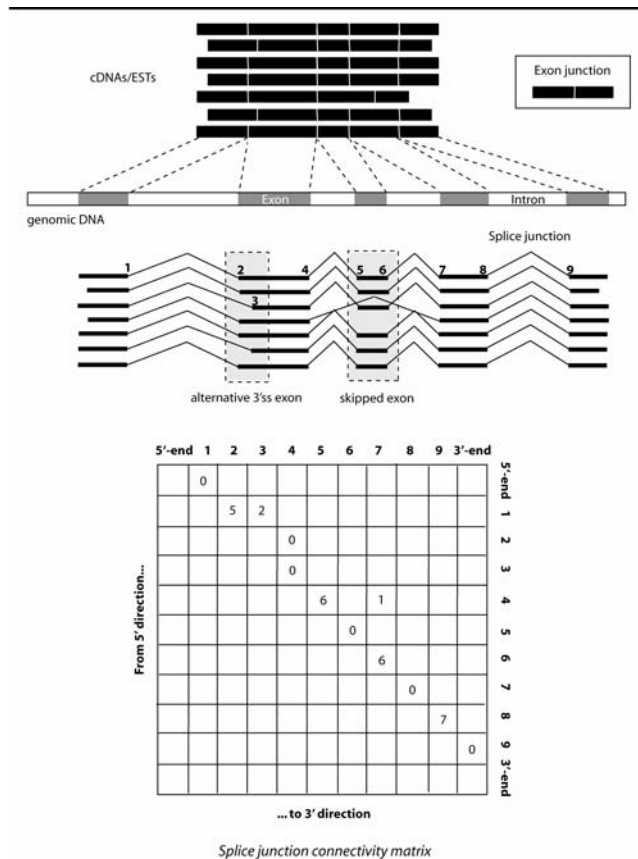
TRANSCRIPT-BASED IDENTIFICATION OF ALTERNATIVE SPLICING

Coupling transcript sequence to assembled genomes and gene loci provides one basic approach to annotate exon-intron structures, based on experimental sequence data, and to detect AS patterns. An alternative approach (when lacking assembled genomic DNA) consists in EST-to-EST multiple alignments. Large-scale spliced alignments of cDNAs and /or ESTs to genomic DNA are conducted using a variety of algorithms (some are listed in Table 2). As example, we consider the principal algorithmic flow that is used in the genome alignment program GENOA (94), similar to those used in other algorithms: (i) identify matches of identity between (repeat-masked filtered) cDNAs/ESTs and genomic DNA; (ii) determine spliced-alignments of significant matches of cDNAs to gene loci; (iii) identify matches of EST to aligned cDNAs, and (iv) splice-align significant matches of ESTs to cDNA-aligned gene loci. Typically, stringent quality filters are applied, e.g., on percent cDNA/EST aligned, percent sequence identity, minimum exon and introns sizes, maximum intron size, splice sites models, exclusion of genes that are subject to frequent DNA rearrangements (e.g., immunoglobulin genes), or the number of multiple hits of cDNAs/ESTs. Treating different transcript sequences (cDNAs, ESTs) with different alignment algorithms and stringency may be instructive, as recently demonstrated by the reannotation of human Chr22. Aligned EST sequences achieved a sensitivity of S_n ($= TP/all\ Positives$) = 0.77, while the specificity was scored as S_p ($= TP/predicted\ Positives$) = 0.42 (Dunham et al (2003), reannotation of human chromosome 22), highlighting the possible low specificity of expressed sequence tags.

After the genome-wide splice-alignment, the annotation of constitutive and alternative exons is the next step. Guided by what is known about different AS events, splice patterns are implemented as computational “rules”, and all transcripts aligned to a gene’s locus are scanned for alternatively spliced exons and retained introns. To this end, one can order observed splice junctions (SJs) and their frequency of occurrence, and construct a SJ matrix, elements of which have positive entries for observed SJs (Figure 17). In this representation, one can think of exons as “nodes” and SJs as “edges” connecting nodes, and by traversing through this (directed acyclic) graph one can capture different AS patterns (95). The above scheme is “exon-centric”, where splice patterns are individually evaluated for each exon. Exon-centric schemes cannot detect MXE events, or clusters of skipped exons. In order to pair-wise compare isoforms generated from one gene against one another, e.g., a heuristic solution is to capture the total number of SJs that differ between two transcripts and normalize it to the total number of SJ, within a region where both transcripts overlap genomic DNA (80). The clearest solution is to store in the SJ matrix in- and outgoing edges, hence allowing the construction of complete isoforms as a representative path through the graph (96). Accumulating splice variants and monitoring the respective frequency of alternatives, one obtains a bimodal distribution of the percentage of events, in particular for skipped exons (SEs). These two sub-populations are commonly referred to as “major-form” exon (one that predominantly includes an SE), and “minor-form” exon (one that predominantly excludes an SE).

Note that (i) the set of rules defines a search space, and hence the types of AS events detected; (ii) constitutive exons are often the “default” label of exons, and this label remains unless one of the rules are met for annotation as an alternative exon; and (iii) all transcript sequences may be treated equal and AS events are defined through pair-wise comparisons, or one may use the concept of a “canonical” sequence, against which the remaining transcripts are contrasted. Because parameters, quality filters, rules, and reference sequences differ between different transcript-based alignments, the inferred frequency of AS events differs as well. Current estimates put the frequency of human AS around 50%, while algorithms

Figure 17 with higher sensitivity boundaries estimate up to about 80% of AS. Clearly, the availability of a larger number of cDNAs and ESTs from a gene increases the chance of observing



alternative isoforms of that gene, the proportion of AS genes will tend to increase with increasing transcript coverage of genes. Probabilistic and sampling strategies have been discussed to correct or circumvent for this bias (80,95,97). Using equal coverage of EST sequences, it was thus found that the total frequency of AS events is roughly equal across animal genomes (97).

Based on compiled EST sequences and inferred AS patterns, two interesting problems have been addressed: tissue-specific splicing, and evolutionary conserved AS events.

Tissue-enriched AS A well-established paradigm in the field of RNA splicing is that usage of the choice of splice sites is often directed by the relative concentrations of specific SR and hnRNP proteins, and that tissue-specific AS genes are regulated by the functional antagonisms between SR and hnRNP proteins, differentially expressed in different cell- or tissue-types (63). In brain tissues, AS is a central genetic program, by which neurons acquire specialized molecular structures, and regulatory pathways are

required to receive and pass on signals. In one likely scenario, protein isoforms involved in neuronal functions are spliced in the same manner in neuronal and other cell types, except that certain alternative splicing pathways are only activated in neuronal cells (14). Consequently, in addition to the basal splicing machinery, there have to be special splicing factors, governing regulation in a cell-type specific manner. The family of Nova proteins is a class of RNA-binding proteins, exclusively detected in neural cell of the brain and localized to the nucleus, relevant for cell survival in postnatal stage of development. Other examples are given by neural-enriched proteins (PTB), or by the family of Hu and CELF proteins (61). In a genome-wide search and computational analysis (81) it was found that in the range of 10-30% of human AS genes display tissue-biased isoforms, where the human brain was exhibiting the highest number of AS genes, and the tissue categories eye/retina, muscle, skin, and testis showed stronger enrichments (% tissue-enriched AS/% total ESTs). Another study (80) compared AS patterns for individual categories (about 20), and found that brain, testis and liver displayed the most among these different types, and the highest rates of AS. When patterns were compared between categories (by a splice-junction "distance" measure), it found that AS clusters of the human brain, pancreas, liver, CNS, placenta, muscle and testis displayed more distinct patterns than the remaining tissues. Computational analyses were also indicative of tissue-biased AS in gene functions, such as transcription factors (92,98), or alternative promoters (73) and polyadenylation across tissues (99). Complementary to transcript-based approaches, AS events have now also been investigated using splicing-sensitive microarrays (21,100-

102), addressing technical features of different platforms as well as biological features of tissue-biased splicing, the functional impact of AS (e.g. via NMD), or disease-associated splicing.

One functional impact of AS is immediate: those skipped exons or splice site extensions of A3Es and A5Es which are not a multiple of three nucleotides, will create a frame-shift and hence lead with high probability to pre-termination stop codons in the alternative or downstream exon (103). Consequently, one can speculate about the different functions of exons that preserve and those that destroy the reading-frame. On the one hand, are frame-destroying exons used to subject transcripts to the NMD pathway? Several lines of evidence confirm the post-transcriptional regulation of gene expression via NMD, based on transcript data and on microarray data; however, data are so far inconclusive about the widespread utilization of this pathway as a regulatory or rather a “cleaning-up” mechanism for non-functional transcripts (101,103,104). On the other hand, frame-preserving skipped exons show an increased proportion among the set of minor form exons and are tissue-biased (105,106).

Table 3 **Evolutionary conserved AS** While initially used as a filter to reduce the number of non-authentic AS events, more recently the focus of comparative genomics has shifted toward

Algorithm	Web-based reference
Predicted about 1,000 AS events conserved between orthologous human-mouse genes, based on feature selection and use of rule-based computational model	http://www.tau.ac.il/~gilast/
Predicted a set of about 160 orthologous splicing-conserved alternative exons in the fly species <i>Drosophila melanogaster</i> and <i>D. pseudoobscura</i> , with 42% validated	http://penguin.uchc.edu/~intron/philipps
Predictive identification of about 2,000 alternative splicing-conserved exons, expressed in orthologous human-mouse genes, based on structural features and motifs trained in ACEScan algorithm	http://genes.mit.edu/acescan
Predicted about novel 50 splicing-conserved skipped exons in human ENCODE regions, and about two dozens conserved retained introns of orthologous human-mouse genes	http://tools.genome.duke.edu/generegulation
Designed kernels for the application of support vector machines to identify skipped exons in <i>C. elegans</i> . When applied to the worm genome, ~200 candidate alternative exons were identified	http://www.fml.tuebingen.mpg.de/~raetsch/

identifying AS events that are (i) splicing-conserved across species (“alternative conserved exons”, or ACEs), (ii) conserved, but species-specific in their splicing patterns, or (iii) newly created alternative exons that are present in orthologous genes of other lineages, respective to their position in the evolutionary tree. Given the differences in primary datasets and protocols for the inference of AS, it is difficult to estimate the fraction of splicing-conserved exons of orthologous human-mouse genes. Based on transcript-inferred and predicted ACEs, the proportion of ACEs is estimated to be larger than 15% (107), while other estimates

determine about 50% or more conservation (108,109). Interestingly, Modrek & Lee (110) observed a correlation between inclusion levels of skipped exons and their splicing-conservation between orthologous human and mouse genes: skipped exons of the major form were found to be highly splicing-conserved (and are thus ACEs), while exons of the minor form were often found to be lineage-specific. This may be indicative of exon-creation/loss events produced by adjustments to evolutionary change, and in the same context other evidence suggests that tissue-biased AS exons appear less abundant in the set of human-mouse ACEs.

There is a caveat to keep in mind here: Noise occurs both on the level of experiment as well as in the cell. The filters imposed on EST-based AS inference discussed above help to reduce experimental noise and eliminate unwanted contamination, e.g., by incompletely spliced products. However, given the weakness of the sequence signals such as splice sites, the splicing machinery itself is inherently noisy and prone to (reproducible) errors, and cellular mechanisms such as NMD have evolved to eliminate erroneously spliced products. As it has been provocatively put, one may be able to observe the alternative use of any possible splice site if one just sequences enough ESTs (104). The question therefore is how many of species specific events actually correspond to a specific function. In any case, wide-spread AS may serve as “evolutionary tunneling” (110) and provide an organism with a mechanism to quickly “explore” new

isoforms, only few of which eventually get fixed. In a larger context, this connects to the question of the evolution of AS and the structure of eukaryotic genes per se (111) (see below).

PREDICTION OF ALTERNATIVE EXONS AND RETAINED INTRONS

Given these problems with transcript-derived isoforms – the question of completeness as well as noise – recent years have seen a number of approaches, which directly aim at the identification of AS isoforms from genomic sequence (and neither transcript nor protein information). This is possible because exons affected by AS have specific characteristics when compared to constitutive ones. In mammals, the skipping of exons (SEs) is the predominant AS type, and sets of alternative splicing-conserved exons of orthologous human-mouse genes (or ACEs) could be identified and analyzed for such functional characteristics (112). Compared to constitutive splicing-conserved exons of orthologous genes, ACEs are predisposed, e.g., to be shorter, exhibit higher sequence-conservation and splice sites differences from the consensus, predominantly skip or include the exon, exhibit higher ESE frequencies, less SNPs, and are under higher natural selection pressure. Algorithms that exploit these features (107,113-115) mostly work on pairs of species (with the exception of (116), a single species approach) and come in two flavors: (i) exon *classification* algorithms, which take a *known* pair of orthologous *exons* to predict whether or not they are subject to AS; and (ii) exon *discovery* algorithms, which take a known pair of *introns* and scan it for the presence of (presumably skipped) exons. For category (i), the exons are known already, and this usually implies that the major isoform is exon inclusion; in comparison, category (ii) exons tend to be excluded in the majority of transcripts, which explains why they have not yet been annotated. Several algorithms have been developed (see Table 3) to predict mostly ACEs, but also conserved retained introns, between orthologous human-mouse genes, by using genomic sequence information. For instance, ACESCAN (a category (i) method based on a regularized least-squares regression algorithm) predicted about 4,000 exons to be ACEs in the human genome, and UNCOVER (a category (ii) algorithm based on a pair HMM) predicted about 50 new splicing-conserved skipped exons in human ENCODE regions and about 8,000 in the whole genome, as well as the surprisingly low number of two dozen of splicing-conserved coding retained introns across all human-mouse orthologous protein-coding genes. ACEs are more likely to preserve the reading frame and less likely to disrupt protein domains, are enriched in genes expressed in the brain, and in genes involved in transcriptional regulation, RNA processing and development.

In summary, an ideal splicing prediction algorithm would take the sequence of the pre-mRNA and be able to automatically predict which isoforms exist, and with additional information on the expression of all splicing factors, how frequently they are generated in a particular context. Compared to computational gene finders which strongly rely on coding content and conservation, such an approach would only make use of the information the cell has available at the time of splicing in the nucleus. In practice, such splicing simulators are only at the beginning, but preliminary successful results have been achieved with approaches that combine models for splice sites and other splicing regulatory motifs, such as ESEs and ESSs (79).

EVOLUTION OF ALTERNATIVE RNA SPLICING

RNA splicing is well conserved throughout evolution from yeasts to humans. The majority of introns are unlikely to have been inherited from recent common ancestral genes, such that multi-exon genes might have formed by gaining introns subsequently, with a sequence preference for MAG/GT and a bias for phase-0 introns, and splicing possibly originated from autocatalytic group II introns. On the one hand, *S.cerevisiae* has few introns (order of hundreds only) and no alternative splicing, and *S.pombe* has introns

in almost half of all its genes, but also does not exhibit alternative splicing. On the other hand, alternative splicing in metazoan genomes is one significant genetic pathway to regulate gene expression – so how did this plasticity come into existence and how did it manifest itself as a regulatory mechanism (9,111,117)? Alternative splicing is controlled at both the basal and the regulatory level. Based on comparative genomics analysis of splice sites and RNA-binding proteins in cells of different organismal complexity, two hypotheses have been made: (I) the more “*cis*-based” one emphasizes the change of produced splice patterns as an outcome of mutations in DNA, while (II) the more “*trans*-based” one emphasizes the evolution and function of splicing factors. Following (I), the production of weak splice sites, deviating from the consensus that base-pairs with U1 snRNA, leads to suboptimal sites that open up differential pathways for the splicing machinery to skip an internal exon during several splicing events. Consequently, the original isoform is kept, while an alternative isoform produces a new transcript. Following (II), the evolution of positive and negative regulatory splicing factors, such as SR proteins and hnRNPs, might have released the natural selection pressure on splice sites, while applying selection pressure on constitutive exons to undergo alternative splicing. Hypothesis (I) and (II) are not mutually exclusive: after the emergence of multi-exon genes (constitutively spliced), alternative exons evolved in a series of mutations that created splice sites with suboptimal recognition by the basal splicing machinery, and the interaction with RNA-binding proteins lead to different splice patterns, depending whether splice sites are paired via exon-definition (vertebrates, but also fly, worm, plants: predominantly creating exon skipping, alternative 5' splice sites or 3' splice sites variations), or intron-definition (fly, worm, and plants: intron retention). This is supported by the observation that splice sites of constitutive exons of higher eukaryotes are similar to lower eukaryotic organisms (yeasts), that splice sites of AS exons are predominantly found in higher eukaryotes, and that longer splice forms are often observed the ancestral one (111). In addition, splicing-regulatory signals and factors also play a role in ensuring the correct linear order of exons, after intron removal, and ESEs can act as barriers to prevent exon skipping (however, e.g., ESEs are also found in intronless genes, introns, and are implicated in other functions besides recruiting SR proteins) (12). Interestingly, the subdivision of introns by recursive splicing at non-exonic splice site sites has also been shown to be advantageous for introns of long size ranges (flies) (118). From a population-genetic perspective (119,120), the creation and retention of introns can be modeled by random genetic drift and weak mutation pressure against intron-containing alleles, and makes testable predictions (e.g., in terms of unicellular and multi-cellular population sizes, intron phase biases, and nonsense-mediated mRNA decay).

Thus, a constitutive exon can undergo different selection and become alternative, or alternative exons can be (newly) created via intron gain, exonization (due to interspersed repeats) and/or cryptic spliced sites of intronic sequences, or exon duplication. A very fruitful playground for such investigations are fungal genomes, for which a large number of completed genomes spanning a variety of evolutionary distances is available, which allows e.g. to distinguish between gain and loss of introns in the genomic sequence (121). Exonization describes an evolutionary pathway to create specific isoform diversity, given the presence and activity of interspersed repeats in a genome lineage. It acts by acquiring mutational changes to create active 5' splice sites. This has e.g. been detected for the primate-specific Alu repetitive sequence elements, which gave rise to about ~5% of AS exons, and is estimated to affect several thousand Alu repeats within the human genome with the potential to create additional alternative exons (122). Exon duplication has been estimated to affect 7-10% of all genes in the human, worm, or fly genome, respectively, where tandemly duplicated exons are often involved in isoforms of mutually exclusive exons (123).

Consequently, one might expect AS to be associated with the recent creation/loss of exons (potential for species-specific exons), and that the gene expression levels of splicing-conserved exons are correlated

across different species and cell- or tissue types. Yet the picture emerging is a bit subtler than that. For instance for exon skipping (SE), which is the predominant vertebrate splice form and been studied in more detail than other splice patterns, coding and splicing-conserved SE (ACEs) show a small preference for being “modular”, i.e., the reading-frame preservation ratio (number of modular SEs/non-modular SEs) is slightly about $\frac{1}{2}$, while minor-form coding ACEs exhibit a larger frame preservation ratio than in major-form; and modular SEs exhibit changes in their percentage of inclusions in a tissue-enriched pattern. That is, with the exception of minor-form exons (low conservation of AS), several lines of evidences gathered from computational analysis so far are indicative of a marked number of AS events that are splicing-conserved from man to mice (estimated to be above 50%). To conceptualize the evolution of AS, Modrek & Lee (110) proposed a model, which is based on fitness landscape adaptive walks. Differential splicing of a newly created exon can “dampen” the impact on the transcript by creating different isoforms, and convert an isoform with a low fitness value toward higher values, given that accumulated mutations become beneficial.

AFFECTS OF RNA SPLICING AND ALTERNATIVE SPLICING ON HUMAN DISEASE

The disruption of specific splicing and/or AS events has been implicated in human RNA pathogenesis. At

Table 4 least 10-15% of genetic diseases may be associated with splicing mutations in cis, affecting constitutive or alternative splice site recognition, or in trans, affecting the splicing machinery or auxiliary splicing factors (12,60).

Splicing mutations associated with altered splice variant

ATP7A (Myotonic dystrophy), *SMN2* (Spinal muscular atrophy), *CFTR* (Cystic fibrosis), *IKBKAP* (Familial disautonomia), *HEXB* (Sandhoff), *ADA*, *ATM*, *BRCA1*, *F8*, *FAH*, *FBN2*, *HPRT*, *MAPT*, *MLH1*, *PDHA1*, *PMM2*, *APC*, *CYP27A1*, *NF1*, *PAH*, *PTS*

Disease associated with changes in relative levels of isoforms

Tau (Fronto-temporal dementia and Parkinsonism linked to Chr17, FTDP-17), *NF2* (Neurofibromatosis II), *WT1* (Wilms tumor), *BRCA1* (Breast/ovarian cancer), *BRCA2* (Breast cancer), *CD44* (Renal, lung and urothelial cancers, Gastric cancer, thyroid cancer), *FHIT* (Lung cancer), *MDM2* (Invasive breast cancer, bone tumor), *Bin1* (melanoma), *Bcl-2* (Prostate cancer, lymphoma, gastric carcinoma), *Bcl-x* (lymphoma, breast cancer), *Bax* (Oral and oropharyngeal cancer)

Disrupted splicing regulatory motifs may be located in either intronic or exonic sequences. The severity of a disease may be associated with the level of correctly and aberrantly spliced mRNA transcribed from genes carrying splicing mutations, and hence with the ratio of isoforms, caused by changes in the level of splicing factors, which modulate the splicing patterns of

disease-associated genes (Table 4). A few examples illustrate disease-associated splicing (see Table 5).

Cystic fibrosis (CF) is an autosomal recessive disorder, caused by the loss-of-function of the CF transmembrane conductance regulator (*CFTR*) gene. Two polymorphisms, located between the

Table 5 branch point and AG/, in *CFTR* are associated to atypical CF phenotypes and directly affect

Splicing trans-acting factors affecting the level of mRNA isoforms leading to disease

Splicing factor	Disease	Affected genes
SC35	Cystic fibrosis	<i>CFTR</i>
SCNM1	Lethal neurological disease	<i>Scn8a</i>
MBNL	Myotonic dystrophy	<i>Clcn1</i> , <i>Tnnt2</i> , <i>Tnnt3</i>
CUG-BP1/CELF1	Myotonic dystrophy	<i>Tnnt2</i> , <i>Mtmr1</i> , <i>Clcn1</i>
ASF/SF2	Lethal cardiomyopathy	<i>CaMKII-Delta</i>
Nova-1	Lethal motor deficit	<i>GlyR-Alpha2</i> , <i>GABA</i> , <i>JNK2</i> , <i>Neogenin</i> , <i>Gephyrin</i>
Htra2-Beta1	Cystic fibrosis	<i>CFTR</i>
SMN1	Spinal muscular atrophy	
SMN2	Spinal muscular atrophy	
TP73Lp	Hay-Wells syndrome	<i>FGFR2</i>

splicing of exon *E9*, by causing exon skipping. Spinal muscular atrophy (SMA) is a neurodegenerative disease, caused by the loss of the survival motor neuron 1 (*SMN1*) gene, which interacts with components of the RNP complexes. Human

individuals possess a linked, nearly identical version of the gene, *SMN2*, which generates a functional SMN protein, but at levels insufficient to compensate for the loss of *SMN1*. Several missense, nonsense and splicing mutations disrupt the *SMN1* protein. In myotonic dystrophy (DM), disease phenotypes have

been directly associated to disrupted regulation of AS. DM is an autosomal dominant disorder with a worldwide incident of about 1 out of 8000 individuals. DM1 is caused by a CTG-repeat expansion in the 3' UTR of the DM protein kinase (*DMPK*), and repeat length correlate with the disease severity. DM2 is caused by a CCTG-repeat expansion in an intron in the *ZNF9* gene. Repeats form hairpins that are bound by the dsRNA-binding protein *MBNL* and subsequently elevate the expression levels of other splicing factors that are members of the family of *CELF* proteins. AS has also been associated with different types of cancers, where different isoforms are differently expressed in normal versus cancerous cells. For instance, *Bcl-x* is a member of the *Bcl-2* family of apoptotic genes, which play central roles in apoptotic response pathways. The *Bcl-x* gene encodes two proteins, a short (*Bcl-x_S*) and a long splice variant (*Bcl-x_L*) by alternative splicing, which have antagonistic functions. The shorter splice variant is highly expressed in many types of cancers. Xu & Lee (81) surveyed about two million human expressed sequence tags derived from normal and tumor tissues for splice variants that are differentially expressed in tumors, and found about 320 genes that display cancer-specific splice variants. Whether AS can cause cancer is difficult to determine as are the several process stages toward cancer in general; yet, it opens up possibilities for diagnostic markers and cancer treatment by restoring function of gene splicing with splicing modulation strategies (e.g., antisense oligomers, RS-domain attachment, overexpression of splicing factors, or compounds that affect phosphorylation of splicing factors).

AVAILABLE RESOURCES

Several heuristic, greedy, and dynamic programming-based algorithms have been implemented and made available to find matches between transcribed sequences and to spliced-align sequences to genomic DNA (Table 2). Conveniently, algorithms have been put up on web servers, or are otherwise available for download or from authors for local installation. In addition, standard genome browsers such as provided by UCSC, Ensembl, and NCBI often integrate alternative mRNA isoforms and provide genome-wide annotations. On the flipside, problems may arise from alignments to non-authentic exons, small exons (about 20-30 base pair long), paralogous genes/exons, steadily increasing numbers of ESTs (about seven million human ESTs), and aberrantly spliced messages.

Design and implementation of bioinformatics tools that reflect the demand of steadily modified and changed models for the regulation of alternative splicing is a significant challenge for bioinformatics. To this end, several databases for storage and retrieval of AS patterns have been developed (Table 1). While initially databases recorded alternative exons, often driven by available transcript data (e.g., from human, mouse, fly), recent ones including ASAP, ASD, or HOLLYWOOD attempt to more adequately reflect the current knowledge and incorporate comparative genomics, splicing-conserved exons, EST-derived tissue types, and the annotation of ESE and ESS elements. Web-based interfaces for browsing AS patterns vary from one database to another, and offer different levels of summary information and viewing, some including linked GenBank, dbEST documents, alternative displays layered onto the UCSC Genome Browser.

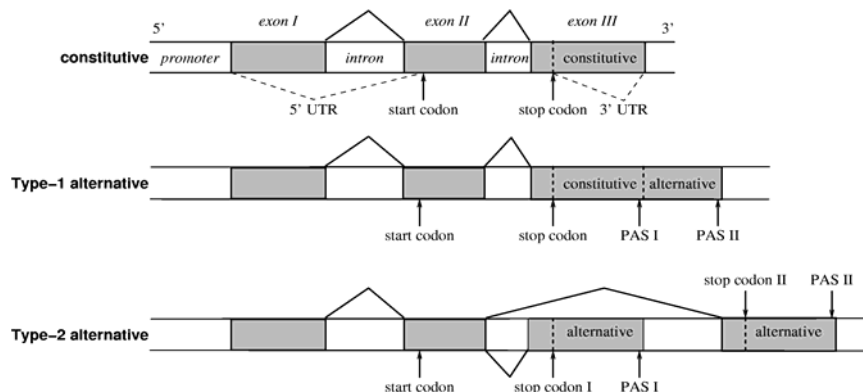
Overview of Other Mechanisms: Beyond Alternative Splicing

The function and regulation of RNA transcripts take place in time and space within the post-transcriptional cellular domain, and constitute the transfer of the genetic information as well as its management. The outcome of gene expression (protein production) is ultimately determined at the post-transcriptional level. Gene regulation does however not stop at the level of alternative splicing, and recent years have seen an increased appreciation for the importance of post-transcriptional regulatory

mechanisms other than splicing (124). Well known mechanisms include control of export, localization, and stability of the messages.

UTRs AND RNA-BINDING PROTEINS

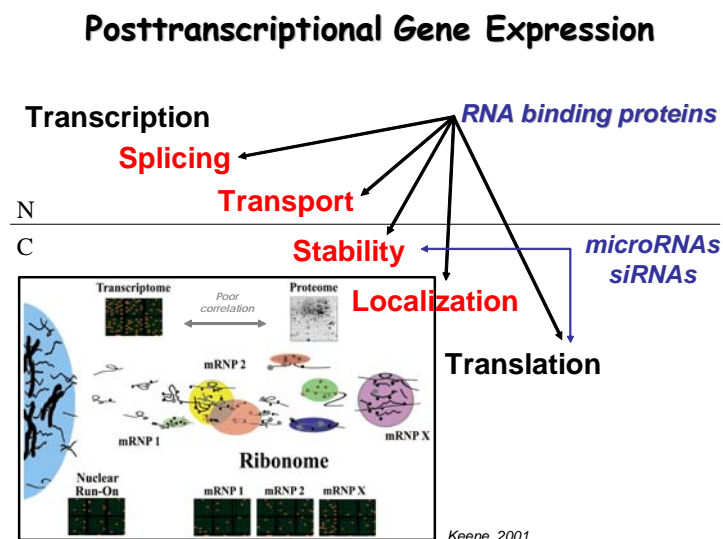
Figure 18 The *cis*-regulatory sequences of many of the known post-transcriptional events are found within the 3'-untranslated regions (3' UTRs) between a stop codon and a polyadenylation site (PAS) of a messenger RNA (125). Alternative polyadenylation can lead to changes in the 3' UTR



portion of an mRNA, which opens up the possibility to specifically include or exclude individual *cis*-regulatory sites in different isoforms of the same gene. The increasing number of reports addressing the high frequency of variation in the

3' UTR (126-128) makes it necessary to address this issue: more than half of the genes are estimated to have variable 3' UTR sequences, and a number of individual cases show that post-transcriptional gene regulation can be affected by alternative polyadenylation. We can distinguish two basic classes of alternative 3' UTRs: those arising from the use of different PAS in the same terminal exons, and those arising from alternative terminal exons (Figure 18). In the first case, the UTR will consist of a 5'-end "constitutive" and a 3'-end "alternative" part. The alternative part can be of different length in case of

Figure 19 | adopted from Keene (2001)



multiple alternative PAS. In the latter case, the alternative 3' UTRs will generally not share common sequence, and the choice of the downstream final exon is likely coupled with suppression of the 3' splice site of the more upstream ones.

Computational biology of post-transcriptional regulatory control is still overshadowed by transcription, but a recent number of studies have performed genome-wide analyses of motifs in 3' UTRs (129-131). In comparison to transcription factor binding sites, RNA binding sites of 3' regulatory trans-factors may also be assisted by structural motifs caused by specific RNA secondary structure. Specific search tools which combine

structural with sequence motifs are available (132,133). Available resources integrate information on UTR sequences and regulatory motifs in them (134,135).

RNA binding proteins have been known for a long time to play an important role as trans-factors interacting with functional sequence elements in the 3' UTR. Similar to chromatin immunoprecipitation to identify targets of transcription factors, several laboratories have devised a procedure called ribonucleoprotein immunoprecipitation (RIP-Chip) which can be used to immunoprecipitate RNA-binding proteins (RBPs), and to identify the associated RNAs using high-throughput platform such as microarrays (136,137). Using this technology, it was demonstrated that mRNAs associated with specific RBPs are functionally related, such as having common RNA stabilities or encoding groups of proteins with related functions (138,139). Figure 19 summarizes the various components of the post-splicing regulatory machinery.

MICRO-RNAs: GENES AND TARGETS

The regulatory mechanism currently in the center of the spotlight is expression control by a class of tiny non-coding RNAs called microRNAs (140,141). MiRNAs are the most prominent example of several classes of non-coding RNA genes identified in the last few years. They are abundant and widely conserved throughout all multi-cellular eukaryotes, and comprise hundreds of members in mammalian organisms alone. The functional mature miRNA is only ~21-25 nucleotides long and is excised out of a precursor stem-loop of about 70 nucleotides, which in turn is part of a potentially several kb long primary transcript. The structure of primary miRNA transcripts shows a surprising variety: miRNA precursor foldbacks can be part of an unspliced larger transcript, in the intronic and exonic portions of non-coding transcripts, as well as in the introns of coding host genes (142). Because of the large number of fortuitous occurrences of foldbacks of an appropriate size for miRNA precursors, the first generation of computational gene finders for miRNAs exploited the wide-spread conservation of this class of genes and relied heavily on features based on cross-species conservation (143-145). Recent predictors have now been adapted to work for on individual genomes, especially important to locate miRNAs in rapidly evolving viral genomes (146). The idea here is to look for stable predictions of secondary structures, which are not altered by a particular choice of genomic sequence window including the foldback.

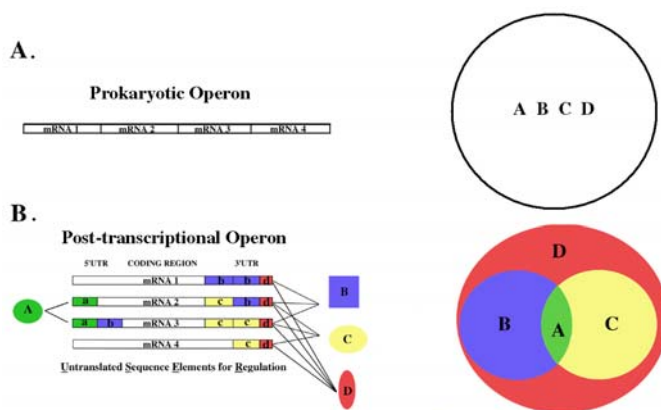
MiRNAs have a regulatory influence on a large fraction of protein-coding genes and are important in a variety of scenarios, with an important role emerging in developmental transitions and differentiation, and establishing cell identity (147-149). According to our current understanding, animal microRNAs will bind to protein coding genes mostly by pairing with complementary "target sites" which are located in the 3' UTR, either leading to degradation of the transcript by the RNAi machinery, repressing of translation by sequestering of transcripts in cellular compartments isolated from the translation machinery, or degradation of the poly(A) tail (150). A given UTR may have several target sites, either of the same or of different miRNAs, implicating the possibility of combinatorial control. Computational target prediction algorithms achieve a high signal to noise ratio of real versus spurious hits by locating a perfect "seed match" complementary to the 5' end of the mature miRNA, and by requiring conservation of this seed match across several species. Target prediction algorithms vary somewhat in the details, e.g., in the exact location of the seed match (usually, to bases 2-7 or 2-8), whether only Watson-Crick base pairs or also G-U base pairs are allowed, and whether the identification of conserved targets relies on pre-computed alignments or is carried out independently in the individual UTRs. Depending on the exact choice of such characteristics, the number of predicted targets and the associated SNR will change somewhat. In any scenario, the fraction of protein coding genes targeted by animal miRNAs is currently estimated at about 30% (151).

NETWORKS OF POST-TRANSCRIPTIONAL GENE-REGULATION: SPLICING AND MICRORNAs

Genes encoding different protein isoforms by means of alternative splicing make use of an RNA code, the molecular determinants of which – “language, words, syntax” – we are just beginning to appreciate and understand. An illustrative, clear example has been revealed in (14): the *trans*-acting splicing factors hnRNP A1 binds to the motif UAGGG[U/A], hnRNP H binds to GGGG monomers, and the code is that the ternary juxtaposition of two exonic UAGG motifs and a 5' splice proximal intronic GGGG motif function cooperatively to silence an exon (via hnRNP A1), while this can be counterbalanced by hnRNP H. More complicated examples exist, but it would be fruitful to decode prototypic modes of this type (causative, less correlative) to advance the computational biology of mechanisms of gene regulation.

Upstream in this picture, splicing factors themselves may be differentially regulated in different tissues or in response to different stimuli at the level of transcription, splicing, or translation, and are frequently regulated by post-transcriptional modifications such as phosphorylation. An instructive example is provided by the negative feedback auto-regulation among splicing factors, e.g., as observed for the neuronal splicing factor *Nova* or the ubiquitously expressed splicing factor PTB, which act on their own expression by acting as a splicing repressors, respectively. In addition, SR proteins shuttle back and forth between nucleus and cytoplasm, possibly acting as “messengers” between two compartments, indicative of an active homeostatic process for the network of isoforms and regulators. Downstream, alternative splicing may affect the stability of other alternative transcripts, e.g., by subjecting messages to the NMD pathway, and frequently alters functional properties of proteins. A genome-wide microarray screen for tissue-specific targets of *Nova-1*, showed that this splicing factor, only found in the brain, controls RNA splicing of about 50 transcripts to produce proteins not found in other tissues, many of them suggestive to work in nerve cells to transmit signals across the synapse.

Figure 20 Protein coding genes are targeted by miRNAs, but the miRNA genes themselves are under control of transcription factors and transcribed by RNA polymerase II, just as “normal” protein coding genes are. MiRNAs can therefore naturally augment regulatory networks which had previously mostly focused on transcription factors, for example as components of negative feedback loops involving transcription factors (152). The availability of expression profiles from microRNA microarrays provides us with a resource to begin deciphering these networks on a genome wide scale (153,154). In addition, many mammalian



microRNAs are located within introns of host genes, but the possible interplay of splicing vs. microRNA processing is largely unexplored at this point.

Strictly focusing on the posttranscriptional level, there is increasing evidence supporting that each RNA regulatory step is physically linked to the next step (4,155). For example, splicing factors associate with nascent transcripts before synthesis is completed and nuclear export factors are recruited by splicing factors during splicing to expedite RNA movement to the cytoplasm. Also, a set of RNA binding proteins termed the exon-junction-complex (EJC) binds near exon-intron junctions of mRNA and plays a role in determining the final fates of mature transcripts later in the cytoplasm (156-158). This interconnectivity

helps maintain continuity of the pathway of posttranscriptional gene expression in the sense of a molecular conveyer belt. For example, RNAs that encode components of macromolecular complexes such as mitochondria or ribosomes are often regulated by a common RBP to coordinate its production, and this phenomenon has been sometimes described as “post-transcriptional operon” (Figure 20), in analogy to the transcriptional operons in prokaryotes. Indeed, these processes can occur synchronously in a coordinated fashion (124,136,137,158-160). Thus, RNA splicing, export, stability, localization and translation are susceptible to being coordinately regulated as the genetic information in the form of RNA is physically organized to encode a proteome that is unique to each cell type.

Posttranscriptional regulation is highly dynamic in response to environmental signals. It was found that RNP complexes become “remodeled” following nutrient depletion or the induction of differentiation (137,161). It strongly suggested that these changes were not solely the result of changes at the levels of transcripts due to effects on transcription, but were the result of specific mRNAs entering into the RNP complexes while other mRNAs exited the complexes. It is believed that RNP remodeling provides a mechanism of coordination of posttranscriptional gene expression. RNA-protein interactions are capable of being “rewired” in evolution to serve different necessary functions (129,138,161). In fact, with the advent of the nuclear membrane in the eukaryotes, the number of RNA binding proteins increased dramatically, 5' and 3' UTR elements expanded, and proteins became increasingly multifunctional (138). An example of rewiring is the yeast RNA binding protein, *Cth2*, which coordinately destabilizes mRNAs involved in the response to iron depletion (161), while its human ortholog, tristetraprolin (*TTP*), coordinately destabilizes mRNAs encoding cytokines in human macrophage (162). Both *Cth2* and *TTP* bind to the identical AU-rich 3' UTR sequence elements in their target mRNAs (which at the same time, serve as target sites of the miRNA miR-16). Acting as analogous interacting components in evolution, these interactions have been used for entirely distinct functional outcomes in evolution to coordinate gene expression in response to different environmental signals. Another example is that the mammalian HuR protein and its counterpart in yeast, Pub1p, both bind to AREs in the 3' UTRs of targeted mRNAs. Mammalian HuR binds and regulates functionally related early response transcripts, while *Pub1p* regulates mRNAs encoding proteins involved in ribosome biogenesis (163). In both of these cases, and in many other examples, this kind of “rewiring” of the adaptor elements provides an evolutionarily exchangeable mechanism to serve many different coordinated functional outcomes across species (129).

Conclusions and Open Issues

With the completion of a number of mammalian genome projects, the attention has shifted from cataloguing the parts list of components toward analyzing how these components actually function mechanistically, interact, and evolved over time. After an initial focus on transcriptional control of gene expression, the control mechanisms on the post-transcriptional level are starting to now emerge as a fruitful arena for computational biology.

To have deciphered the RNA splicing code is to be able to explain the relationship between exon-intron structures, splice site choice, *cis*-regulatory elements and their interaction to *trans*-acting factors in different cell-types, developmental stages, and response to extracellular signals. To this end, splicing analysis has exploited the abundance of cDNA and EST sequences, and began by inferring and classifying AS patterns, and monitoring their frequency of occurrences in different species; subsequently, patterns were studied for the evolutionary conservation of AS derived from orthologous genes and for differences in normal versus disease-associated splicing; beyond expressed sequence tags, and more

recently, splicing-sensitive microarrays have provided means to observe “exon expression” across different tissues.

Building upon this, problems that might call for the computational biologist’s attention and initiative include: (1) **Treatment of noise inherent in biological systems and data acquisition**, it is still hard to distinguish authentic from pseudo events, and functional from non-functional ones; (2) **Combinatorial control**, linking singular AS events to the generation of functional RNA isoforms; (3) **Complete catalogue of functional splicing-regulatory elements**, the identification and verification of signals in exons and introns, and their interrelatedness, are still in an early stage of sufficient description and are required for defining the “words” of the molecular language; (4) **Continuous versus discrete models**, AS events are typically regulated on a continuous scale and not as “switch” decisions, and consequently the level of each RNA isoform, in different cell- or tissue types, and not solely its mere presence is important. This is a consequence of levels of several splicing factors, where the relative ratios to each other determine the outcome of the competition for splice site choice; (5) **Mechanistic models of splicing regulation**, splicing factors create “zones of silencing” for negative regulation of splicing, e.g., by looping out exons or coating exons, steric interference, spliceosome incompatibility, or docker-selector sequence pairing, and the nature of these modes of action need be appropriately reflected into the “functionality” of sequence elements; (6) **Models and pathways**, many steps of interconnected gene regulation challenge the development of appropriate computational models and require that individual contributions of each component can be integrated; (7) **Evolution of splicing and alternative splicing**, the evolutionary history of nucleotides used for the production and function of isoforms; (8) **Simulation of splicing pathways *in silico***, the knowledge of gene structure, functional elements and expression of splicing factors need to be translated into computer programs that simulate the correct removal of introns and the annotation of alternative exons. Much needs to be done for computational biologists to get from “a combination of staring at the sequences for months and sheer luck” (though admirable) as a cited method for, e.g., the identification of functional elements (164) to computational pipelines that can integrate the transcriptional and post-transcriptional levels to provide accurate models of gene regulation in complex eukaryotes.

Acknowledgements

UO is an Alfred P Sloan fellow in Molecular Biology, and would like to thank his group and collaborators at Duke University. DH would like to acknowledge support from the Institute of Molecular Pathology (IMP), and would like to thank Barry Dickson (IMP). Both DH and UO are grateful to Chris Burge and their former colleagues in the Burge lab (Department of Biology, MIT).

Website references

<http://genome.cs.ucsc.edu> - GoldenPath Genome Browser
<http://www.ensembl.org> - Ensemble Genome Browser
<http://exon.cshl.org.edu/ESE> - ESEfinder web server
<http://genes.mit.edu/burgelab> - RESCUE-ESEs and ESS web server
<http://www.sanger.ac.uk/Software/Rfam/> - RNA family database, including microRNAs
<http://www.ba.itb.cnr.it/UTR/> - UTR resource (*cis*-regulatory elements and *trans*-acting factors)
<http://polya.umdj.edu/> - database of (alternative) polyadenylation sites
<http://pictar.bio.nyu.edu/> - miRNA target prediction server (fly, worm, vertebrates)
<http://genes.mit.edu/targetscan/> - miRNA target prediction database (mammals)
<http://www.tbi.univie.ac.at> - Vienna RNA structure prediction package
<http://bioinfo.mat.rpi.edu> - mfold prediction server
http://www.ncbi.nlm.nih.gov/dbEST/dbEST_access.html - dbEST data
<http://mgc.nci.nih.gov> - Mammalian gene collection
<http://www.ncbi.nlm.nih.gov/RefSeq> - Reference sequences
<http://www.ncbi.nlm.nih.gov/CCDS> - Consensus protein coding sequences
<http://dbtss.hgc.jp> - Database of transcriptional start sites
<http://www.peptideatlas.org> - Peptide sequence data
<http://www.ncbi.nlm.nih.gov/geo> - Gene expression data
http://fantom31p.gsc.riken.jp/cage_analysis - Capped gene expression data
<http://cgap.nci.nih.gov/SAGE> - Serial analysis of gene expression data
<http://www.geneontology.org> - Gene Ontology
http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html - Donor splice site scoring program
http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html - Acceptor program
<http://www.genet.sickkids.on.ca/~ali/splicesitescore.html> - Splice site scoring program

References

1. Maniatis, T. and Reed, R. (2002) An extensive network of coupling among gene expression machines. *Nature*, **416**, 499-506.
2. Black, D.L. (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*, **72**, 291-336.
3. Lopez, A.J. (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu Rev Genet*, **32**, 279-305.
4. Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236-243.
5. Zhou, Z., Licklider, L.J., Gygi, S.P. and Reed, R. (2002) Comprehensive proteomic analysis of the human spliceosome. *Nature*, **419**, 182-185.
6. Rappsilber, J., Ryder, U., Lamond, A.I. and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res*, **12**, 1231-1245.
7. Nilsen, T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147-1149.
8. Jurica, M.S. and Moore, M.J. (2003) Pre-mRNA splicing: awash in a sea of proteins. *Mol Cell*, **12**, 5-14.
9. Burge, C., Tuschl, T. and Sharp, P. (1999) In Gesteland, R., Cech, T. and Atkins, J. (eds.), *The RNA world*. 2 ed. Cold Spring Harbor Laboratory Press, pp. 525-560.
10. Maniatis, T. (1991) Mechanisms of alternative pre-mRNA splicing. *Science*, **251**, 33-34.
11. Matlin, A.J., Clark, F. and Smith, C.W. (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, **6**, 386-398.
12. Cartegni, L., Chew, S.L. and Krainer, A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, **3**, 285-298.
13. Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet*, **17**, 100-107.
14. Grabowski, P.J. (2004) A molecular code for splicing silencing: configurations of guanosine-rich motifs. *Biochem Soc Trans*, **32**, 924-927.
15. Schwerk, C. and Schulze-Osthoff, K. (2005) Regulation of apoptosis by alternative pre-mRNA splicing. *Mol Cell*, **19**, 1-13.
16. Nakahata, S. and Kawamoto, S. (2005) Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. *Nucleic Acids Res*, **33**, 2078-2089.
17. Lareau, L.F., Green, R.E., Bhatnagar, R.S. and Brenner, S.E. (2004) The evolving roles of alternative splicing. *Curr Opin Struct Biol*, **14**, 273-282.

18. Kornblihtt, A.R. (2005) Promoter usage and alternative splicing. *Curr Opin Cell Biol*, **17**, 262-268.
19. Calvo, O. and Manley, J.L. (2003) Strange bedfellows: polyadenylation factors at the promoter. *Genes Dev*, **17**, 1321-1327.
20. Proudfoot, N. (1996) Ending the message is not so simple. *Cell*, **87**, 779-781.
21. Yeakley, J.M., Fan, J.B., Doucet, D., Luo, L., Wickham, E., Ye, Z., Chee, M.S. and Fu, X.D. (2002) Profiling alternative splicing on fiber-optic arrays. *Nat Biotechnol*, **20**, 353-358.
22. McCullough, A.J. and Berget, S.M. (1997) G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol Cell Biol*, **17**, 4562-4571.
23. Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J Biol Chem*, **270**, 2411-2414.
24. Bell, M.V., Cowper, A.E., Lefranc, M.P., Bell, J.I. and Screaton, G.R. (1998) Influence of intron length on alternative splicing of CD44. *Mol Cell Biol*, **18**, 5930-5941.
25. Eperon, I.C., Makarova, O.V., Mayeda, A., Munroe, S.H., Caceres, J.F., Hayward, D.G. and Krainer, A.R. (2000) Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Mol Cell Biol*, **20**, 8303-8318.
26. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
27. Amir-Ahmady, B., Boutz, P.L., Markovtsov, V., Phillips, M.L. and Black, D.L. (2005) Exon repression by polypyrimidine tract binding protein. *Rna*, **11**, 699-716.
28. Ibrahim el, C., Schaal, T.D., Hertel, K.J., Reed, R. and Maniatis, T. (2005) Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. *Proc Natl Acad Sci U S A*, **102**, 5002-5007.
29. Nussinov, R. and Jacobson, A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, **77**, 6309-6313.
30. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, **101**, 7287-7292.
31. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31**, 3406-3415.
32. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, **9**, 133-148.
33. McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105-1119.
34. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, **31**, 3429-3431.
35. Ding, Y., Chan, C.Y. and Lawrence, C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna*, **11**, 1157-1166.

36. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911-940.
37. Zuker, M., Jaeger, J.A. and Turner, D.H. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res*, **19**, 2707-2714.
38. Rivas, E. and Eddy, S.R. (2000) The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, **16**, 334-340.
39. Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, **2**, 919-929.
40. Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
41. Baldi, P. and Brunak, S. (2001) *Bioinformatics: The Machine Learning Approach*. 2nd ed. MIT Press, Cambridge.
42. Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.
43. Mount, D.W. (2004) *Bioinformatics: Sequence and Genome Analysis* 2nd ed. CSHL Press.
44. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, **8**, 967-974.
45. van Nimwegen, E., Paul, N., Sheridan, R. and Zavolan, M. (2006) SPA: A Probabilistic Algorithm for Spliced Alignment. *PLoS Genetics*, **2**, e24.
46. Dewey, C., Wu, J.Q., Cawley, S., Alexandersson, M., Gibbs, R. and Pachter, L. (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res*, **14**, 661-664.
47. Meyer, I.M. and Durbin, R. (2002) Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, **18**, 1309-1318.
48. Alexandersson, M., Cawley, S. and Pachter, L. (2003) SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res*, **13**, 496-502.
49. McAuliffe, J.D., Pachter, L. and Jordan, M.I. (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, **20**, 1850-1860.
50. Siepel, A. and Haussler, D. (2004) Computational identification of evolutionarily conserved exons. *Annual Int'l Conf. on Research in Computational Biology (RECOMB)*, **8**, 177-186.
51. Iseli, C., Jongeneel, C.V. and Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*, 138-148.
52. Cawley, S.L. and Pachter, L. (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19 Suppl 2**, II36-II41.

53. Burges, C.J. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167.
54. Schoelkopf, B., Tsuda, K. and Vert, J.-P. (2004) *Kernel Methods in Computational Biology* MIT Press, Cambridge.
55. Lanckriet, G.R., De Bie, T., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004) A statistical framework for genomic data fusion. *Bioinformatics*, **20**, 2626-2635.
56. Zheng, Z.M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci*, **11**, 278-294.
57. MacDougall, C., Harbison, D. and Bownes, M. (1995) The developmental consequences of alternate splicing in sex determination and differentiation in *Drosophila*. *Dev Biol*, **172**, 353-376.
58. Modafferi, E.F. and Black, D.L. (1997) A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. *Mol Cell Biol*, **17**, 6537-6545.
59. Xie, J. and Black, D.L. (2001) A CaMK IV responsive RNA element mediates depolarization-induced alternative splicing of ion channels. *Nature*, **410**, 936-939.
60. Faustino, N.A. and Cooper, T.A. (2003) Pre-mRNA splicing and human disease. *Genes Dev*, **17**, 419-437.
61. Dredge, B.K., Polydorides, A.D. and Darnell, R.B. (2001) The splice of life: alternative splicing and neurological disease. *Nat Rev Neurosci*, **2**, 43-50.
62. Garcia-Blanco, M.A., Baraniak, A.P. and Lasda, E.L. (2004) Alternative splicing in disease and therapy. *Nat Biotechnol*, **22**, 535-546.
63. Hanamura, A., Caceres, J.F., Mayeda, A., Franza, B.R., Jr. and Krainer, A.R. (1998) Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. *Rna*, **4**, 430-444.
64. Venables, J.P. (2004) Aberrant and alternative splicing in cancer. *Cancer Res*, **64**, 7647-7654.
65. Krawczak, M., Reiss, J., Cooper, D.N. (1992) The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Human Genetics*, **90**, 41-54.
66. Cartegni, L. and Krainer, A.R. (2003) Correction of disease-associated exon skipping by synthetic exon-specific activators. *Nat Struct Biol*, **10**, 120-125.
67. Zhu, J., Shendure, J., Mitra, R.D. and Church, G.M. (2003) Single molecule profiling of alternative pre-mRNA splicing. *Science*, **301**, 836-838.
68. Soret, J., Bakkour, N., Maire, S., Durand, S., Zekri, L., Gabut, M., Fic, W., Divita, G., Rivalle, C., Dauzonne, D. *et al.* (2005) Selective modification of alternative splicing by indole derivatives that target serine-arginine-rich protein splicing factors. *Proc Natl Acad Sci U S A*, **102**, 8764-8769.
69. Blanchette, M., Labourier, E., Green, R.E., Brenner, S.E. and Rio, D.C. (2004) Genome-wide analysis reveals an unexpected function for the *Drosophila* splicing factor U2AF50 in the nuclear export of intronless mRNAs. *Mol Cell*, **14**, 775-786.
70. Park, J.W. and Graveley, B.R. (2005) Use of RNA interference to dissect the roles of trans-acting factors in alternative pre-mRNA splicing. *Methods*, **37**, 341-344.

71. Burge, C.B., Padgett, R.A. and Sharp, P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol Cell*, **2**, 773-785.
72. Ladd, A.N., Charlet, N. and Cooper, T.A. (2001) The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing. *Mol Cell Biol*, **21**, 1285-1296.
73. Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*, **16**, 55-65.
74. Jensen, K.B., Dredge, B.K., Stefani, G., Zhong, R., Buckanovich, R.J., Okano, H.J., Yang, Y.Y. and Darnell, R.B. (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, **25**, 359-371.
75. Rahman, L., Bliskovski, V., Reinhold, W. and Zajac-Kaye, M. (2002) Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles. *Genomics*, **80**, 245-249.
76. Jin, Y., Suzuki, H., Maegawa, S., Endo, H., Sugano, S., Hashimoto, K., Yasuda, K. and Inoue, K. (2003) A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *Embo J*, **22**, 905-912.
77. Fairbrother, W.G., Yeh, R.F., Sharp, P.A. and Burge, C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007-1013.
78. Zhang, X.H. and Chasin, L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*.
79. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M. and Burge, C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831-845.
80. Yeo, G., Holste, D., Kreiman, G. and Burge, C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol*, **5**, R74.
81. Xu, Q., Modrek, B. and Lee, C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res*, **30**, 3754-3766.
82. Black, D.L. (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**, 367-370.
83. Pagani, F. and Baralle, F.E. (2004) Genomic variants in exons and introns: identifying the splicing spoilers. *Nat Rev Genet*, **5**, 389-396.
84. Neves, G., Zucker, J., Daly, M. and Chess, A. (2004) Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet*, **36**, 240-246.
85. Ullrich, B., Ushkaryov, Y.A. and Sudhof, T.C. (1995) Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron*, **14**, 497-507.
86. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O. and Zhang, M.Q. (2000) An alternative-exon database and its statistical analysis. *DNA Cell Biol*, **19**, 739-756.

87. Shah, P., Jensen, L.J., Boue, S, Bork, P. (2005) Extraction of transcript diversity from scientific literature. *PLoS Computational Biology*, **1**, 67-72.
88. Zheng, C.L., Kwon, Y.S., Li, H.R., Zhang, K., Coutinho-Mansfield, G., Yang, C., Nair, T.M., Gribskov, M. and Fu, X.D. (2005) MAASE: an alternative splicing database designed for supporting splicing microarray applications. *Rna*, **11**, 1767-1776.
89. Boguski, M.S. (1995) The turning point in genome research. *Trends Biochem Sci*, **20**, 295-296.
90. Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res*, **29**, 2850-2859.
91. Clark, F. and Thanaraj, T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum Mol Genet*, **11**, 451-464.
92. Taneri, B., Snyder, B., Novoradovsky, A. and Gaasterland, T. (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol*, **5**, R75.
93. Yeo, G., Holste, D., Kreiman, G. and Burge, C.B. (2004) Variation in alternative splicing across human tissues. *Genome Biol*, **5**, R74.
94. Holste, D., Huo, G., Tung, V. and Burge, C.B. (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res*, **34**, D56-62.
95. Kan, Z., States, D. and Gish, W. (2002) Selecting for functional alternative splices in ESTs. *Genome Res*, **12**, 1837-1845.
96. Lee, C. and Wang, Q. (2005) Bioinformatics analysis of alternative splicing. *Brief Bioinform*, **6**, 23-33.
97. Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002) Alternative splicing and genome complexity. *Nat Genet*, **30**, 29-30.
98. Lopez, A.J. (1995) Developmental role of transcription factor isoforms generated by alternative splicing. *Dev Biol*, **172**, 396-411.
99. Zhang, H., Lee, J.Y. and Tian, B. (2005) Biased alternative polyadenylation in human tissues. *Genome Biol*, **6**, R100.
100. Johnson, J.M., Castle, J., Garrett-Engle, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R. and Shoemaker, D.D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, **302**, 2141-2144.
101. Pan, Q., Saltzman, A.L., Kim, Y.K., Misquitta, C., Shai, O., Maquat, L.E., Frey, B.J. and Blencowe, B.J. (2006) Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. *Genes Dev*, **20**, 153-158.
102. Srinivasan, K., Shiue, L., Hayes, J.D., Centers, R., Fitzwater, S., Loewen, R., Edmondson, L.R., Bryant, J., Smith, M., Rommelfanger, C. *et al.* (2005) Detection and measurement of alternative splicing using splicing-sensitive microarrays. *Methods*, **37**, 345-359.

103. Lewis, B.P., Green, R.E. and Brenner, S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A*, **100**, 189-192.
104. Baek, D. and Green, P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci U S A*, **102**, 12813-12818.
105. Resch, A., Xing, Y., Alekseyenko, A., Modrek, B. and Lee, C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res*, **32**, 1261-1269.
106. Xing, Y. and Lee, C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet*, **1**, e34.
107. Yeo, G.W., Van Nostrand, E., Holste, D., Poggio, T. and Burge, C.B. (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A*, **102**, 2850-2855.
108. Nurtdinov, R.N., Artamonova, I.I., Mironov, A.A. and Gelfand, M.S. (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet*, **12**, 1313-1320.
109. Thanaraj, T.A., Clark, F. and Muilu, J. (2003) Conservation of human alternative splice events in mouse. *Nucleic Acids Res*, **31**, 2544-2552.
110. Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, **34**, 177-180.
111. Ast, G. (2004) How did alternative splicing evolve? *Nat Rev Genetics*, **5**, 773-782.
112. Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res*, **13**, 1631-1637.
113. Ohler, U., Shomron, N. and Burge, C.B. (2005) Recognition of unknown conserved alternatively spliced exons. *PLoS Comput Biol*, **1**, 113-122.
114. Philipps, D.L., Park, J.W. and Graveley, B.R. (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. *Rna*, **10**, 1838-1844.
115. Sorek, R., Shemesh, R., Cohen, Y., Basechess, O., Ast, G. and Shamir, R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res*, **14**, 1617-1623.
116. Ratsch, G., Sonnenburg, S. and Scholkopf, B. (2005) RASE: recognition of alternatively spliced exons in *C.elegans*. *Bioinformatics*, **21 Suppl 1**, i369-i377.
117. Boue, S., Letunic, I. and Bork, P. (2003) Alternative splicing and evolution. *Bioessays*, **25**, 1031-1034.
118. Hatton, A.R., Subramaniam, V. and Lopez, A.J. (1998) Generation of alternative Ultrabithorax isoforms and stepwise removal of a large intron by resplicing at exon-exon junctions. *Mol Cell*, **2**, 787-796.
119. Lynch, M. (2002) Intron evolution as a population-genetic process. *Proc Natl Acad Sci U S A*, **99**, 6118-6123.

120. Lynch, M. and Kewalramani, A. (2003) Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol Biol Evol*, **20**, 563-571.
121. Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B. and Galagan, J.E. (2004) Patterns of intron gain and loss in fungi. *PLoS Biol*, **2**, e422.
122. Sorek, R., Ast, G. and Graur, D. (2002) Alu-containing exons are alternatively spliced. *Genome Res*, **12**, 1060-1067.
123. Letunic, I., Copley, R.R. and Bork, P. (2002) Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet*, **11**, 1561-1567.
124. Hieronymus, H. and Silver, P.A. (2004) A systems view of mRNP biology. *Genes Dev*, **18**, 2845-2860.
125. Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) Untranslated regions of mRNAs. *Genome Biol*, **3**, REVIEWS0004.
126. Zhang, H., Hu, J., Recce, M. and Tian, B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res*, **33**, D116-120.
127. Beaudoin, E. and Gautheret, D. (2001) Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Res*, **11**, 1520-1526.
128. Le Texier, V., Riethoven, J.J., Kumanduri, V., Gopalakrishnan, C., Lopez, F., Gautheret, D. and Thanaraj, T.A. (2006) AltTrans: transcript pattern variants annotated for both alternative splicing and alternative polyadenylation. *BMC Bioinformatics*, **7**, 169.
129. Chan, C.S., Elemento, O. and Tavazoie, S. (2005) Revealing Posttranscriptional Regulatory Elements Through Network-Level Conservation. *PLoS Comput Biol*, **1**, e69.
130. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338-345.
131. Foat, B.C., Houshmandi, S.S., Olivas, W.M. and Bussemaker, H.J. (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A*, **102**, 17675-17680.
132. Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A. and Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res*, **29**, 4724-4735.
133. Grillo, G., Licciulli, F., Liuni, S., Sbisà, E. and Pesole, G. (2003) PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res*, **31**, 3608-3612.
134. Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C. and Pesole, G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*, **33**, D141-146.
135. Jacobs, G.H., Stockwell, P.A., Tate, W.P. and Brown, C.M. (2006) Transterm--extended search facilities and improved integration with other databases. *Nucleic Acids Res*, **34**, D37-40.

136. Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M.D. and Hughes, T.R. (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol*, **24**, 5534-5547.
137. Tenenbaum, S.A., Carson, C.C., Lager, P.J. and Keene, J.D. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A*, **97**, 14085-14090.
138. Keene, J.D. and Lager, P.J. (2005) Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res*, **13**, 327-337.
139. Keene, J.D. and Tenenbaum, S.A. (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell*, **9**, 1161-1167.
140. Ambros, V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350-355.
141. Bartel, D.P. and Chen, C.Z. (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet*, **5**, 396-400.
142. Kim, V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, **6**, 376-385.
143. Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol*, **4**, R42.
144. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B. and Bartel, D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, **17**, 991-1008.
145. Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P. and Burge, C.B. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *Rna*, **10**, 1309-1322.
146. Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E. and Zavolan, M. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.
147. Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B. and Bartel, D.P. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, **310**, 1817-1821.
148. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. and Cohen, S.M. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, **123**, 1133-1146.
149. Sood, P., Krek, A., Zavolan, M., Macino, G. and Rajewsky, N. (2006) Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A*.
150. Valencia-Sanchez, M.A., Liu, J., Hannon, G.J. and Parker, R. (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev*, **20**, 515-524.
151. Rajewsky, N. (2006) microRNA target predictions in animals. *Nat Genet*, **38 Suppl 1**, S8-S13.
152. Johnston, R.J., Jr., Chang, S., Etchberger, J.F., Ortiz, C.O. and Hobert, O. (2005) MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision. *Proc Natl Acad Sci U S A*, **102**, 12449-12454.

153. Baskerville, S. and Bartel, D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *Rna*, **11**, 241-247.
154. Thomson, J.M., Parker, J., Perou, C.M. and Hammond, S.M. (2004) A custom microarray platform for analysis of microRNA gene expression. *Nat Methods*, **1**, 47-53.
155. Orphanides, G. and Reinberg, D. (2002) A unified theory of gene expression. *Cell*, **108**, 439-451.
156. Le Hir, H., Izaurralde, E., Maquat, L.E. and Moore, M.J. (2000) The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *Embo J*, **19**, 6860-6869.
157. Gudikote, J.P., Imam, J.S., Garcia, R.F. and Wilkinson, M.F. (2005) RNA splicing promotes translation and RNA surveillance. *Nat Struct Mol Biol*, **12**, 801-809.
158. Moore, M.J. (2005) From birth to death: the complex lives of eukaryotic mRNAs. *Science*, **309**, 1514-1518.
159. Gerber, A.P., Herschlag, D. and Brown, P.O. (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol*, **2**, E79.
160. Vemuri, G.N. and Aristidou, A.A. (2005) Metabolic engineering in the -omics era: elucidating and modulating regulatory networks. *Microbiol Mol Biol Rev*, **69**, 197-216.
161. Puig, S., Askeland, E. and Thiele, D.J. (2005) Coordinated remodeling of cellular metabolism during iron deficiency through targeted mRNA degradation. *Cell*, **120**, 99-110.
162. Blackshear, P.J. (2002) Tristetraprolin and other CCCH tandem zinc-finger proteins in the regulation of mRNA turnover. *Biochem Soc Trans*, **30**, 945-952.
163. Duttagupta, R., Tian, B., Wilusz, C.J., Khounh, D.T., Soteropoulos, P., Ouyang, M., Dougherty, J.P. and Peltz, S.W. (2005) Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. *Mol Cell Biol*, **25**, 5499-5513.
164. Graveley, B.R. (2005) Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell*, **123**, 65-73.



Computational Biology of Post-transcriptional Gene Regulation

Dirk Holste

Institute of Molecular Pathology, Austria
holste@imp.ac.at

Uwe Ohler

Duke University, US
uwe.ohler@duke.edu

ISMB 2006 - Fortaleza, Brazil - August 6-10, 2006

Biology a data and computational intensive discipline



- More than 200 completely **sequenced genomes** show conservation and evolution of genes (*alignments*)
- **Gene expression** arrays provide snapshots of gene activity (*clustering, machine learning*)
- ChIP/CLIP data reveal protein-DNA and protein-RNA **interactions** (*functional elements/motifs*)
- Protein-protein interactions look into **interrelationships** of regulatory complexes (*graphs, belief networks*)



Introduction

2

Biology a data and computational intensive discipline



- Overall size of about 3,000 million base pairs with about one-third genic and two-thirds intergenic DNA sequences (*it's large*)
- Encodes **more than 25,000 genes** that express more than **100,000 proteins** in more than **250 cell types** (*many cell types, many unknown genes, many more proteins*)
- About **5%** of DNA are estimated to be under **natural selection**, with **less than 2%** protein encoding regions

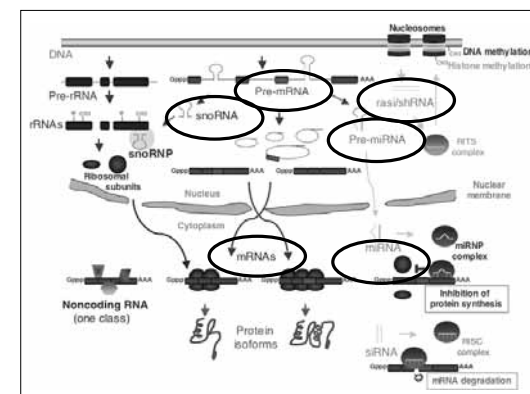


Introduction

3

Post-transcriptional regulation of gene expression

The
expanding
world of
types of
RNA and
non-coding
genes



Introduction

Soares & Valcarcel (2006)

4

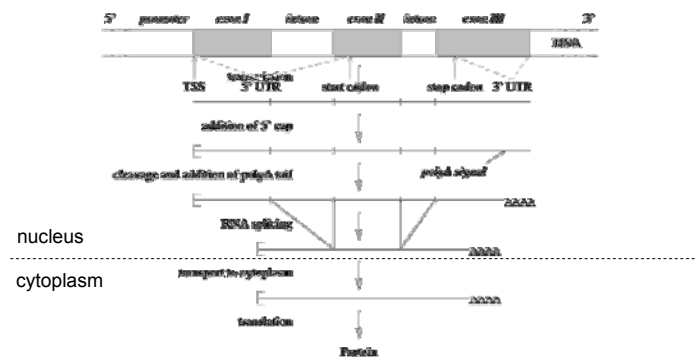
The RNA world

- RNA is thought to be the “original” molecule of life, predating DNA as the so-called “ancient RNA world”
- RNA in modern organisms was thought to be
 - only an intermediary product: the **messenger RNA**
 - a structural component: **rRNA**
 - involved in translation: **tRNA**
 - but not to have an active functional role: the *Central Dogma*
- The “modern RNA world” recognizes that **RNA molecules** have a variety of important **functions**
 - challenging the central dogma and notion of “genes”

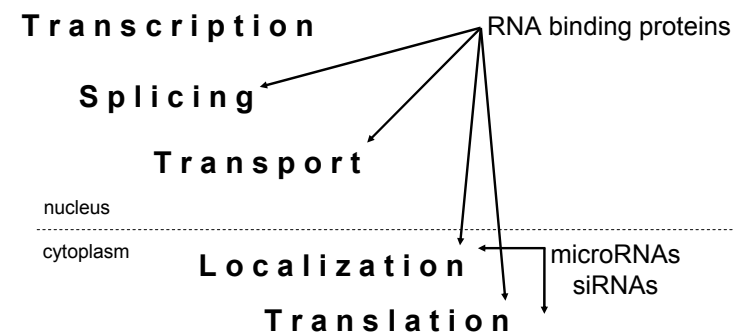
The framework

- What are we interested in?
 - Identification of all **transcriptional variants**: tissue, cell-type specific, developmental stage, disease-specific
 - Understanding the **roles of alternative splicing** (AS), in generating transcript diversity, gene regulation and disease
 - **Simulating RNA splicing/AS** in single organisms
- Different levels of detail and data
 - **Sequences**: genomes, cDNAs, ESTs, SAGE, CAGE, ...
 - **Expression**: splicing-specific microarrays
- Splicing is **just one** post-transcriptional mechanism

Different steps in gene regulation

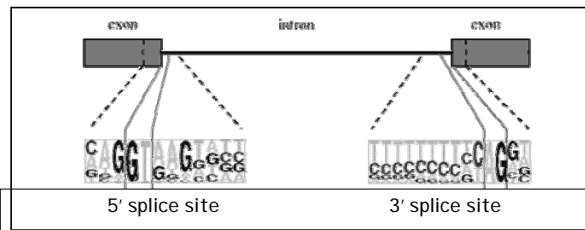


Different steps in gene regulation

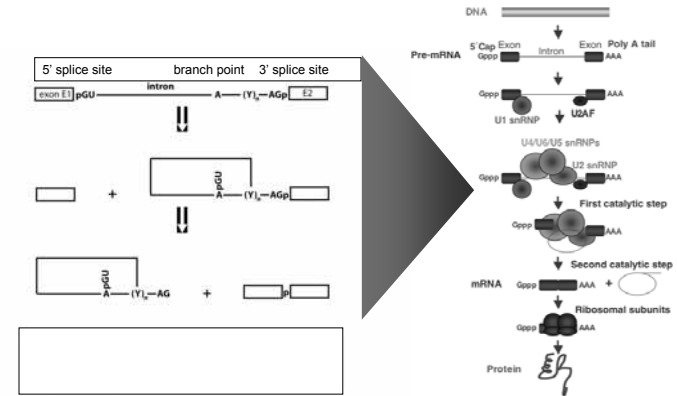


Protein coding genes and splice sites

- Most eukaryotic protein-coding genes have a **split gene structure** of exons and introns
- Conserved **sequence motifs** mark beginning and end of exons

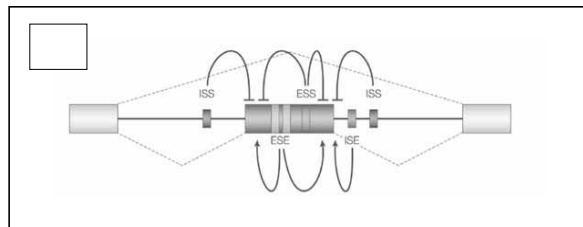


Steps in RNA splicing

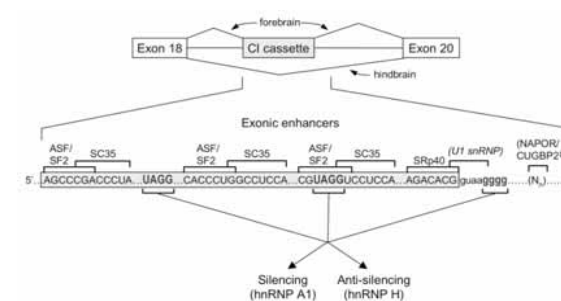


The reality is of course more complicated

- In higher eukaryotic genomes (beyond yeast), splice sites alone often do not contain sufficient information for accurate splicing, compensated for by **splicing regulatory elements** in both exons and introns



Illustrative example: the *NMDA*-type glutamate receptor



RNA-binding proteins attracted to multiple sequence elements:
SR proteins, hnRNPs many (but not all) are conserved

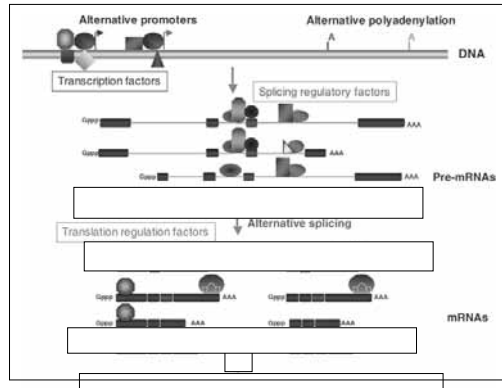
Posttranscriptional regulation of gene expression

Alternative mRNA isoforms from a single gene locus

Alternative promoters/first exons

Alternative polyadenylation/last exons

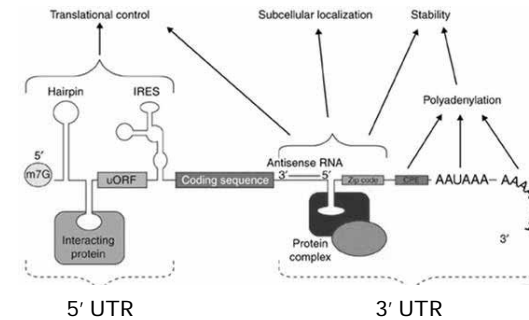
Alternative splicing/internal exons



Introduction

13

The bigger picture

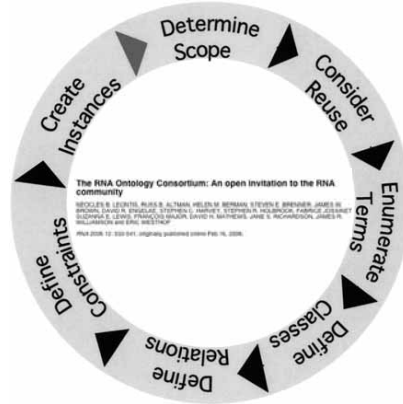


Text...

Introduction

14

Time for a vocabulary



Introduction

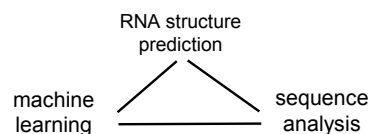
15

Today's menu

- I. Computational methods for RNA structure prediction, sequence alignments, and machine learning
- II. Concepts of experimental methodology at the gene level and genome-wide
- III. RNA splicing and alternative splicing: biology, signals, patterns, spliced-alignments, from transcripts back to the genome, and resources
- IV. Beyond splicing: UTRs and RNA-binding proteins, microRNAs, and networks of regulation
- V. Open issues

16

I. Computational methods



Computational Methodology

• RNA secondary structure prediction

- After all, splicing (and other post-transcriptional mechanisms) takes place on the pre-mRNA level
- Structure plays a role both for *cis*-regulatory elements on RNA sequences, as well as for active RNA *trans*-acting factors
- Today: Focus on individual mRNA (but also some basic ideas for ncRNA prediction)

• Sequence analysis

- Examples for specific alignment algorithms
- Basic principles of classification: HMMs, SVMs

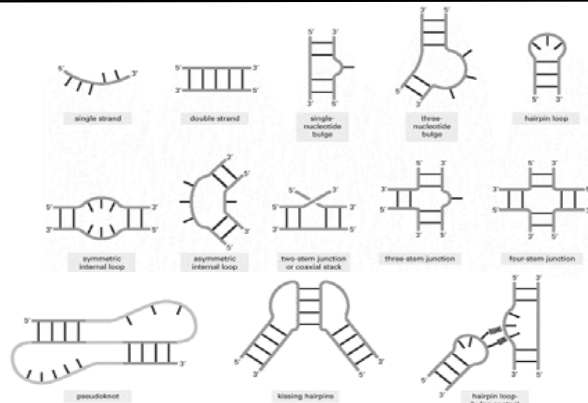
Classes of (small) functional RNAs

- Small nuclear RNAs (**snRNAs**) ↔ Spliceosome
 - Recognize the splice sites / branch point
- Small nucleolar RNAs (**snoRNAs**) ↔ Modification
 - Lead to changes in the sequence of r/sn/m?RNAs
- Micro RNAs (**miRNAs**) ↔ gene regulation
 - Repress translation of target mRNAs
- small interfering RNAs (**siRNAs**) ↔ silencing
 - Degrade mRNAs / induce heterochromatin / RNAi
- Emerging picture:
specific targeting of individual other RNAs

Structure is important

- RNA is a single-stranded molecule, and can fold back onto itself: **secondary structure**
 - G:C > A:U > G:U (Q: consequence of G:U?)
- Apart from “independent” RNA transcripts, secondary structure often plays a **role in cis**
 - Splicing ↔ recognition of splice sites
 - Riboswitches ↔ obstruction of start codon
 - Coding sequence ↔ efficiency of translation
 - RNA editing ↔ change of coding sequence
- Tertiary structure refers to interactions based on the secondary structure, not the 3-D structure...

RNA secondary & tertiary structure

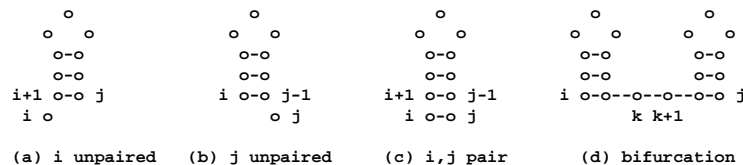


Nussinov algorithm: idea

- Premise: The more nucleotides are paired in a structure, the more stable is the structure
- Simple idea Find the secondary structure with the highest possible number of pairs
- Naïve approach: enumeration (have fun...)
- Instead: use Dynamic Programming
 - Align the sequence to itself
 - Count C:G, A:U, G:U pairs as one, singletons as zero
 - Compute global alignment
 - But...

Nussinov algorithm: operations

- Position (i, j) in the alignment:
 - Best substructure from i to j
 - Fill matrix up to $(1, N)$ and we are done
- At each position in the matrix W , we maximize over four basic cases



- Consequence of (d) on the complexity?
- Runtime: $O(N^3)$

Nussinov algorithm: traceback

- We potentially have *nested* substructures, so we need to use a FIFO stack for traceback
- ```

init: push (1,N)
repeat
 pop (i,j); if (i>=j) continue;
 // done in this substructure
 else if (W(i+1,j) = W(i,j)) push (i+1,j);
 // unpaired
 else if (W(i,j-1) = W(i,j)) push (i,j-1);
 // unpaired
 else if (W(i+1,j-1)+s(i,j) = W(i,j)) push(i+1, j-1);
 // base pair
 else for (k=i+1 to j-1)
 // split substructure
 if W(i,k) + W(k+1,j) = W(i,j)
 push (k+1, j); push (i,k); break;

```



## Folding energy parameters

- Simply counting a match as one and a mismatch as zero is not very close to reality
- Instead, stacking energy parameters have been (and continue to be) estimated
  - Decrease in free energy by stacking one pair of nucleotides on top of the previous pair
    - Means: Markov order 1
  - Increase by various kinds and lengths of unpaired sequences: bulges, internal, terminal/hairpin loops
    - <http://www.bioinfo.rpi.edu/~zukerm/cgi-bin/efiles-3.0.cgi>
  - Incorporate qualitative restrictions, e.g., minimum hairpin loop size



RNA structure

25

## Parameter examples

Stacking energy in stem, X:Y following A:U

|              |       |       |       |
|--------------|-------|-------|-------|
| 5' --> 3' AX |       |       |       |
| 3' <-- 5' UY |       |       |       |
| .            | .     | .     | -0.90 |
| .            | .     | .     | -2.20 |
| .            | -2.10 | .     | -0.60 |
| -1.10        | .     | -1.40 | .     |

Terminal mismatch in hairpin loop, X:Y following A:U

|              |       |       |       |
|--------------|-------|-------|-------|
| 5' --> 3' AX |       |       |       |
| 3' <-- 5' UY |       |       |       |
| -0.30        | -0.50 | -0.30 | -0.30 |
| -0.10        | -0.20 | -1.50 | -0.20 |
| -1.10        | -1.20 | -0.20 | 0.20  |
| -0.30        | -0.30 | -0.60 | -1.10 |



RNA structure

26

## Zuker algorithm: idea

- Start from DP maximal base pairing algorithm, but use free energy parameters instead
  - Becomes quickly complicated:
    - two matrices needed: overall best energy, paired energy at  $i, j$  (similar to insertion/deletion in primary sequence alignment)
    - Tracking of different types of unpaired regions
    - Size restrictions of unpaired/paired regions
  - Extensions allow to find suboptimal structures
    - Current assessment: For only about ~60-70% of structures, the minimal energy structure (i.e., the base pairs) is the correct one *according to the current parameter estimates*
    - Modifications to standard DP algorithm allow to predict suboptimal structures within x% of the optimal one



RNA structure

27

Zuker & Stiegler (1981), Zuker (1989)

## Probabilistic interpretation

- Using the **statistical thermodynamics**, we can relate energies to probabilities:
  - The probability of forming a region with free energy  $\Delta G$  is proportional to  $e^{-\Delta G/kT}$  k... constant; T... temperature
  - The probability of one structure a (e.g., the best one) can be compared to the total probability of all possible structures (the "partition function")
 
$$P(a | s) = \frac{e^{(-\Delta G_a/kT)}}{\sum_b e^{(-\Delta G_b/kT)}}$$
    - Viterbi algorithm vs Forward/Backward in HMMs
    - The probability that  $i, j$  will be paired in *any* possible structure can be assessed



RNA structure

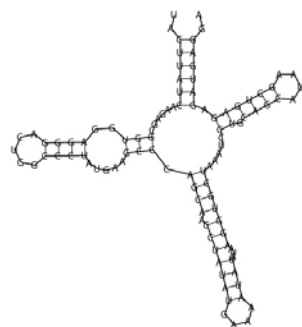
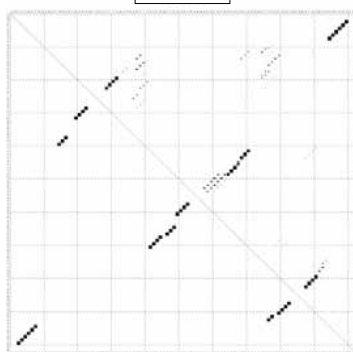
28

McCaskill (1990)



## Probabilistic interpretation: example

- A “probabilistic dot plot” of a clostridium riboswitch RNA



## ncRNA gene finding

- In an ideal world, we would like to **predict RNA genes** independent of their function (just like protein coding genes)
- Bad news first: A good and fancy secondary structure does **not** imply a functional RNA
  - Large enough foldbacks occur frequently by chance
- Remedy I: use **comparative algorithms**
  - Require conservation of the gene across species, either with a formal model or *ad hoc*
- Remedy II: additional **class-specific features**

## General principles for ncRNA finding

- RNAs from the same class show specific *primary* sequence features
- RNAs from the same class show similar *secondary* structure
- Other features: length, position, conservation patterns
  - Common problem: Availability of training/test data
- Combine these in a model, search for highly probable regions in the genome
  - Due to the complexity of structure prediction, this is often done in a sliding window

## Sequence analysis: Motifs

- Cis*-regulatory sequence elements (“**motifs**”) such as splice sites are usually represented by a model
  - Popular: **weight matrix**
  - Calculate the *relative frequency* (occurrences of each symbol at each position, divided by total)
  - Each column: *discrete probability distribution* (entries sum up to 1 and are greater/equal zero)

- Add “pseudo counts”/priors (Bayesian estimation), especially in case of small datasets

| pos | 1   | 2   | 3   | 4   | 5   | 6   |
|-----|-----|-----|-----|-----|-----|-----|
| A   | 0.8 | 0.0 | 0.8 | 1.0 | 0.0 | 0.0 |
| C   | 0.0 | 0.8 | 0.2 | 0.0 | 0.0 | 0.8 |
| G   | 0.0 | 0.2 | 0.0 | 0.0 | 0.2 | 0.2 |
| T   | 0.2 | 0.0 | 0.0 | 0.0 | 0.8 | 0.0 |



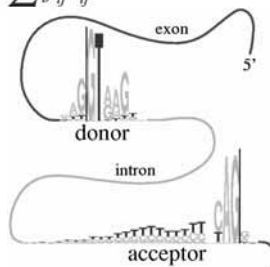
## Sequence logos and information content

- To adjust for background (i.e. nucleotide composition), convert relative frequencies to **log-odds scores**

$$s_{ij} = \log_2(f_{ij}) - \log_2(b_j)$$

- Information content (bit):**  $IC_i = \sum_j f_{ij} s_{ij}$

- Gives impression how often motif can be expected to occur in a sequence
- Sequence logos:** Depict motifs with each position scaled by IC
- Pictogram:** All positions stretched to the same height



## Modeling splice sites

- Alternative models for motifs
  - E.g. splice sites: lots of data available for many species
- Higher order weight matrices
  - Probability of a site:  $P(w_1 \dots w_n) = P(w_1)P(w_2 | w_1) \dots P(w_n | w_1 \dots w_{n-1})$
  - In standard WMM, all positions are assumed independent (i.e., of Markov order 0)
  - Better: Model dinucleotides (Markov order 1) to capture dependencies between adjacent nucleotides
    - 16 instead of 4 parameters per position
- Maximum Dependency Decomposition (MDD)
  - Partition data set and train several models depending on the presence of correlated (non-adjacent) dinucleotides

## Modeling splice sites

- Learn the joint distribution
  - Maximum Entropy
    - Framework that generalizes WMM and higher-order models
    - If nothing else is known, the best distribution is the one leading to the highest entropy (or, adjusted in our case, information content)
    - Distribution has to fulfill constraints of the data, here: low-order (empirical) distributions, i.e. mono/dinucleotide frequencies
  - Alternative: Bayes Networks
    - Any position can depend on all previous ones; model conditional independencies explicitly
  - [Neural Networks]
    - Dependencies can be taken into account either in the input encoding or the connections in the hidden layer (somewhat outdated)
- Experiments confirming splice site strength predictions: Logitlinear models

## A side note on... Terminology

- cDNA:** copy DNA
  - Result of reverse transcribing a DNA sequence from an isolated (partial) mRNA (more precise: transcribed/poly-adenylated RNA)
- EST:** expressed sequence tag
  - Fragment of a mRNA/cDNA
  - Length typically corresponds to one run on a sequencer (~600 nt); often single pass: low quality
- 5'/3' EST:** beginning/end of cDNA
- Libraries:** ESTs are often extracted from specific conditions; development, tissues, cancer



## Sequence analysis: alignment

- The common problem of optimally aligning two sequences also arises in the context of splicing
  - Pre-genomic era: align **EST sequences to each other** to obtain longer/complete gene structures
  - Now: Aligning **cDNAs/ESTs to the genome**
- Specifics:
  - Sequences should ideally be exact matches --- after all, one is a copy of the other
    - Use faster algorithms than simple DP (e.g. BLAT)
    - This is often necessary as genomes can be big...
  - But: **ESTs are spliced**, while **genome is complete**

## EST alignments

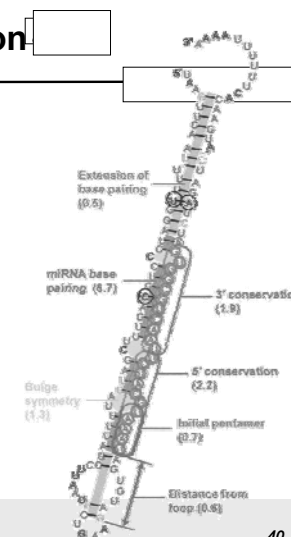
- In particular, special situation for gaps:
  - If arising from intron:
    - Penalty according to intron length
    - Flanking sequence has to match splice sites (maybe on the reverse strand)
- Low quality sequence:
  - "Normal" gaps with standard gap penalties
  - Alignment can take low quality of sequence into account
    - Adjust match/mismatch parameters to common sequence errors or using raw quality scores
- Problems with orthologous/paralogous genes
  - Uniqueness of solution?

## Classification (in one slide)

- Basic principles:
  - We have several *classes* and want to distinguish between them
  - Data shows specific distinct *features*
- Distribution-based**
  - Determine distributions of the feature values for each class from *training data*
    - Supervised:** Class labels known
    - Discrete (e.g. histogram) or continuous (e.g. Gaussian)
  - Determine the probability/likelihood for unseen *test* data (which was not used during training), based on their features
  - Simplest case: features independent, i.e. uncorrelated
    - Separate distribution for each feature
    - Probabilities can be multiplied
    - Class with highest probability wins (naïve Bayes)

## Example: miRNA prediction

- miRNAs are regulatory short RNAs (see later)
- Are excised from so-called precursor foldbacks/hairpins
- Real precursor foldbacks have distinct features
- Classifier can distinguish real miRNAs from ubiquitous foldbacks of similar size





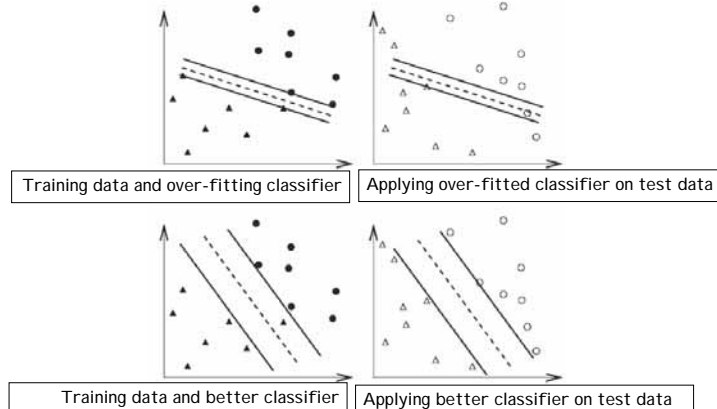
## Distribution-free approaches

- The above approach of classification is based on the specific *distribution* of features
  - Problem: Data does not follow a specific distribution
  - Problem: Not enough data to estimate high-dimensional distribution
- As alternative, learn a classification/decision *function* between two or more classes
  - Samples can be seen as points in a high-dimensional vector space; ideally, classes limited to sub-spaces
  - “Dual” problem: Define separation function by appropriate members of the class, i.e., those closest to the boundary (“support vectors”)

## Support Vector Machines

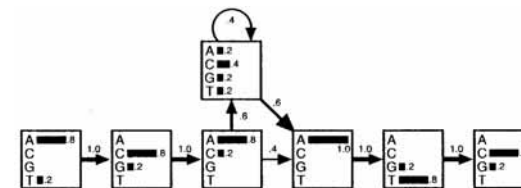
- Simplest case: linear separation of 2 classes
  - Training: determine set of SVs and their weight
  - Similarity: given by dot product  $\langle x, y \rangle$
  - Many functions to separate classes possible, pick the one that *maximizes the margin*
    - Allow for *slack* to include mislabeled/unseparable cases
    - Additional condition to penalize on the number of SVs
- More elaborate classifiers by specific *kernels*
  - Mapping function from “input” to “feature” space: implies linear classification in higher-dimensional space
    - Replace  $\langle x, y \rangle$  by  $K(x, y) = \langle \phi(x), \phi(y) \rangle$
    - Examples:  $d$ -order polynomial; sigmoidal; Gaussian
    - Sequence features: spectrum kernel (substrings w/length  $< k$ )

## SVM illustrated



## Hidden Markov Models

- A probabilistic, i.e., **distribution-based model**, consisting of states connected by transitions
  - States contain feature distributions
  - Here: Discrete distributions on nucleotides (WMM!)
  - Transitions set the probability how often states are used (more than one way to generate one sequence)



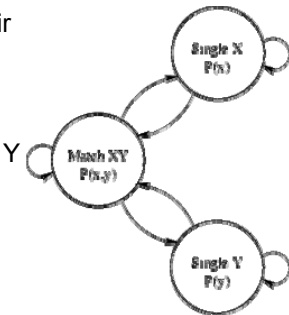


## Applications of HMM

- Several HMMs can be used for *classification*
  - Each HMM represents a different class
  - Evaluate total likelihood that HMM generated test sample by any path through the model
- Alternatively, an HMM can be used to *parse* a sequence
  - I.e., label the sequence with the state IDs which were used on the “best” (i.e. most likely) path: Viterbi-Algorithm
  - E.g. gene finding: Label input genomic sequence with location of splice sites, exons, ...
- Pair HMMs: features are *aligned* nucleotides

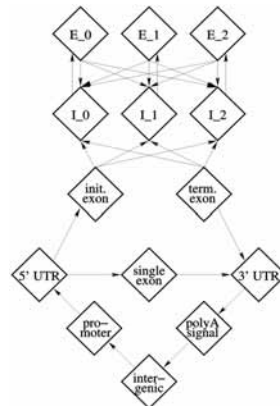
## Pair HMMs

- Emits a *pair* of (aligned) sequences X, Y
- Stereo/*match* state: Output of a pair of aligned nucleotides from a joint distribution (i.e., a distance matrix)
- Mono/*single* states: insertion of unaligned characters in either X or Y
- Transition probabilities specify gap opening/extension
- Standard HMM algorithms can be adapted

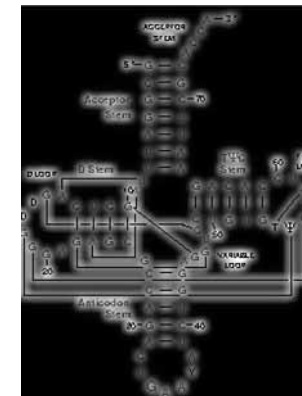


## More complex HMMs

- States represent different functional *regions*
  - Popular application: gene finding
    - Regions, such as coding/intronic
    - Signals, such as splice sites (one state per position), at the transitions
    - Generalized HMM/semi-Markov model

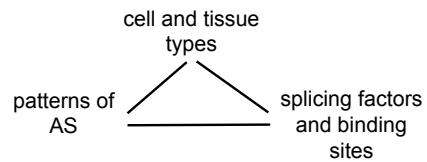


## QUESTIONS FOR THIS SECTION





## II. RNA splicing and alternative splicing



## One gene, multiple messages of mRNA isoforms

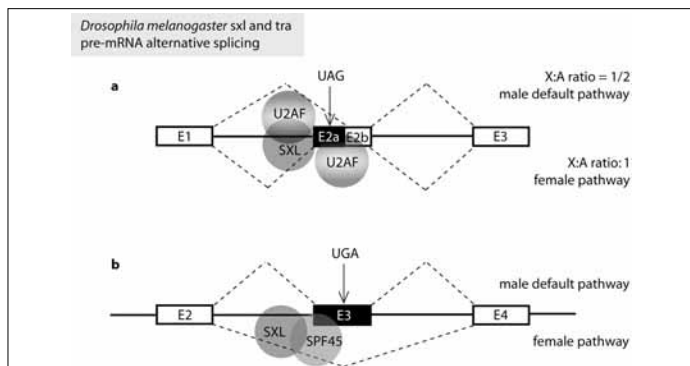
Decoding of primary transcript structure



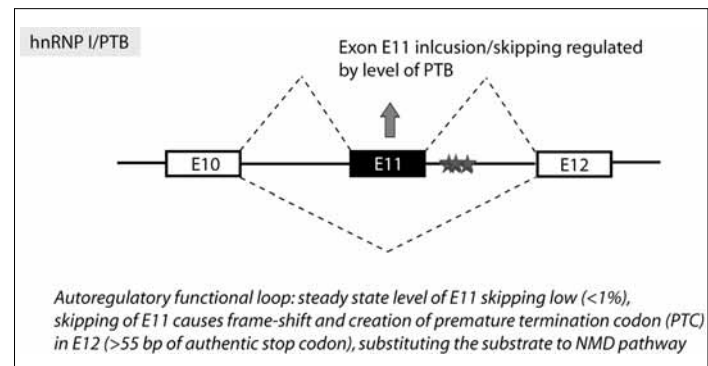
Presence of overlapping codes

- Most eukaryotic protein-coding genes have a split gene structure of exons and introns
- Processing of eukaryotic primary transcripts to generate mRNAs that will direct protein synthesis is often variable, producing multiple alternatively spliced mRNA isoforms

## *Drosophila* Sxl and Tra alternative splicing and sex determination

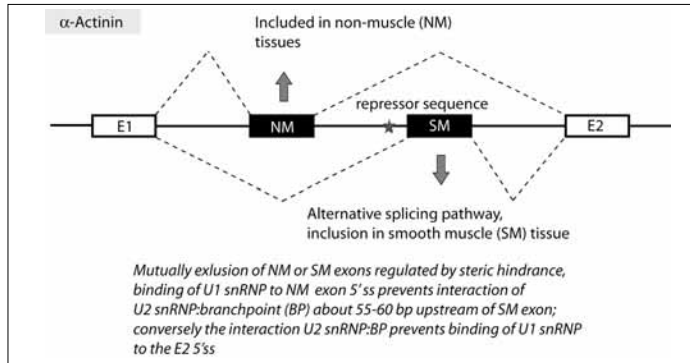


## Splicing regulation of PTB exon skipping by the its own protein

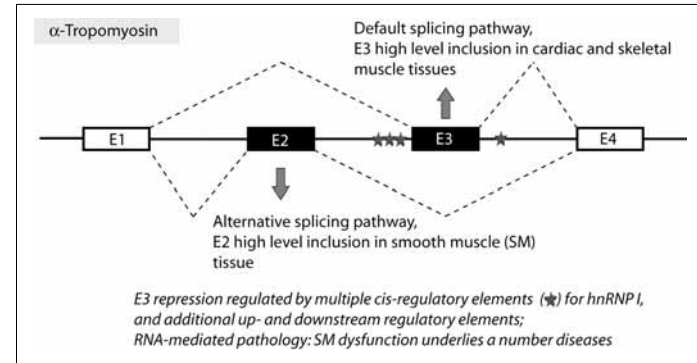




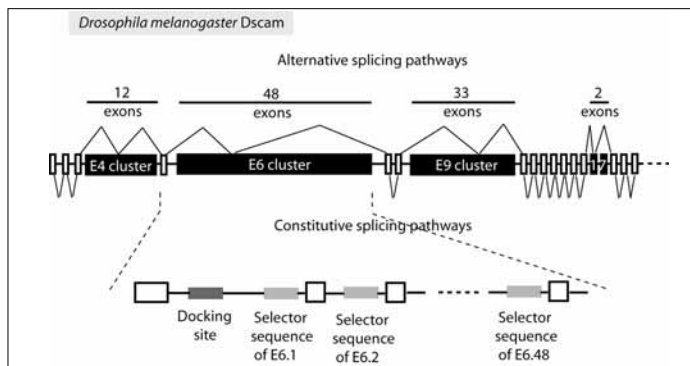
## Tissue-specific splicing switch using steric hindrance



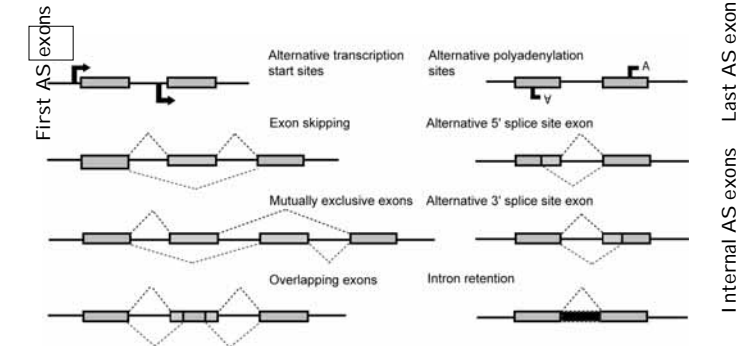
## Tissue-specific exon switch using *trans*-regulatory factors



## Clusters of alternative exons in *Drosophila*'s *Dscam* gene



## A repertoire of alternative splicing patterns





## The machinery at work

### The spliceosome: the most complex macromolecular machine in the cell?

Timothy W. Nilsen

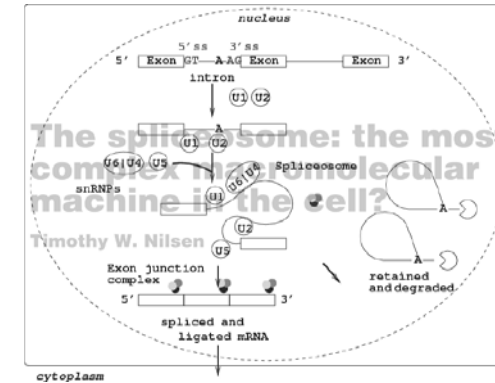
Potentially up ~200 factors involved



Alternative splicing

57

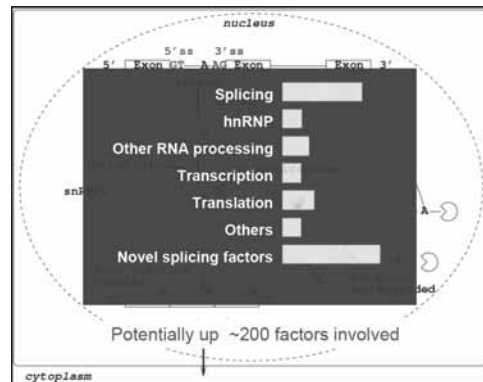
## The machinery at work



Alternative splicing

58

## The machinery at work



Rappalier et al (2002)  
Reed Lab (HMS) Moore  
Lab (Brandeis) Sperling  
Lab (Tel Aviv)

Alternative splicing

59

## Functional roles and consequences of alternative splicing

### RNA splicing and Plasticity of alternative splicing

| Spliceosome function                                                                                                                                                                   | RNA processing                                                                                                                                                                                                                                                                                                                       | Protein functional diversity                                                                                                                                                                         | Splicing errors in human disease                                                                                                                                                                                                                                                                                                                                                                           |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>- Precise recognition of splice sites among many pseudo-sites</li> <li>- Removal of introns</li> <li>- Production of correct message</li> </ul> | <ul style="list-style-type: none"> <li>- Coupling interaction of RNA splicing with gene expression:                             <ul style="list-style-type: none"> <li>&gt; Transcription</li> <li>&gt; Capping and Polyadenylation</li> <li>&gt; mRNA export</li> <li>&gt; Surveillance and mRNA degradation</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>- Cell and tissue-specific mRNA isoforms</li> <li>- Developmental stage regulated isoforms</li> <li>- Inducible control and expression of isoforms</li> </ul> | <ul style="list-style-type: none"> <li>- Inherited mutations and effects of genetic background affecting                             <ul style="list-style-type: none"> <li>&gt; Authentic splice sites</li> <li>&gt; Alternative splice sites</li> <li>&gt; Basal splicing machinery</li> <li>&gt; Trans-acting regulators of AS</li> </ul> </li> <li>- Tumorigenic phenotypes and progression</li> </ul> |



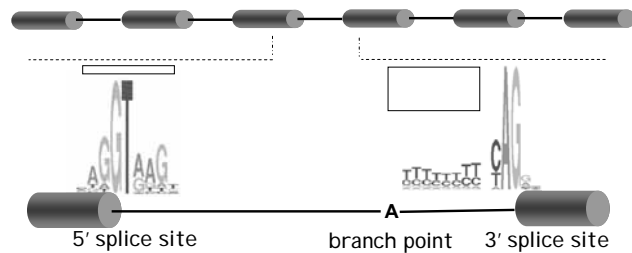
Alternative splicing

adapted from Grabowsky (2004)

60



## Signals: RNA binding sites recognized by the spliceosome



| Signal/Species        | 5'ss | BP | 3'ss  |
|-----------------------|------|----|-------|
| <i>S.cerevisiae</i>   | 11   | 12 | 7 30  |
| <i>C.elegans</i>      | 8    | 5  | 11 24 |
| <i>D.melanogaster</i> | 9    | 5  | 10 24 |
| <i>H.sapiens</i>      | 8    | 5  | 8 21  |

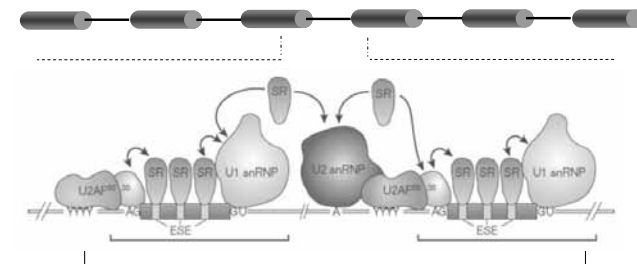
Weaker splice sites, yet more signals enhancing (ESE) or silencing (ESS) exons in higher eukaryotes



Splicing regulatory elements

61

## Signals and *trans*-acting factors



| Signal/Species        | 5'ss | BP | 3'ss  |
|-----------------------|------|----|-------|
| <i>S.cerevisiae</i>   | 11   | 12 | 7 30  |
| <i>C.elegans</i>      | 8    | 5  | 11 24 |
| <i>D.melanogaster</i> | 9    | 5  | 10 24 |
| <i>H.sapiens</i>      | 8    | 5  | 8 21  |

Weaker splice sites, yet more signals enhancing (ESE) or silencing (ESS) exons in higher eukaryotes



Splicing regulatory elements

62

## Splicing regulatory elements

- Splicing accuracy/efficiency is achieved and modulated by "classical" and additional signals in exons and introns
  - Classical: 5'ss MAG/GTRAGT (exonic and intronic nucleotides), branch point [C/T]NCTRA<sub>C</sub> and 3'ss Y(n)NCAG/G (both intronic)
- Additional signals: either **enhancing** or **silencing** a splicing reaction --- in constitutive *and* alternative exons
  - Context-dependent:** exonic splicing enhancer (ESE)/silencer (ESS), intronic splicing enhancer (ISE)/silencer (ISS)
  - Vague elements:** short (often 6 base pairs long), degenerate (lack well-defined consensus), compositional bias (R-rich ESE elements, e.g. GAGAA, AC-rich ESE elements, GT-rich ESS elements), can be overlapping, are under natural selection
  - Multiple occurrences, **positional bias** (closer to splice junctions)
  - Some elements exhibit a splice **site preference**



Splicing regulatory elements

63

## Exonic splicing enhancers (ESEs)

- Splicing signals function similar to transcription factor binding sites (TFBS) and TFBS:TF interactions
- ESEs: Thought to be **utilized as RNA-binding sites for specific arginine-serine-rich (SR) proteins**
  - SR proteins are a class of structurally related and conserved splicing *trans*-acting factors, exhibit several RNA recognition motifs (RRMs)
  - SR proteins bound to ESEs can promote **exon-definition** (using the exon as the unit), by mediating interaction across an exon
  - But... ESE motifs are also found in introns, in intronless genes, and function in roles different from splicing (transport, NMD, ...)
  - Order of several dozen ESEs are proved functional, several hundreds have been ID'ed as candidate elements
  - Can function in complex interaction (e.g., *NMDA*, N1 exons of *c-src* gene) and can have antagonistic context-dependent function



Splicing regulatory elements

64



## Exonic splicing silencers (ESSs)

- Silencers are thought to recruit negative regulators of splicing; those often belong to the family of nuclear ribonucleoproteins (**hnRNP**) --- diverse RNA-binding proteins associating with newly generated pre-mRNA
  - Several dozen hnRNPs known; e.g., hnRNP A1, which is widely expressed in many cell types, and the poly(Y)-tract binding protein (**hnRNP I**), which occurs in tissue-specific form (**nPTB**) in neurons
  - Mechanisms of action are not yet fully understood, but potentially include **competition with ESEs** with overlapping binding sites, multi-merization of hnRNPs
- Decision to make a certain **splice choice** is a **composite** of individual signals/strength of **splice sites** and the **combinatorial effects of enhancing/silencing elements**

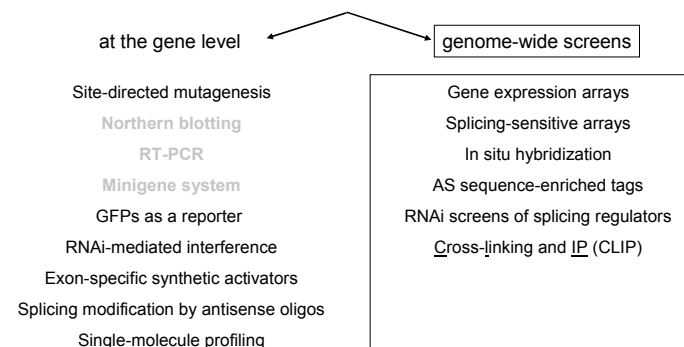


Splicing regulatory elements

65

Fairbrother Yeh et al (2002); Cartegni & Krainer (2002); Black (2003)

## A side note on... Experimental methodology



Experimental methodology

66

## Experimental methodology: Northern blotting

- Determine whether **alternative splice variants of different sizes are present** (and estimate the size of the putative protein), by “running” an RNA blot (a “Northern”)
  - Isolate total or poly(A)+ RNA from cells or whole tissues/organ
  - Run RNA on a denaturing agarose gel, and separate RNAs by their sizes through electrophoresis (smaller RNA will move faster)
  - Transfer separated RNAs to paper/filter, and incubate gel in a solution to hybridize with radioactive-labeled single-stranded probe, complementary to the mRNA of interest
  - After binding (if present) and washing off of any unbound RNA, labeled samples can be detected by autoradiography (exposure to X-ray film)
  - Signal is proportional to amount of target RNA in the population
  - (Low-abundance RNA: use RNase protection assay)

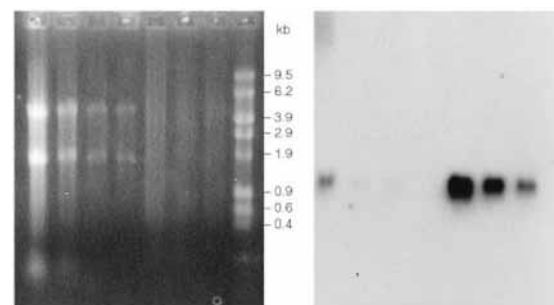


Experimental methodology

67

## Experimental methodology: Northern blotting

Gel analysis of RNA      Northern blot



Experimental methodology

68

www.promega.com/

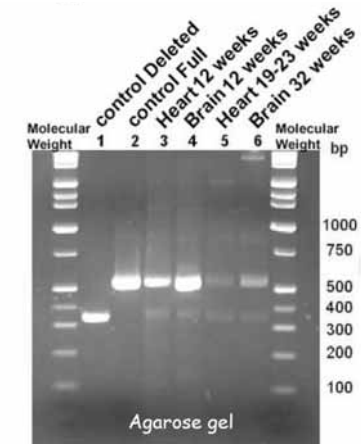


## Experimental methodology: RT-PCR

- Reverse-transcriptase polymerase chain-reaction: Determine whether a gene is transcribing mRNA in a particular cell or tissue-type
  - Convert purified mRNA into complementary DNA** by the enzyme RT and use RNase to destroy the RNA template after first-strand cDNA synthesis
  - Primer hybridization** (to one end of target sequence) by a sequence complementary to mRNA fragment of interest in the cDNA population
  - Extension:** second strand is complemented by the DNA Taq polymerase
  - Denaturation:** target dsDNA is denatured and the second primer (hybridizing to opposite end) is added
  - Repeat cycling** (denaturation-hybridization-extension) to amplify product rapidly

## Experimental methodology: RT-PCR

**DSCAM expression in human fetal heart and brain | RT-PCR**  
 was performed using primers which flank the transmembrane domain of *DSCAM*. The reactions were templated on cDNAs encoding the deleted (1) and full-length (2) transcripts as controls, and on cDNAs reverse-transcribed from human fetal heart (3) and brain (4) at 12 weeks of development, from pooled human fetal heart samples at 19-23 weeks of development (5) and from human fetal brain at 32 weeks of development (6).



## Experimental methodology: Minigenes

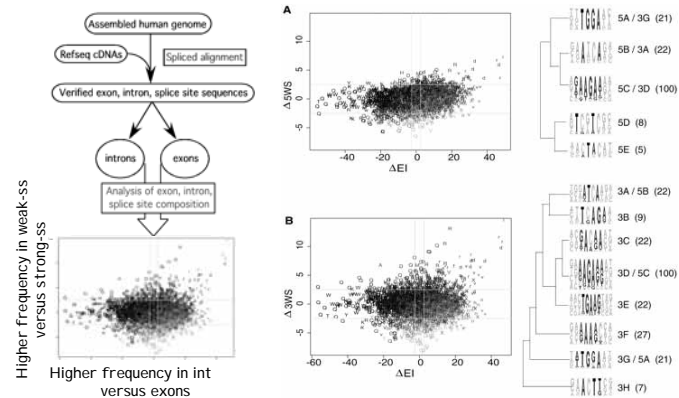
- Transient expression of minigenes is used to study the intrinsic gene **features that direct exon usage: cis-regulatory** elements for constitutive/alternative splicing
  - Contains a parts list of...
    - Transcriptional enhancer/promoter, upstream exon and 5'ss
    - Cloned genomic fragment from gene of interest with alternative exon and flanking regions
    - Downstream exon and 3'ss
    - Cis elements for 3'-end formation
  - Genomic fragment is amplified by PCR from genomic DNA
  - Transfection into cells suitable for the study of the gene (e.g., cell lines HeLa, HEK293, ...)
  - Downstream analysis: RT-PCRs, RNase protection assays, Northern, ...

## Identification of ESE and ESS elements

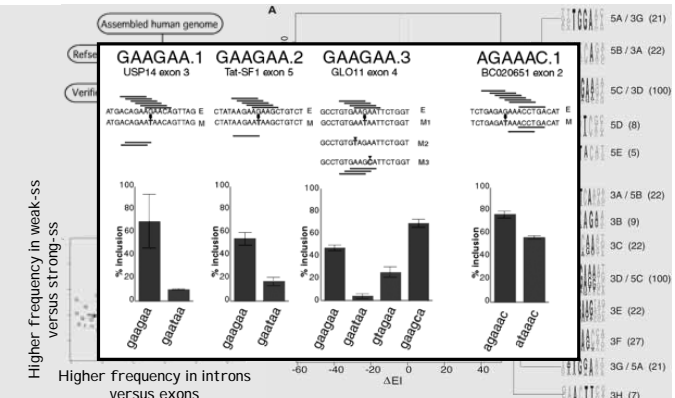
- Computational approaches**
  - Frequency of occurrences of over- or under-represented *k*-mers in exons versus introns (Fox-1 binding site TGCATG in downstream introns of skipped exons for genes expressed in the brain)
  - RESCUE-ESE: *k*-mers integrated into context of weak/strong splice sites (idea: compensatory effects)
  - Putative ESE (PESE) and PESS: non-coding exons in UTR sequences (avoid codon bias)
- Disease-associated** disruption of ESE or ESS elements and site-directed mutagenesis
- Functional selection assay**



## ESE elements: identification of exonic splicing enhancers (RESCUE)



## ESE elements: identification of exonic splicing enhancers (RESCUE)

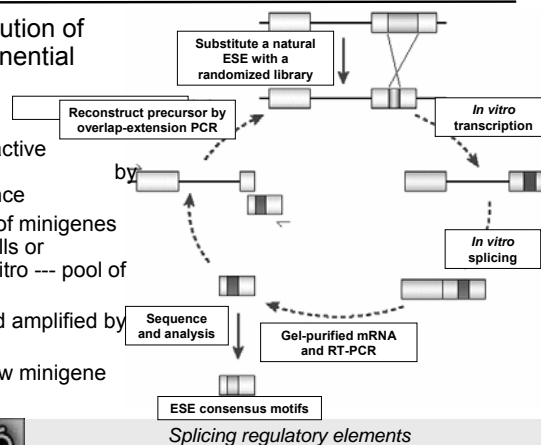


## Identification of ESE elements via selection assays

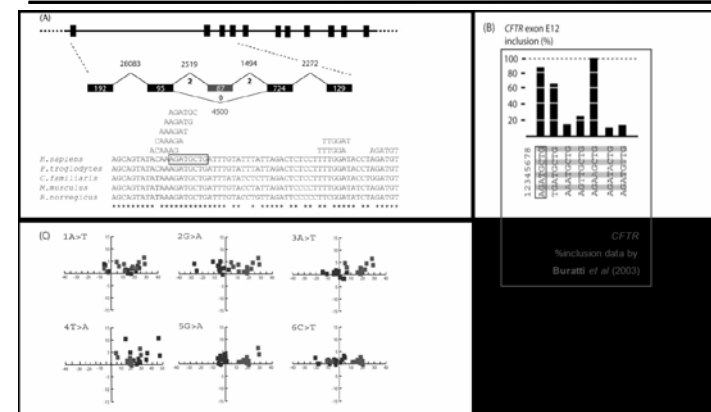
Systematic evolution of ligands by exponential enrichment

{ iterate }

- Minigene with active ESE replaced random sequence
- Transfect pool of minigenes into cultured cells or transcribed in vitro --- pool of mRNAs
- Gel-purified and amplified by RT-PCR
- Reconstruct new minigene templates



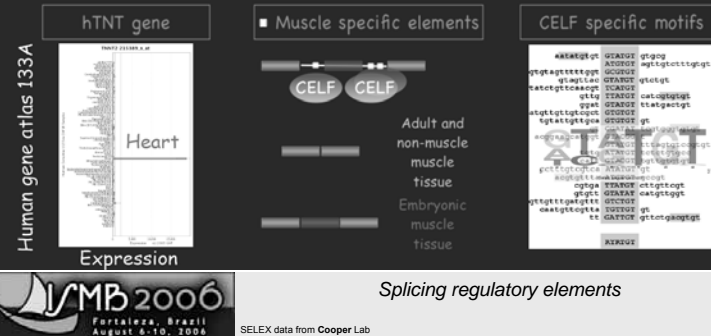
## Example of ESE activity in human CFTR gene



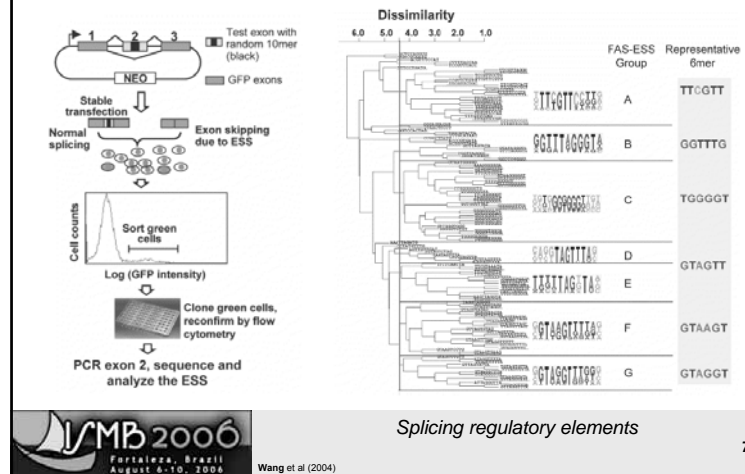


## Example of ISE (intron splicing enhancer) activity in human *cTNT* gene

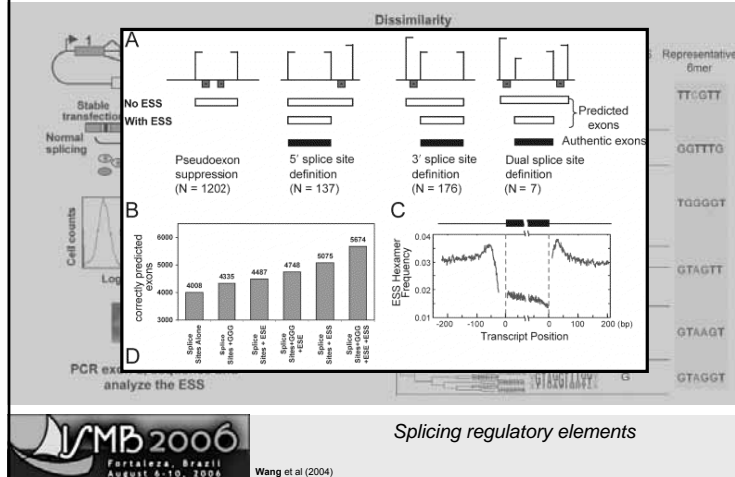
### Trans-acting splicing factors: family of CELF proteins



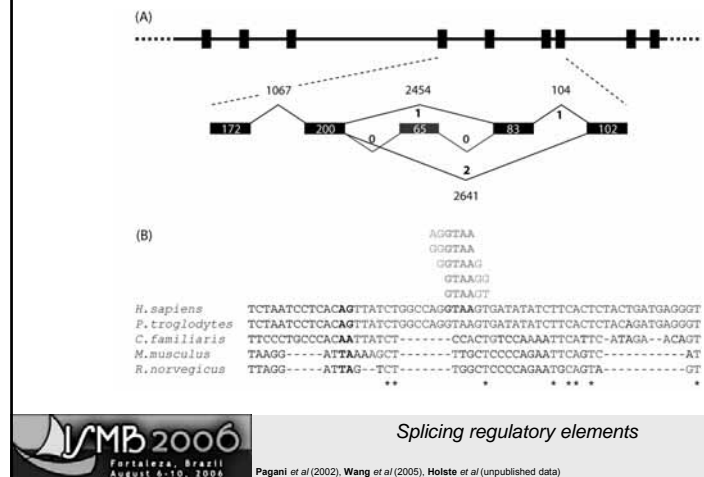
## Identification of exonic splicing silencer (ESS) elements



## Identification of exonic splicing silencer (ESS) elements



## Example of ESS activity in human *ATM* gene

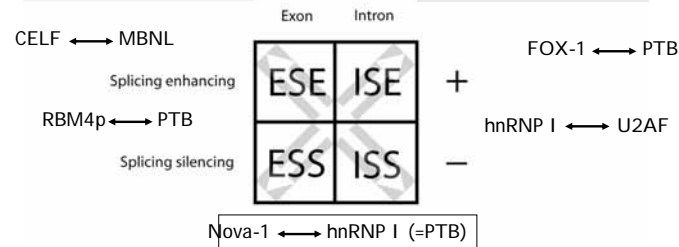




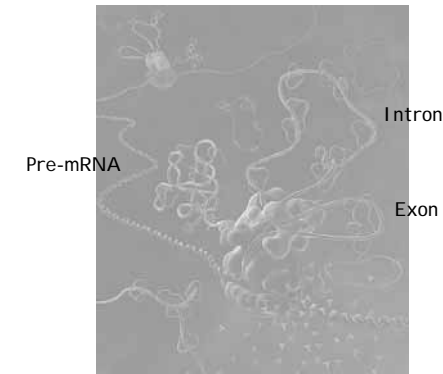
## Antagonism between splicing enhancing and silencing *trans*-acting factors

- Enhancers promote utilization of proximal splice site through exon definition mechanism  
- Purin(R)-rich ESEs recognized by members of SR dipeptide-rich protein family

- Silencers may interfere with access of activators to enhancers via competitive binding  
- ESS elements recognized by heterogeneous nuclear ribonucleoproteins (hnRNPs)



## QUESTIONS FOR THIS SECTION

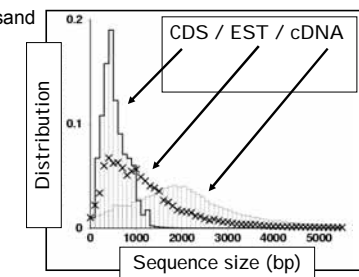


## Identification of alternative mRNA isoforms

- Total **RNA**, or nuclear/cytoplasmic mRNA, derived from diverse conditions (e.g., strains, cell lines, tissues, development, disease), **RT-PCR** and **sequencing**
  - Very precise, but very laborious
- Robotics: large-scale **EST** and **cDNA sequencing** projects measure digital profile of mRNA isoforms
  - Higher throughput, but less specific, less accurate (ESTs)
- Splicing-sensitive **microarrays** measure abundance of single exons or exon-exon junctions present
- Automatic **literature searches** pool experimental outcomes and knowledge into databases

## Transcript data: types and sources

- Large-scale **cDNA** libraries and sequencing as well as individual submissions to **GenBank**, **DBJ**, **EMBL**
  - human & mouse #cDNAs several 100,000
  - *Drosophila* #cDNAs up to ten thousand
  - high-quality data
- Sources: Mammalian Gene Collection (**MGC**), Cancer Genome Anatomy Project (**CGAP**), RIKEN (mouse, rice), RefSeq, consensus CDS (**CCDS**)

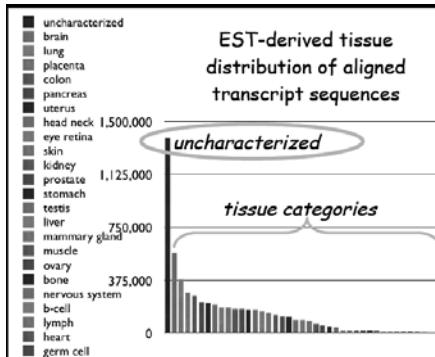




## Transcript data: types and sources

ESTs are very abundant  
(millions for vertebrates)

- **Disadvantage**  
Short, lower in quality (~3% error), 5'-end bias, GC-bias, wt/disease, many tissues, normalized libraries, incompletely spliced transcripts, nuclear pre-mRNAs, reverse strand
- **Advantage**  
short, abundance, wt/disease, many tissues, stages of development



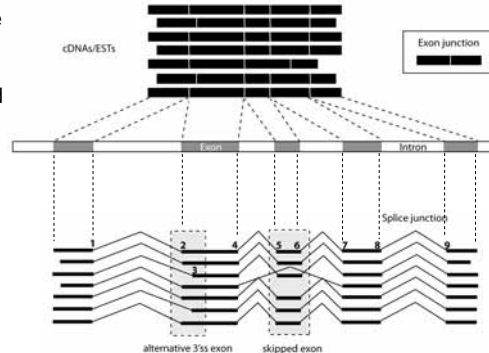
## Pipeline for building an alternative exon database

1. Genome assembly of gene loci UCSC, Ensembl, NCBI 25,000 human genes
2. Mask interspersed repeats in cDNAs Blast, RepeatMasker, Smith-Waterman
3. Determine approximate locations of repeat-masked cDNAs: *uniqueness, aligned length, locus overlap* Blast, Blat Several hundred repeats
4. Splice-align original cDNAs against candidate gene loci: *percent identity and cDNA coverage* Blat, Spidey, Sim4, Exalign, Gmap, ASPIC, SPA, mRNAsGen ~200,000 cDNAs
5. Determine significant matches of similarity between ESTs and repeat-masked version of successfully splice-aligned cDNAs Blast, Blat
6. Splice-align candidate ESTs against gene loci covered by cDNA alignments: *percent identity, EST coverage, and library information* Spidey, Sim4, Exalign, Gmap, est2genome, ASPIC, SPA ~700,000 ESTs
7. Classify genes into loci with single and multiple transcripts ~15,000 genes
8. Determine types and frequencies of occurrence of alternative splicing events on exon level: *high stringency filter (problem: many false positives)*
9. Database for storage, maintenance, querying, retrieval of AS data PostgreSQL

## Gene structure assembly

Given an incomplete and complement transcript sequence (cDNA/ mRNA), find the complete genomic sequence (primary transcript structure) and determine the gaps (introns)

- Align sequences with gaps and determine 'best' match
- Gaps start/end with 5'ss and 3'ss signals

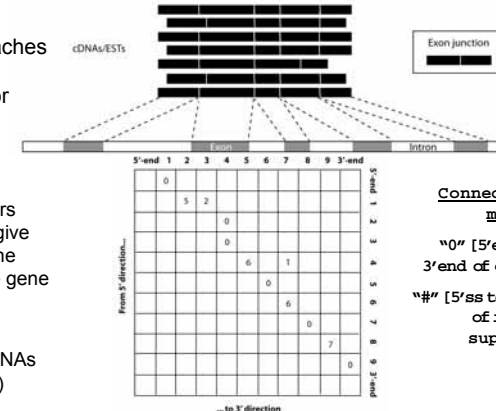


## Gene structure assembly II

Align sequences by

1. heuristic approaches
2. dynamic programming, or
3. probabilistic (Bayesian) approaches

- Different parameters and/or algorithms give rise to different gene structures (just like gene finding)
- Algorithms can be tailored to align cDNAs and ESTs (quality!)



Connectivity matrix  
"0" [5'end to 3'end of exon]  
"#" [5'ss to 3'ss of intron support]



## Algorithms for gene structure assembly

| Algorithm              | Web-based reference                                                                                                 |
|------------------------|---------------------------------------------------------------------------------------------------------------------|
| ASPIC                  | <a href="http://aspic.algo.disco.unimib.it/aspic-devel/">http://aspic.algo.disco.unimib.it/aspic-devel/</a>         |
| BLAST                  | <a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>                                               |
| BLAT                   | <a href="http://genome.ucsc.edu/~kent">http://genome.ucsc.edu/~kent</a>                                             |
| DDS/GAP2               | <a href="http://www.tigr.org/software/alignment.shtml">http://www.tigr.org/software/alignment.shtml</a>             |
| Ensembl genome Browser | <a href="http://www.ensembl.org">http://www.ensembl.org</a>                                                         |
| EST_GENOME             |                                                                                                                     |
| EXALIGN                | <a href="http://blast.wustl.edu/exalign">http://blast.wustl.edu/exalign</a>                                         |
| GMAP                   | <a href="http://www.gene.com/share/gmap">http://www.gene.com/share/gmap</a>                                         |
| MGAlign                | <a href="http://origin.bic.nus.edu.sg/mgalign/">http://origin.bic.nus.edu.sg/mgalign/</a>                           |
| mRNAsGen               | <a href="http://genes.mit.edu/genoa">http://genes.mit.edu/genoa</a>                                                 |
| Sim4                   | <a href="http://globin.cse.psu.edu">http://globin.cse.psu.edu</a>                                                   |
| Spidey                 | <a href="http://www.ncbi.nlm.nih.gov/spidey">http://www.ncbi.nlm.nih.gov/spidey</a>                                 |
| GeneSeqer              | <a href="http://globin.cse.psu.edu/">http://globin.cse.psu.edu/</a>                                                 |
| TAP                    | <a href="http://sapiens.wustl.edu/~zkan/TIP/">http://sapiens.wustl.edu/~zkan/TIP/</a>                               |
| UCSC Genome Browser    | <a href="http://genome.ucsc.edu">http://genome.ucsc.edu</a>                                                         |
| WABA                   | <a href="http://www.cse.ucsc.edu/~kent/intronator">http://www.cse.ucsc.edu/~kent/intronator</a>                     |
| SPA                    | <a href="http://www.biozentrum.unibas.ch/personal/nimwegen/">http://www.biozentrum.unibas.ch/personal/nimwegen/</a> |



Genomics of splicing

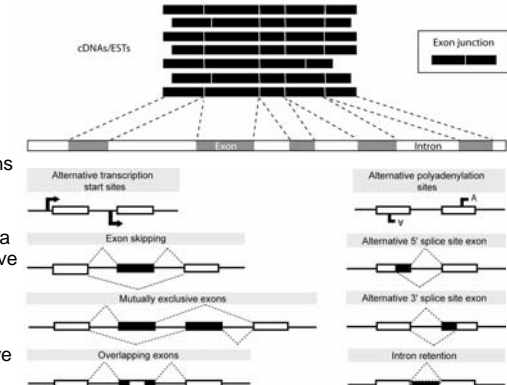
89

## Transcript-derived categories of alternative splice patterns

Four basic AS patters

1. Exon **skipping**
2. Alternative **5'ss** exon
3. Alternative **3'ss** exon
4. Intron **retention**

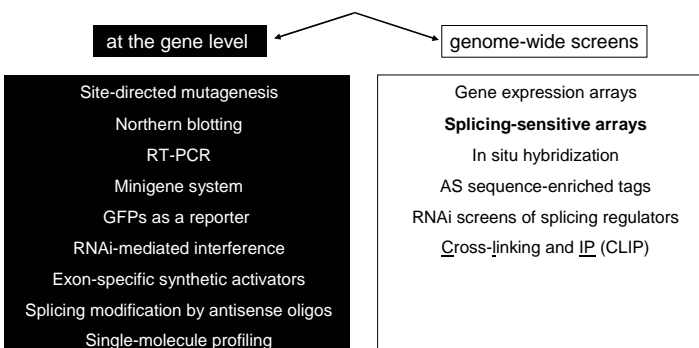
- Mutually exclusive exons are a special type of exon skipping
- Overlapping exons are a special type of alternative 5'ss and 3'ss exons
- Different start/ polyadenylation sites correspond to alternative first/last exons



Genomics of splicing

90

## A side note on... Experimental methodology



Experimental methodology

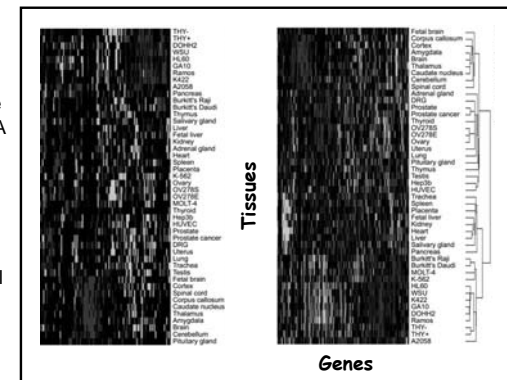
91

## Experimental methodology: Microarrays

### Gene expression and splicing- sensitive arrays

Microarrays measure overall level of mRNA from a gene, either "absolute" or "competitive" (with different fluorescent dyes)

Tiling arrays cover all non-random portions of genomes, in addition to genes



Experimental methodology

Gene Atlas (2004)

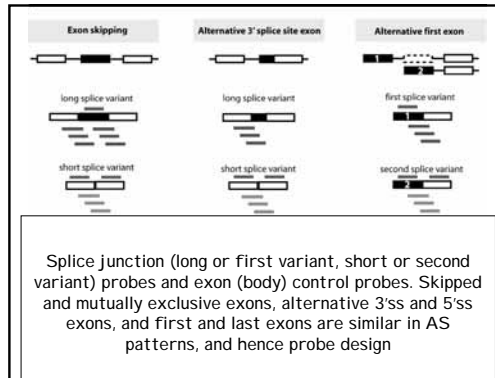
92



## Experimental methodology: Microarrays

### Gene expression and splicing-sensitive arrays

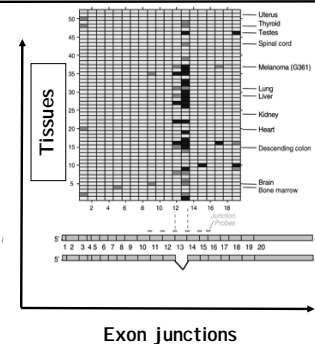
Microarrays can be adopted to alternative splicing, requiring a reference exon-intron structure from prior knowledge of spliced-alignments and adopted probe design



## Experimental methodology: Microarrays

### To address...

1. *Sn and Sp of measured splice types (tissues/cell lines, duplicated probes, prior AS information, gene expression)*
2. *Probe design: oligo-size, melting temperature, secondary structure, low information content uniqueness/cross-hybridization, complex AS forms*
3. *Normalization (noise reduction, different expression in different tissues)*
4. *Rule- or score-based decision of splice choices*
5. *Different flavors of arrays: oligo-based, RASL, ...*



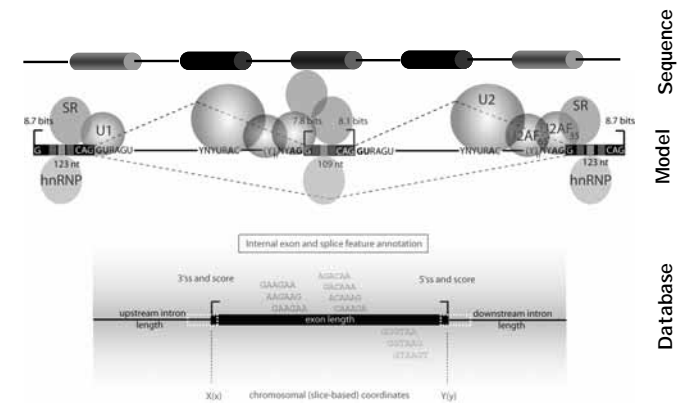
AS prediction scores of splice junction probes (yellow to blue: increase in deficit of observed to expected intensity)

## Experimental methodology: Microarrays

### What were chips used for? Examples...

- **Rat genes** (~1,600) interrogates across 10 normal tissue types (standard Affymetrix array)
  - Yeast mutants:
    - Genome-wide effects on splicing caused by loss of different splicing factors (~20) to study functional relationships between factors
- **Human transcriptome:**
  - About 10,000 multi-exon genes across ~50 diverse tissue samples (some including two stages of development), and cell lines to indicative of more than 70% AS genes
  - NCI 60 cancer cell lines
- **Mouse transcriptome:**
  - About 3,100 AS events (skipping) across ~2,600 genes and a panel of about a dozen tissue types

## Conceptualizing sequences to data models





## The “best of” of all databases

- Simple and clear **query forms**, accepting gene names and synonyms or gene identifiers
- **Summary** page of gene and identified splicing patterns
- **Characterization of alternative splice forms** and prototypic transcript sequences used to identify the presented patterns, in textual and graphical representation
- Additional information: **exon** sequence, **splice sites**, upstream and downstream **intron** sequences, frequency of occurrence (**inclusion and exclusion**), allele or strain, genotype data, protein **domains**, promoter, **RNA-binding sites**, **polyadenylation** signal, reading frame, stop codon and any premature termination codons (**PTCs**), embeddings in Ensembl, UCSC, NCBI, EST-derived tissue information, expression and splicing array data, wt/ disease association, sequence-conservation, **splicing conservation**, phenotypic data
- Linked tools for bioinformatics (CLUSTALW, MEME, Gibbs, clustering,...)



Databases

97

## Developed and implemented databases for alternative exons

Database/Dataset Web-based reference

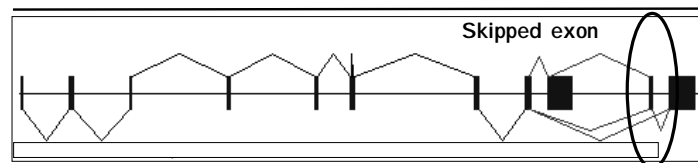
|                                          |                                                                                                                                   |
|------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Collection of EST-inferred splice sites  | <a href="http://www.ebi.ac.uk/~thararaj/splice.html">http://www.ebi.ac.uk/~thararaj/splice.html</a>                               |
| Collection of literature-based AS events | <a href="http://cpgsma.cshl.org/new_alt_exon_db2/">http://cpgsma.cshl.org/new_alt_exon_db2/</a>                                   |
| ASDB                                     | <a href="http://cbcg.nersc.gov/asdb/">http://cbcg.nersc.gov/asdb/</a>                                                             |
| ISIS (intron database)                   | <a href="http://isis.biol.uq.edu.au/isa_splicers.html">http://isis.biol.uq.edu.au/isa_splicers.html</a>                           |
| ASManDB                                  | <a href="http://166.111.30.65/ASManDB">http://166.111.30.65/ASManDB</a>                                                           |
| SpliceDB (U2/U12 splice sites)           | <a href="http://www.softberry.com">http://www.softberry.com</a>                                                                   |
| SpliceNest                               | <a href="http://splicecest.molgen.mpg.de">http://splicecest.molgen.mpg.de</a>                                                     |
| PAL-Sdb                                  | <a href="http://palsdb.ym.edu.tw">http://palsdb.ym.edu.tw</a>                                                                     |
| AltExtron                                | <a href="http://www.ebi.ac.uk/asd/altextron">http://www.ebi.ac.uk/asd/altextron</a>                                               |
| DEDB (AS patterns in fly)                | <a href="http://pubmed.ncbi.nlm.nih.gov/pubmed/1406474/">http://pubmed.ncbi.nlm.nih.gov/pubmed/1406474/</a>                       |
| AS graphs                                | <a href="http://www.cse.ucsd.edu/groups/bioinformatics/ESTs">http://www.cse.ucsd.edu/groups/bioinformatics/ESTs</a>               |
| ASAP                                     | <a href="http://www.bioinformatics.ucla.edu/ASAP">http://www.bioinformatics.ucla.edu/ASAP</a>                                     |
| ProSplicer                               | <a href="http://bioinfo.csie.ncu.edu.tw/ProSplicer">http://bioinfo.csie.ncu.edu.tw/ProSplicer</a>                                 |
| AS patterns in plants                    | <a href="http://plants.genomics.org.cn">http://plants.genomics.org.cn</a>                                                         |
| ASD                                      | <a href="http://www.ebi.ac.uk/asd">http://www.ebi.ac.uk/asd</a>                                                                   |
| Splicing conserved AS patterns           | <a href="http://www.ebi.ac.uk/asd/signet/psdb/frameshift/psdb.html">http://www.ebi.ac.uk/asd/signet/psdb/frameshift/psdb.html</a> |
| AS patterns across different tissues     | <a href="http://genes.mit.edu/genoa">http://genes.mit.edu/genoa</a>                                                               |
| ASG, browsing splice patterns            | <a href="http://atlasgen.ncsu.edu/asg">http://atlasgen.ncsu.edu/asg</a>                                                           |
| SpliceInfo                               | <a href="http://spliceinfo.msh.edu.tw">http://spliceinfo.msh.edu.tw</a>                                                           |
| ECgene                                   | <a href="http://genome.ewha.ac.kr/ECgene">http://genome.ewha.ac.kr/ECgene</a>                                                     |
| STACdb                                   | <a href="http://www.sanra.ac.za/STACdb.html">http://www.sanra.ac.za/STACdb.html</a>                                               |
| EASED                                    | <a href="http://easedb.bioinfo.mdc-berlin.de">http://easedb.bioinfo.mdc-berlin.de</a>                                             |
| Collection of alternative poly-A sites   | <a href="http://physics.nyu.edu/~y272/altA">http://physics.nyu.edu/~y272/altA</a>                                                 |
| LSAT (literature-based)                  | <a href="http://www.bork.embl.de/LSAT">http://www.bork.embl.de/LSAT</a>                                                           |
| MAASE (literature-based)                 | <a href="http://maase.genomics.purdue.edu">http://maase.genomics.purdue.edu</a>                                                   |
| HOLLYWOOD                                | <a href="http://hollywood.mit.edu">http://hollywood.mit.edu</a>                                                                   |



Databases

98

<http://www.bioinformatics.ucla.edu/ASAP>



### Fragile X mental retardation-related protein 1 (FRX1)

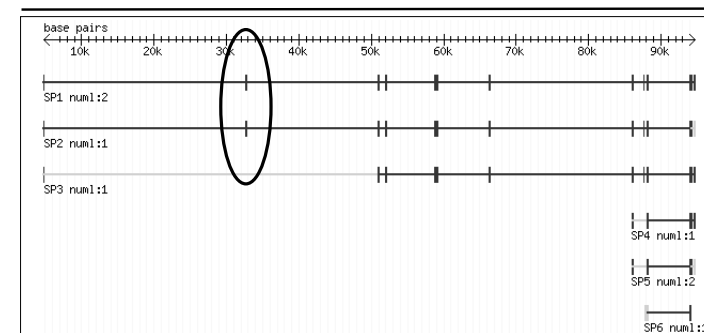
FRX1 directly associated with FMR1, an RNA-binding protein; FMR1 protein and the FRX1 (and FRX2) form a family with functional similarities, such as RNA binding, polyribosomal association and nucleocytoplasmic shuttling; it is widely expressed in mammals and its expression pattern is complex with several mRNA variants and protein isoforms; it's an autoimmune antigen in patients



Databases

99

<http://www.ebi.ac.uk/asd>



### Poly pyrimidine tract binding protein (PTB)

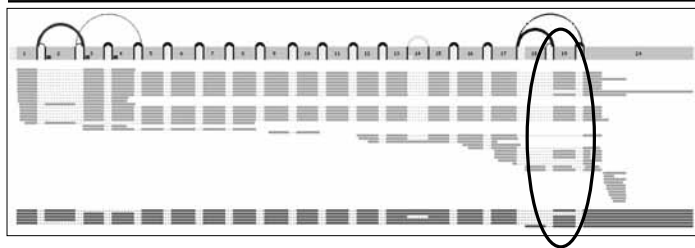


Databases

100



<http://stagen.ncsu.edu/asg>



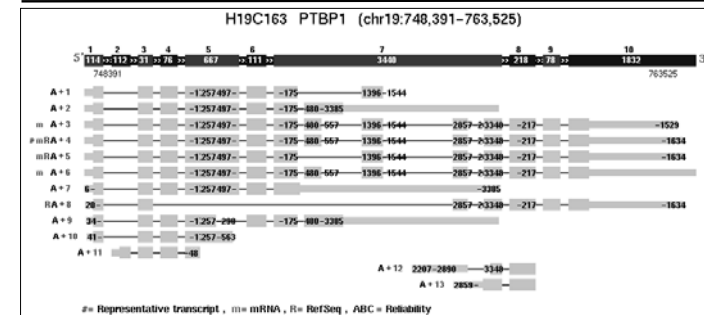
Fragile X mental retardation-related protein 1 (FRX1)



Databases

101

<http://genome.ewha.ac.kr/ECgene>



Poly pyrimidine tract binding protein (PTB)



Databases

102

<http://bioinfo.csie.ncu.edu.tw/ProSplicer>



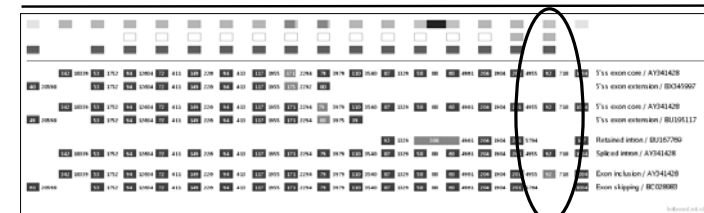
Poly pyrimidine tract binding protein (PTB)



Databases

103

<http://hollywood.mit.edu>



Fragile X mental retardation-related protein 1 (FRX1)



Databases

104



## A constant challenge for bioinformatics development



"The design of effective databases to support experimental and computational investigations of AS is a significant challenge. The effort to integrate accurate exon and splice site annotation with current knowledge about splicing regulatory elements and predicted AS events, and to link information about the splicing of orthologous genes in different species, requires ongoing developments..."

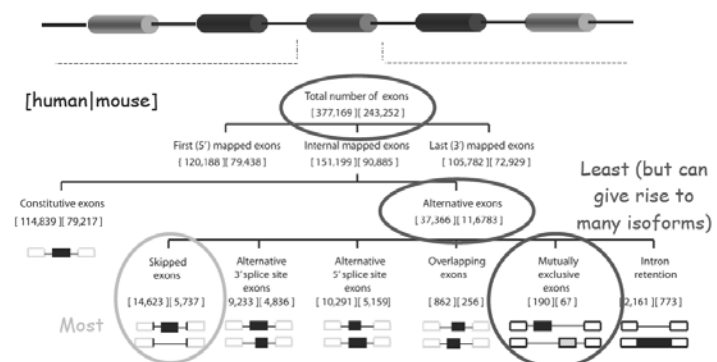
## QUESTIONS FOR THIS SECTION



## Using AS databases for the analysis and understanding of splicing

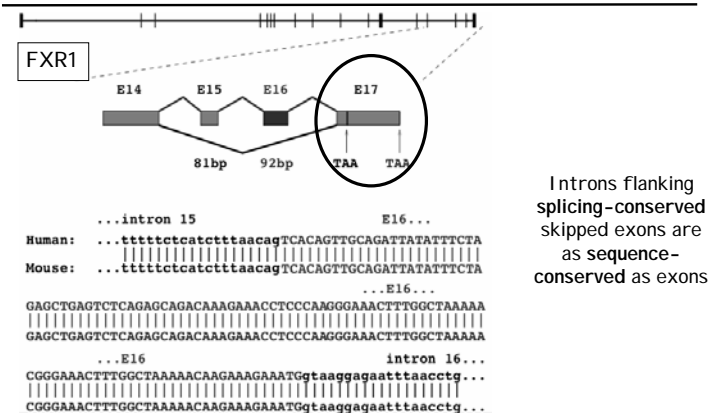
- **Descriptive feature statistics**
  - What AS types are produced by the splicing machinery and which types have the highest/lowest proportions? Do different tissues differ in their usage of AS types and proportions? Which tissues/stages are most distinct from others in the spectrum of AS isoforms they express? To what extent correlate splicing factors with the observed AS patterns? Can we detect candidate binding sites for tissue-specific splicing factors?
- What sequence **features** are different in **constitutive and alternative exons**?
- What AS types are **splicing-conserved** (e.g. between human and mouse)?
- Can we **predict AS types from sequence** without transcript information? (similar to *ab initio* gene prediction)
- What do we learn about the **evolution of AS**?
- Can we **simulate splicing** (the spliceosome) *in silico*?

## Types and abundances of mammalian AS patterns





## FXR1 gene splicing pattern of exon E16



Introns flanking  
 splicing-conserved  
 skipped exons are  
 as sequence-  
 conserved as exons



Computational analysis of alternative splicing

109

Yeo et al (2005)

## Tissue-enriched alternative splicing events

Mammalian AS events are not evenly distributed but enriched in certain tissues, associated with conservation/creation/loss of alternative exons

- Digital EST counts: (human, mouse)
  - Enriched number of AS forms demonstrated in the **human brain**, also in the **eye/retina, muscle, skin, testis, and liver** (e.g., transcription factors exhibit different isoforms across tissues)
  - 10-30% of AS genes undergo tissue-specific splicing (Xu et al)
- Splicing-sensitive microarrays: (mouse)
  - Tissue-specificity of genes (~3,100) regulated by AS occurs to a large extent independent of the tissue-specificity of transcription, and vice versa
  - %Inclusion levels depend on splicing- and exon-conservation (genome-specific exhibit generally weak %inclusion levels)
  - About 240/2,100 (with 2,100/3,100 inclusions in more than two tissues) “tissue-switched” AS exons were indicative of increased preservation of the reading-frame (defined by marked changes between “major” (>66%) and “minor” form (<34%) of %inclusion across tissues)



Computational analysis of alternative splicing

110

Xu et al (2002), Taneri et al (2004), Yeo et al (2005), Ying & Lee (2005)

## Different types and abundances of AS patterns in different tissues

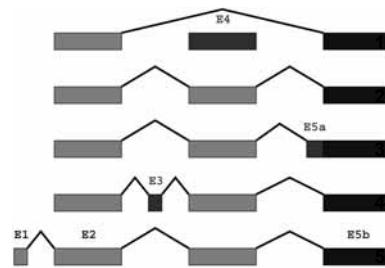
Measuring differences between EST (short, incomplete)

$d(i,j)$  = #splice junctions that differ between transcripts  $i,j$

$t(i,j)$  = total #splice junctions in  $(i,j)$

$SJD(i,j) = d(i,j)/t(i,j)$

| $i$ | $j$ | $SJD(i,j)$ |
|-----|-----|------------|
| 1   | 2   | 3/3 = 1    |
| 2   | 3   | 2/4 = 0.5  |
| 2   | 4   | 3/5 = 0.6  |
| 1   | 4   | 4/4 = 1    |
| 2   | 5   | 0/4 = 0    |



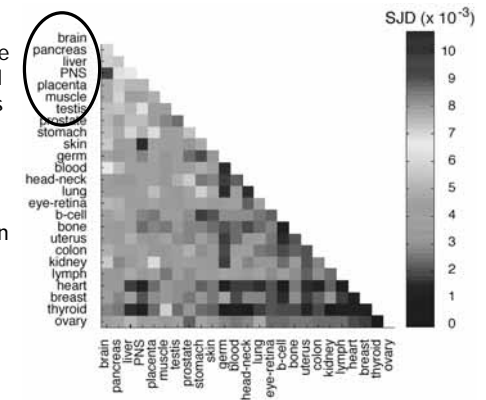
Computational analysis of alternative splicing

111

Yeo et al (2005)

## Different types and abundances of AS patterns in different tissues

Compute the difference between ESTs derived from different tissues (SJD ratio), across all genes with sufficient pair-wise EST tissue coverage (blue to red: increase in different usage of splice choices)



Computational analysis of alternative splicing

112

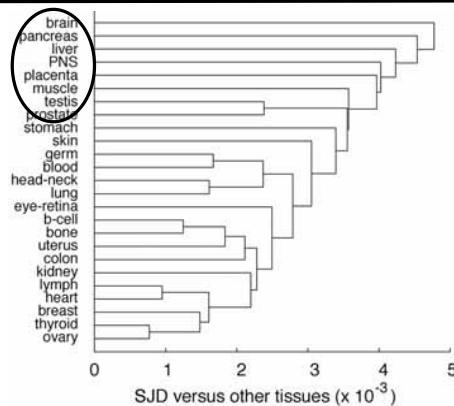
Yeo et al (2005)



## Different types and abundances of AS patterns in different tissues

Tree relationship based upon the SJD ratio across all tissues

The human brain/nervous system, pancreas, placenta, liver, and muscle are singled-out groups, several cluster resemble organismal relationships

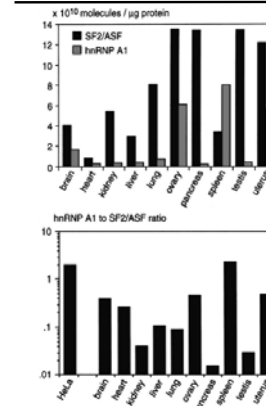


Computational analysis of alternative splicing

113

Yeo et al (2005)

## Gene expression of *trans*-acting splicing factors across tissues



Explore differences in splicing factor expression:

20 'classical' splicing factors of the SR, SR-related and hnRNP protein families

Quantify variation in gene expression between pairs of tissues by computing 20-dimensional correlation coefficient  $\rho$  between 16 plus 10 additional tissues

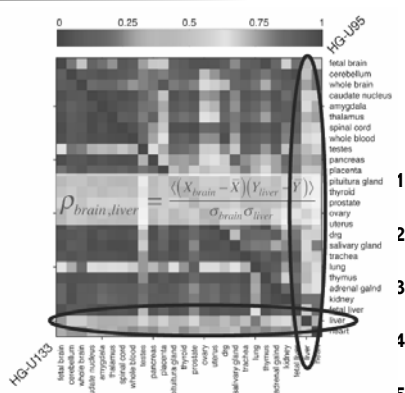
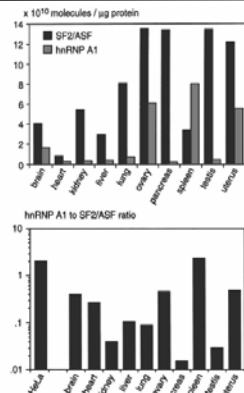


Computational analysis of alternative splicing

114

Krainer Lab (data: hnRNP A1, ASF/SF), Yeo et al (2005)

## Gene expression of *trans*-acting splicing factors across tissues



Computational analysis of alternative splicing

115

Krainer Lab (data: hnRNP A1, ASF/SF), Yeo et al (2005)

## Splicing-conservation of AS events

"With enough data, aberrant splicing events may cause virtually all genes to appear alternatively spliced"

Baek & Green PNAS (2005)

- Are many reported AS events due to noise?  
(on experimental as well as reproducible cellular "biological" level)  
Or: Does AS provide a frequent shortcut in evolution?
- Are we still ignoring a lot of AS events despite millions of ESTs?
- EST alignments deliver a relatively small number of splicing-conserved AS events (compared to #AS exons) across species (up to ~1,000), and often only the more frequent isoform is splicing-conserved Modrek & Lee Nat Genet (2003)
- Conserved AS** is likely to be more important/functional
  - Against species-specific AS, but can be advantageous to filter spurious events

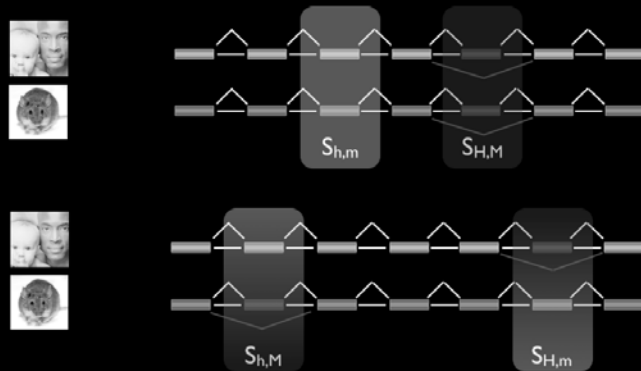


Computational analysis of alternative splicing

116



## Evolutionary conservation of alternative splicing events



Computational analysis of alternative splicing

117

## Evolutionary conservation of alternative splicing events



**Databases:** events of alternative splicing-conserved exons (ACEs) in human and mouse transcriptomes --- is the encoded sequence information sufficient to distinguish between ACEs and constitutive exons? Can we identify & utilize intrinsic features and build computational models that predict ACEs?

"Despite the high fidelity of exon recognition in vivo, it is currently impossible to accurately predict alternative exons" -Thanaraj & Stamm *Prog Mol Subcell Biol* (2003)

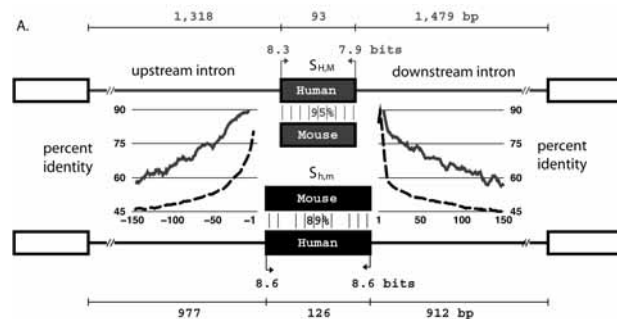
"Prediction of AS events is not yet possible" -Bour, Letunic, and Bork *BioEssays* (2003)



Computational analysis of alternative splicing

118

## Distinct sequence features of alternative conserved exons (ACEs)



Generic **sequence features** evaluated by comparing splicing-conserved constitutive and alternative exons



Computational analysis of alternative splicing

119

## Computational models for predicting AS events (skipping, retention) from sequence

| Algorithm                                                                                                                                                                                          | Web-based reference                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| Predicted about 1,000 AS events conserved between orthologous human-mouse genes, based on feature selection and use of rule-based computational model                                              | <a href="http://www.tau.ac.il/~gilast/">http://www.tau.ac.il/~gilast/</a><br>(Boolean, SVM)                           |
| Predicted a set of about 160 orthologous splicing-conserved alternative exons in the fly species <i>Drosophila melanogaster</i> and <i>D. pseudoobscura</i> , with 42% validated                   | <a href="http://penguin.uchc.edu/~intron/philipps">http://penguin.uchc.edu/~intron/philipps</a><br>(Similarity-based) |
| Predictive identification of about 2,000 alternative splicing-conserved exons, expressed in orthologous human-mouse genes, based on structural features and motifs trained in ACEScan algorithm    | <a href="http://genes.mit.edu/acescan">http://genes.mit.edu/acescan</a><br><b>AceScan (SVM-like)</b>                  |
| Predicted about 50 splicing-conserved skipped exons in human ENCODE regions, and about two dozens conserved retained introns of orthologous human-mouse genes                                      | <a href="http://www.philip@delaware.edu">www.philip@delaware.edu</a><br><b>UNCOVER (pHMM)</b>                         |
| Designed kernels for the application of support vector machines to identify skipped exons in <i>C. elegans</i> . When applied to the worm genome, ~200 candidate alternative exons were identified | <a href="http://www.fmi.tuebingen.mpg.de/~raetsch/">http://www.fmi.tuebingen.mpg.de/~raetsch/</a><br>(SVM)            |



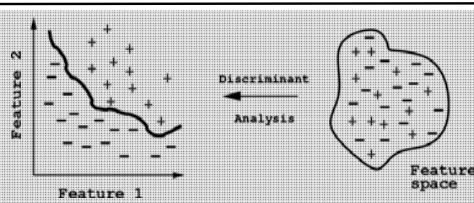
Computational analysis of alternative splicing

120



## Classification of known exons

Look at one sequenced-conserved exon and decide whether it is skipped



### Issues

- The problem is 'ill-posed'
- High-dimensional space
- Avoid over-fitting data
- Included-conserved exons may be ACEs

### Addressed

- Regularized least-squares classifier
- Introduced parameter to weight ACEs higher
- Feature selection



Computational analysis of alternative splicing

121

## Identification of alternative conserved exons (ACEs) from sequence by AceScan

| Upstream intronic region |         |             | Exon region |         |             | Downstream intronic region |         |  |
|--------------------------|---------|-------------|-------------|---------|-------------|----------------------------|---------|--|
| Unaligned                | Aligned | Exon 5' end | Unaligned   | Aligned | Exon 3' end | Unaligned                  | Aligned |  |
| GGCCG                    | UGCAU   | CCUCC       | UGUAG       | UAGGG   | CCUCC       | GCAUG                      | UGCAU   |  |
| UUUCC                    | GCAUG   | CUCCC       | GUAGU       | CUCCG   | CCUCC       | UGCAU                      | GCAUG   |  |
| UUUCU                    | CGGGG   | CCUCC       | ACUAG       | CGGGG   | CCUCC       | CGAUC                      | UGCAU   |  |
| UUUCU                    | ACACU   | CAUCC       | UAGAA       | UACGA   | CUCCC       | CAUGC                      | UUGUC   |  |
| UUUCC                    | UUAAC   | CUCCC       | UUCCU       | UUCCU   | UUCCU       | CAUUG                      | CAUUG   |  |
| CUUCU                    | ACUAC   | CCCCC       | CGAGG       | AGUAG   | GUCCC       | GUUUG                      | UGCCG   |  |
| CUUUC                    | CUUAC   | ACAAA       | CCGGG       | CCGGG   | UUCCU       | CUAAC                      | UCCGG   |  |
| UUUUU                    | UUCAU   | CGGGG       | CUACG       | AGUUA   | UUCCC       | CUAAC                      | CUAAC   |  |
| CGCCG                    | CUUAC   | CCUCC       | UUACG       | UUAAU   | GAAGG       | CAUUC                      | GUUUG   |  |
| UUUCU                    | UUUAC   | UUUCC       | UUUAC       | UUUAC   | CCUCC       | CAUUC                      | GUUUG   |  |
| CGCCU                    | GUAGG   | UUUCU       | UUUAC       | UUUAC   | CGAGG       | CGAGG                      | CGAGG   |  |
| UGCAU                    |         | CGGGG       | UAGUG       | UUCCU   | AAAGG       | UAGU                       | GAGCA   |  |
| CUUUC                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| UUUUC                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| GUUUC                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| GGACA                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| UGGAG                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| GGUUG                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |
| GACAG                    |         | UUUCC       | UUUCC       | UUUCC   | UUAAA       | UAGAA                      | AGUAG   |  |

Oligonucleotide features over- or under-represented in exons and introns (sequence-aligned and unaligned)



Computational analysis of alternative splicing

Yeo et al (2005)

122

## Regularized least-squares classification

$y_i \in \{-1, +1\}$  binary labels for examples  $\mathbf{x}_i$  ( $i = 1, 2, \dots, L$ )

$$\min_{\mathbf{f}} (1/L) \sum_i w_i (y_i - \mathbf{f}(\mathbf{x}_i))^2 + \lambda \|\mathbf{f}(\mathbf{x})\|_{\mathcal{H}}^2$$

$\mathbf{f}$ , solution

$\|\mathbf{f}(\mathbf{x})\|_{\mathcal{H}}^2$  induced function norm

$\lambda$ , generalization/over-fitting tradeoff

$w_i = \beta$  if  $y_i = 1$ , otherwise  $w_i = 1$

Assuming a solution  $\mathbf{f}^*$  to the above problem has the form:

$$\mathbf{f}^*(\mathbf{x}) = \sum_i c_i K(\mathbf{x}, \mathbf{x}_i)$$

where  $K(\mathbf{x}, \mathbf{x}_i) = \langle \mathbf{x}, \mathbf{x}_i \rangle$  and  $c_i$  are coefficients, after substitution one obtains

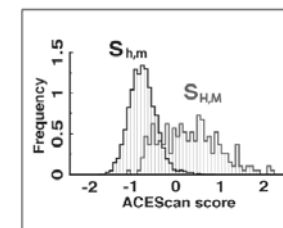
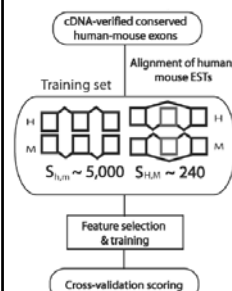
$$(K(\mathbf{x}_i, \mathbf{x}_j) + \lambda L \mathbf{W}^{-1}) \mathbf{c} = \mathbf{y}$$


Computational analysis of alternative splicing

Rifkin et al (2003), Yeo et al (2005)

123

## Evolutionary conservation of alternative splicing events



### Learning



### Prediction



### Validation

**Constant features:** exon & intron lengths, conservation, splice sites (~10)

**Selected features:** most discriminative 4- and 5-mers in exonic and intronic regions flanking splice site (~200)



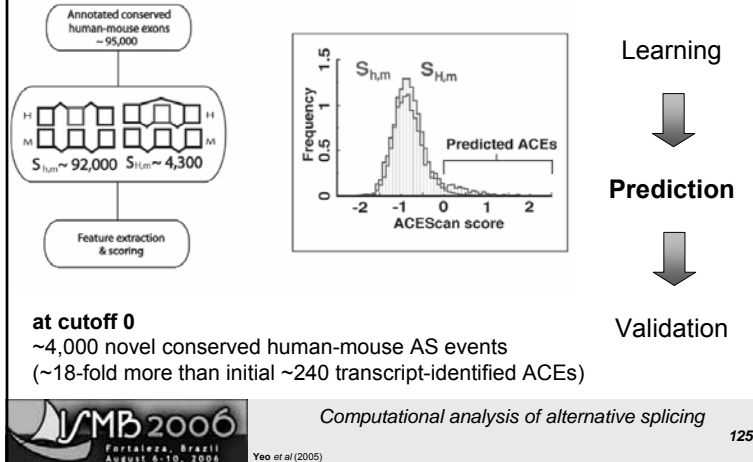
Computational analysis of alternative splicing

Yeo et al (2005)

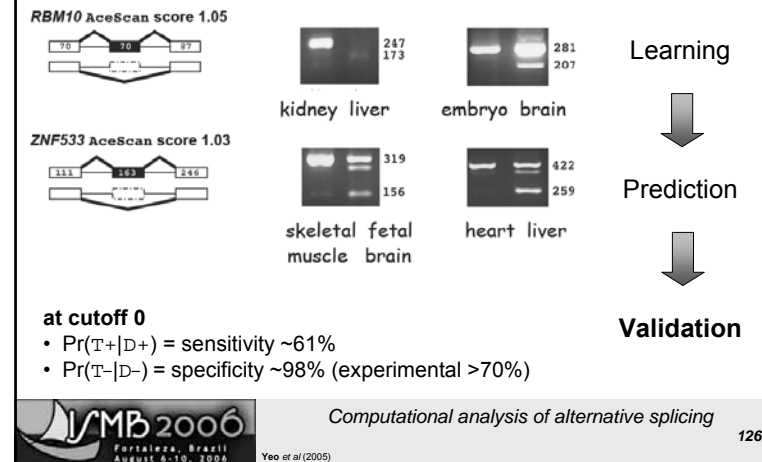
124



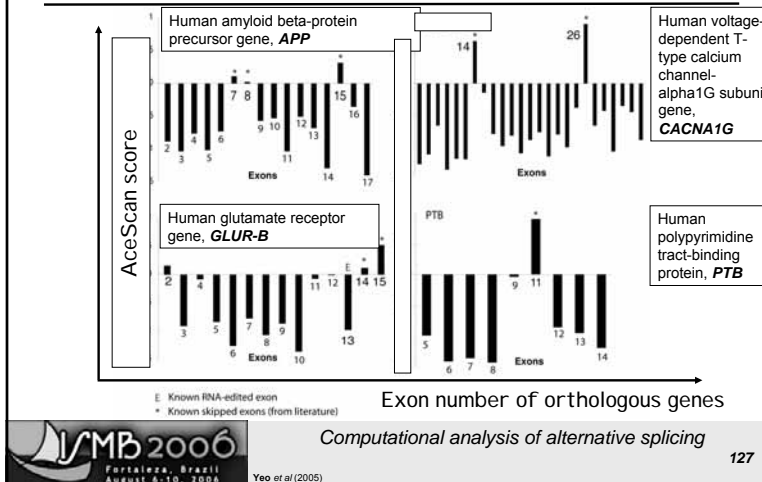
## Evolutionary conservation of alternative splicing events



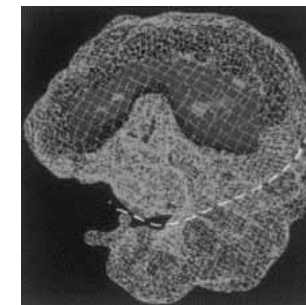
## Evolutionary conservation of alternative splicing events



## Evolutionary conservation of alternative splicing events

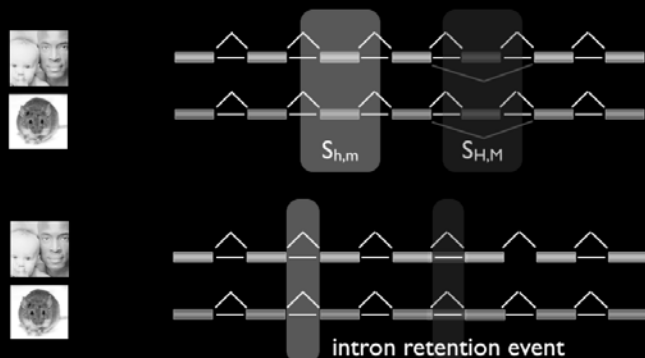


## QUESTIONS FOR THIS SECTION





## Identification of rare AS events: retention-type introns

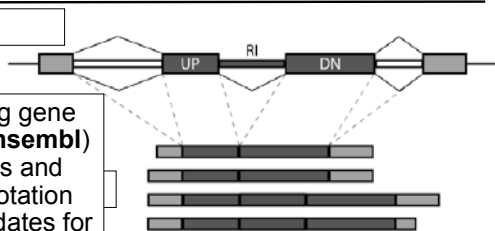


## Identification of retention-type introns

- Transcript-based annotation of AS events show the type "exon skipping" as most frequent (~40%), while **retention-type introns** are much less frequent
  - Retention-type introns are least studied AS event
  - Anecdotal evidence exists (fly *Msl2*; human *Klk* gene family, Pro1B gene developmentally regulated, hnRNP M)
  - 30% of AS events in *Arabidopsis* estimated to be retention-type events, estimates in human transcriptome up to 15% (Galante *et al*)
  - Problematic type of AS: unspliced/partially-spliced pre-mRNAs are "noise" rather than purposefully (functional) retained introns
  - Filter: evolutionary conservation of splice form (retained, spliced)
- Idea: look locally and examine orthologous introns of already known genes

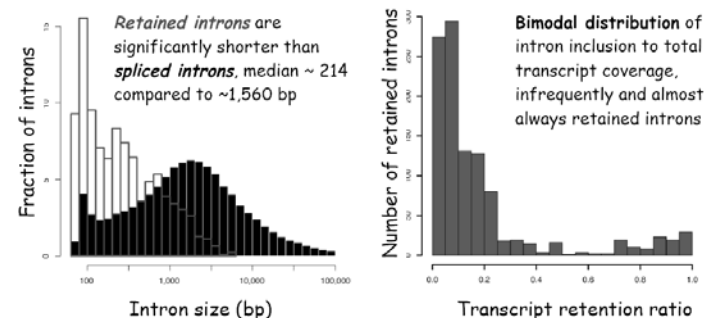
## Identification of retention-type introns

- Use pre-existing gene models (e.g. **Ensembl**) and map cDNAs and ESTs onto annotation to detect candidates for retention-type events



- Identify ~1,300 human AS events, characterize ORF alteration, features, and search for evolutionary conserved genes harboring retention-type introns

## Feature analysis of retention-type introns



- Two **discriminative features** of retention-type introns: length distribution and inclusion of introns in transcript coverage
- Others: splice sites, coding potential, repetitive sequence elements, ...



## Classification of three types of retained introns based on reading-frame

| Category | Reading frame preservation or disruption |                                                                                                                                 |
|----------|------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| RF++     | +                                        | Retention and splicing maintain "bi-functional" ORF (~130 introns, average %inclusion, no repeats, median splice site strength) |
| RF+-     | +                                        | Retention-only maintains ORF (~50 introns, high %inclusion, lower splice site strength)                                         |
| RF-+     | -                                        | Splicing-only maintains ORF (~900 introns, low %inclusion, stronger splice site strength)                                       |



Computational analysis of alternative splicing II

133

van Nostrand, Holste, Burge (unpublished data)

## RT-PCR validation of selected intron-retention-type introns

| Reading frame class | RT-PCR validated |            |            | Example gel |
|---------------------|------------------|------------|------------|-------------|
|                     | Splicing         | Retention  | Both       |             |
| RF++                | 6/6              | 4/6        | 4/6        | hnRNP M     |
| RF+-                | 2/6              | 5/6        | 2/6        | POLR3C      |
| RF-+                | 5/6              | 2/6 (*3/6) | 2/6 (*3/6) | P2RX7       |

- All three classes of human retention-type introns are produced, but only a small number (~30) of such AS events are co-identified in the transcripts of orthologous human-mouse genes (why not?... rare AS event, species- or allele-specific, disease-associated, aberrant)



Computational analysis of alternative splicing II

134

van Nostrand, Holste, Burge (unpublished data)

## Identifying unknown skipped exons & retained introns

Q: 541 gccgcagctgcagacagcccgctggaacaagaggtggcttcgtgctcaacttccatgcg  
S: 541 gccgcagcc---gacagcccgccggcaccgcggcggttctcgtgctcaacttccacgca

Q: 601 gacacggaactg---ggcaagaagaaggcgccctcttcctcggtgggttccttctcggc  
S: 598 gacgctgagctagcgggcaagaagaaggcgccctcttcggagggttccttcttggga

- Such patterns are the basis of comparative gene finding --- can they be used to find new cases of alternatively spliced exons?
- Idea: look locally and examine orthologous introns of already known genes
- Will mostly lead to new minor form AS events

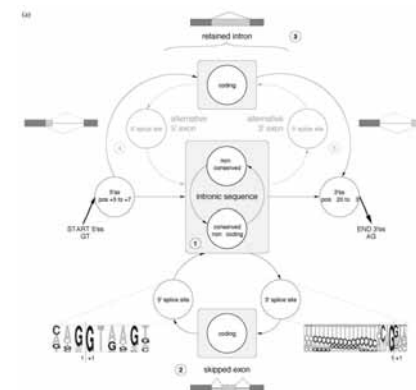


Computational analysis of alternative splicing II

135

van Nostrand, Holste, Burge (unpublished data)

## UNCOVER: a pHMM to predict AS



Computational analysis of alternative splicing II

136

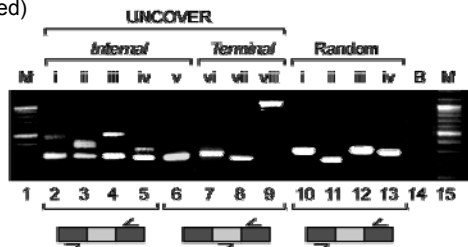
Ohler et al (2005), also: Raetsch et al. (2005)



[illegible]

- 150,238 human introns in **Ensembl** (July 2005)
- 80,028 of these are orthologous in mouse
  - BLAST hit spanning across 30 nt exon junctions
- UNCOVER **pipeline**:
  - Pre-screen intron pairs by BLAST with cutoff 1e-10
  - Repeat-masking, no low-complexity filtering
  - Initial 9,547 candidates; 8,836 not yet annotated
- 8,589 human/mouse prediction pairs (in 3,430 introns) with no prior annotation in either species
  - Combination with AceScan: 70% (~6,000) of UNCOVER predictions score positive (!)
  - Estimate of total of **~10,000 ACES in the human genome**

- Test of 20 ENCODE predictions flanked by strong splice sites
  - Several primer pairs to flanking exons and/or flanking plus predicted exon
  - RT-PCR of several human adult tissues (brain, liver, plus 7 more; 15/20 expressed)



| Name                                    | Spliced EST | In-frame | Mod 3 | Protein domain |
|-----------------------------------------|-------------|----------|-------|----------------|
| <i>Known Ensembl / EST / Vega genes</i> |             |          |       |                |
| HNRPL                                   | X           | -        | X     | X              |
| FXR1                                    | X           | X        | X     | X              |
| Q7M4L6                                  | X           | X        | X     |                |
| HNRPM                                   | X           | X        | X     | X              |
| HOX A1                                  | X           | X        |       |                |
| TRPC4                                   | X           | X        | X     | X              |
| PJA1                                    | X           | X        | X     | X              |
| BCL2L11                                 | X           | X        | X     |                |
| NM18178                                 | X           | X        | X     |                |
| GRIN1                                   |             | X        |       |                |
| SALL3                                   |             | X        | X     |                |



- Analysis of the whole genome
- Only handful of conserved cases known



## Conserved intron retention II

| Name                   | Spliced EST | In-frame | Mod 3 | Protein domain |
|------------------------|-------------|----------|-------|----------------|
| <i>New predictions</i> |             |          |       |                |
| BAT8                   | X           | -        |       |                |
| NM174926               | X           | X        | X     | X              |
| novel                  | X           | X        | X     |                |
| PAX6                   |             | ?        |       | X              |
| FAP                    |             | ?        |       |                |
| PCDH17                 |             | X        | X     |                |
| EBF3                   |             | -        | X     |                |
| PPARGC1                |             | -        |       |                |
| novel                  |             | ?        |       |                |
| novel                  |             | ?        |       |                |
| HTATIP                 |             | -        |       |                |
| LACE1                  |             | -        |       |                |



– Intron retention does not play a major role in mammalian coding alternative splicing

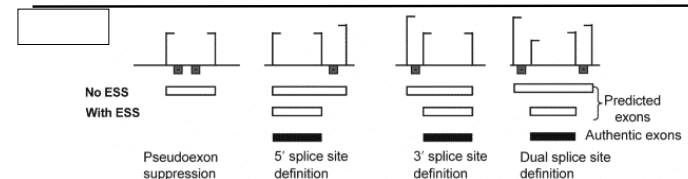
## Summarizing some points on alternative splicing covered so far

- Only a fraction of EST derived AS events is conserved
- Nevertheless, a large number of conserved AS events/exons is still unknown
  - but can be predicted
- Splicing has moved from mere EST alignments to more diverse set of computational problems
- Large-scale experimental studies will provide a wealth of functional data on AS
- Ideal computational scenario: A splicing *simulator*

## Simulating splicing

- Predict splicing from the viewpoint of the cell:  
Given a pre-mRNA, what does the processed transcript (or set of transcripts) look like?
  - Does *not* use:
    - conservation, measures of coding potential etc.
  - Does use:
    - splice site models which cover the region of the splice site interacting with the spliceosome
    - Enhancer/silencer motifs
    - Context, i.e., distance between splice sites, enhancers and splice sites, ...
  - Current success: ~70% for short transcripts (<10 kB)

## Simulating splicing: ExonScan



Predicting exon boundaries from genomic signals (no reading-frame information)

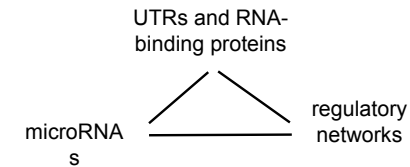
| Features used in ExonScan | Improvement in accuracy relative to SS only (+/=2%) |         |          |         |        |         |
|---------------------------|-----------------------------------------------------|---------|----------|---------|--------|---------|
|                           | <10kbp                                              |         | 10-30kbp |         | >30kbp |         |
|                           | Exact                                               | Partial | Exact    | Partial | Exact  | Partial |
| SS only                   | 0                                                   | 0       | 0        | 0       | 0      | 0       |
| SS+GGG                    | ++                                                  | ++      | ++       | ++      | +      | +       |
| SS+ESE                    | ++                                                  | ++      | ++       | ++      | +++    | ++++    |
| SS+ESS                    | ++++                                                | ++++    | ++++     | ++++    | ++++   | ++++    |
| SS+ESE+ESS                | +++++                                               | +++++   | +++++    | +++++   | +++++  | +++++   |
| SS+ESE+GGG+ESS            | +++++                                               | +++++   | +++++    | +++++   | +++++  | +++++   |



## QUESTIONS FOR THIS SECTION



## IV. Beyond splicing

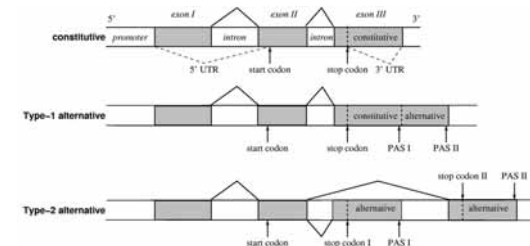


## Other mechanisms

- Increased appreciation for **RNA regulatory mechanisms** other than splicing
  - Polyadenylation (can exhibit alternative poly(A) sites)
  - Translation
  - Export/Localization
  - Stability
- Driven by realization of importance of regulatory non-coding RNAs
  - Caution: not universal across species, e.g., eukaryotic riboswitches occur only in plants, miRNAs not in yeast, ...

## 3'-untranslated regions (3' UTRs)

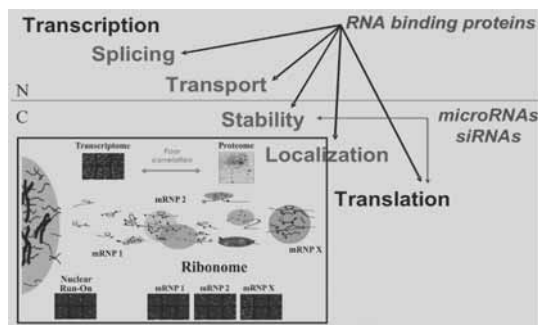
- Many PT *cis*-regulatory elements are located in the 3' untranslated region
  - Between the stop codon and the cleavage site
  - Usually completely contained within the terminal exon
  - Can be quite extensive; up to several tens of kb





## Trans-acting factors

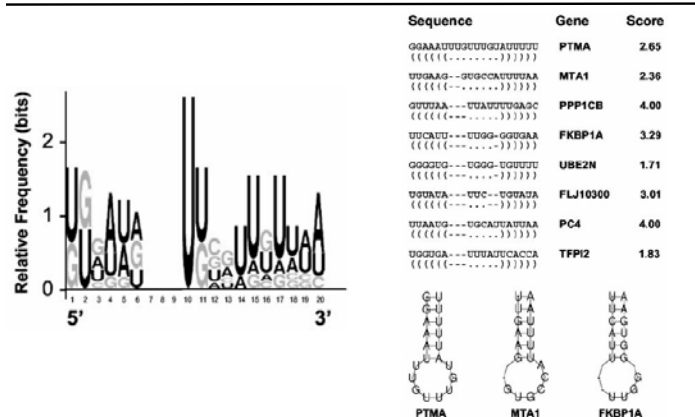
- Similar to transcription factors, a large number of RNA binding proteins (RNPs) can **interact with motifs** in 3'UTRs



## Example: Predicting *HuR* binding sites

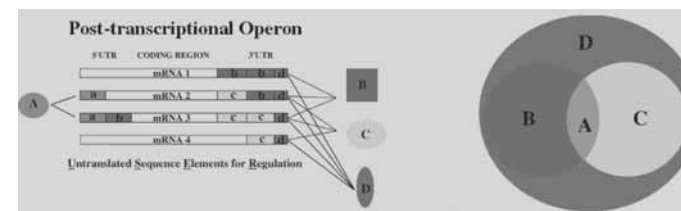
- Array analysis identified **57 strong targets** of the RNA-binding protein *HuR* (a “stabilizer”)
- Sequences were searched for over-represented **RNA motifs** with Foldalign (Gorodkin *et al.*, 1997)
  - Secondary structures confirmed by MFOLD
- Curated set was used to **train a stochastic context free grammar model** (Eddy & Durbin, 1994)
  - Parameters for single nucleotides and paired dinucleotides
  - Hits are preferentially located in 3'UTRs and conserved in other mammals
  - New predictions were experimentally validated

## *HuR* motif



## Combinatorial genetic control

- PT events are **co-regulated**
  - RIP-chip: Analogous to ChIP-chip, to identify target genes bound by specific RNP
  - Diverse function, e.g., localization or stability
  - Combinatorial control



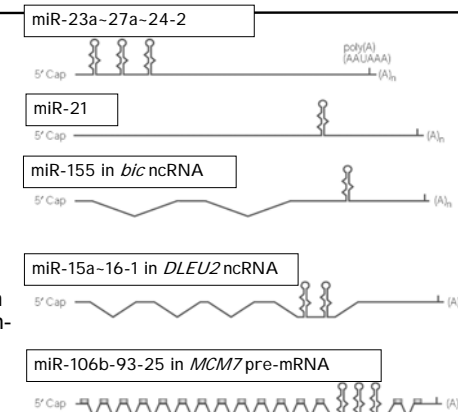


## Non-coding regulatory genes

- Last few years have uncovered new regulatory layer
- Prominent and increasingly well understood case: **miRNAs**
  - Small regulatory RNAs which repress target genes
  - More than 30% of human genes is estimated to be influenced
- Processing is in parallel to protein coding genes:
  - Primary transcripts (several kb; nucleus; RNA pol II)
  - Precursor foldbacks (70 nt; nucleus; Drosha)
  - Mature miRNA (20-25 nt; cytoplasm; Dicer)

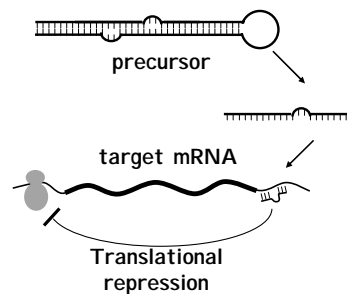
## Location of miRNA genes

- miRNAs come in a variety of disguises
  - Can be independently transcribed
  - Can be intron/exon of a non-coding transcript
  - Can be part of an intron of a protein-coding gene
  - Can come in clusters



## What is the function of microRNAs?

- Animal miRNAs target protein coding genes through complementary sequence regions in their 3' UTR
- “Natural counterpart” to siRNAs/RNAi



- One miRNA influences many target genes
- One miRNA can have several target sites in one UTR
- One UTR can have multiple miRNA targets
- miRNA genes and targets are well conserved

## Identification of miRNA genes (I): conserved foldbacks

Example: original miRscan flowchart (Lim et al 2003)

1. Scan *C.elegans* genome for potential RNA hairpin structures
  - Fold every 110 base segment in genome using RNAfold (Vienna RNA software package, Hofacker et al)
2. Identify hairpins with homology to *C.briggsae* shotgun traces
  - WU-BLAST cutoff E1.8; RNAfold *C.briggsae* sequence
3. Align *C.elegans* and *C.briggsae* hairpins
  - Pair must have certain secondary structure similarity
4. Classify foldback into miRNA/no miRNA using features representative of miRNAs



## Identification of miRNA genes (II)

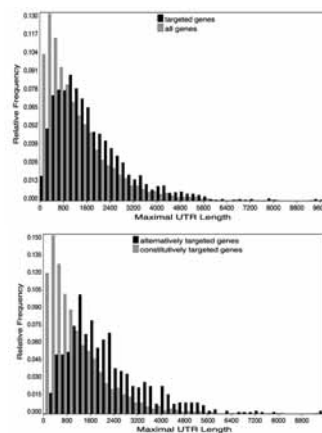
- “**Reverse engineering**” approach (Xie *et al*, 2005)
  - Look for conserved UTR motifs
  - Predict conserved those foldbacks as miRNAs which contain a complementary stretch
- **Single-genome** approach (Pfeffer *et al*, 2005)
  - Conservation gives strongest signal, but cannot be used in all circumstances (i.e. viral genomes)
  - Idea: miRNA foldbacks are stable and should not be influenced too much by varying prediction parameters
    - E.g. size/positioning of sliding window

## Identification of miRNA targets

- All current well-known target predictors are based on conserved “**seed**” matches in 3' UTR
  - Positions 2-7 in the miRNA were found to be crucial for targeting, and also show strongest conservation
  - Some variations in whether/ how to take complementarity outside the seed into account
  - Some variations which species to include/ whole genome alignments versus independent
- Very few experimentally, *functionally* verified targets
  - Predictions evaluated on “random” sequences or seeds with similar base composition

## Alternative UTRs versus miRNAs

- **PT co-regulation:**  
Targeted genes have greater UTR length
- Targeted genes have alternative poly(A) sites more frequently
- In general, some genes subject to many PT events, while others not at all



## Open issues

- How exactly do **miRNA/target interactions** work?
  - Seed matches vs extended near-perfect complementarity
  - Degradation vs blocking of targets
  - Down-regulation vs stabilization
- Are functional targets always in the UTRs only and why?
- How well conserved is this? Are we missing stuff?
- How to clearly define miRNAs? How to keep them apart from other regulatory ncRNAs?



## V. Summary and open issues

## A summary of the topics in four slides, I

- **Splicing** has moved from mere EST alignments to more **diverse set of computational problems**
- Large-scale experimental studies will provide a wealth of data on AS
  - High throughput binding assays for regulatory factors
  - RIP-Chip for RNA binding proteins
- In particular, **splicing arrays** have scaled up from exploratory to commercial products
  - Current caveat:  
How were the probes derived that are on the chip?
  - High resolution tiling arrays as alternative?

## A summary of the topics in four slides, II

- Towards an **annotation of post-transcriptional regulatory elements**
- Identification of functional sequences (splicing silencers/enhancers)
  - Computational combined with experimental approaches have extended the available vocabulary for exonic regulatory sequence elements
  - Correlation between those mechanisms (e.g., alternative promoters/ alternative splicing/ alternative poly-adenylation)

## A summary of the topics in four slides, III

- A wide variety of **specialized databases**
  - Many follow stringent criteria of quality to arrive at reliable sets of isoforms
  - **Exon-centric vs transcript-centric**
  - “Value-added” databases become more frequent: Display functional information in addition to mere isoforms
  - Good starting point for computational analyses
  - Still --- How easy are they to handle? How valuable to the wet-bench researcher?



## A summary of the topics in four slides, IV

- *Ab initio* based **prediction methods** have become available
  - Prediction of exon skipping/intron retention solely from the genomic sequence
  - Will lead to a more comprehensive set of isoforms – hard to find even the last AS event from ESTs alone
  - Most of them use comparative features – how much AS is conserved and functional?
  - Some approaches can use one genome only; successful for *C.elegans*, promising for human
  - Splicing simulators have begun to systematically address this



Summary

165

## Some open issues: splicing as a controlling cellular processes

- Like everything else in biology, splicing is not a 100% error proof process
- “Nonsense-mediated mRNA decay” (**NMD**) is able to detect premature termination codons (**PTCs**)
  - Have to occur upstream of the terminal exon
  - Transcripts get degraded and do not result in truncated proteins
- Important in relation to alternative splicing
  - Many **alternative transcripts lead to PTCs**
  - To which extent is this a regulated effect?



Open issues

Lewis et al (2003), Pan et al (2006)

166

## Open issues: subtle and complex splice variants

- Many EST- and all sequence- based approaches heavily centered on exon level
- Less attention to **subtle 5'/3' splice site choices**, micro-exons of few nt lengths
  - Example NAGNAG splice sites: function or noise?
- Extension to >2 species will help (but not solve)
  - Increase confidence in predictions
  - Detect subtle variants: alternative 5'/3', short exons
- Conversely, how to predict a whole alternative transcript instead of only alternative splice sites?



Open issues

Hiller et al (2004, 2006), Chem et al (2006)

167

## Open issues: evolution of splicing

- Splicing present in all eukaryotes
  - But: *S.cerevisiae* only has about 200 introns total, and *S.pombe* no AS
  - **Complexity increases in higher organisms:** Canonical splice sites degenerate, new functional elements emerge
  - Interplay of *cis*-element evolution and diversification of *trans*-acting factors
- How did the eukaryotic gene structure emerge?
  - How much plasticity is there?
  - Ongoing studies, e.g., in fungal genomes



Open issues

Ast (2005), Lynch (2006), Stajich & Dietrich (2006)

168



## Open issues: regulatory networks

- Cell has to coordinate many different control mechanisms acting on a transcript
- **Interconnections** appear on many different levels
  - Auto-regulation, i.e., splice factors which influence their own isoforms (e.g. PTB)
  - Coordination of transcription initiation/termination and splicing
  - Degradation by NMD
  - microRNA processing vs splicing
  - Potential binding sites of both stabilizers (HuR) and destabilizers (microRNAs)



Open issues

169

## Open issues: computation

- **Networks** and their **properties**, emergence
  - Combinatorial control
  - Rich probabilistic models
  - Discrete vs continuous
- **Integrating different types** of biological data:
  - Sequence (genes, alleles, comparative genomics)
  - Expression and decay
  - Functional association (linkage)
  - Physical interaction (binding, interaction)
  - Time: stages of development; Context: specific tissues
  - Transport and localization
  - Activity (phosphorylation)



Open issues

170

## Attempt to look into the near and more distant future

- **Computational sequence analysis**
- **Functional genomics data:** splicing microarrays
  - Exon-junction and whole genome tiling
- **Computational modeling**
  - Identification of all functional and aberrant transcriptional variants: tissue, cell-type specific
  - Predicting AS events from sequence/sequence-expression data/sequence-protein levels
  - Understanding the different roles of AS in different cell-types, stages of development, extracellular stimuli
  - Predicting the complete set of isoforms from pre-mRNA sequence alone (simulating RNA splicing)



Future issues

171

## Attempt to look into the near and more distant future

- **Clinical data: disease and therapy**
  - Libraries of diseased tissue for enriched for alternative exons
  - Splice variants as cancer biomarkers
  - Design and functional impact of antisense oligonucleotides
- **Data coming up and cool to have at hand:**
  - ESTs > 10,000,000 (SAGE, CAGE)
  - Libraries of normal/diseased tissue for enriched for alternative exons
  - More arrays in all flavors, coding and non-coding genes
    - Standard, tiling, splicing-sensitive
    - Chip-on-chip
    - Protein-binding arrays
  - MassSpec/peptide libraries
  - Selection (SELEX)
  - Experimental 3D structures of RNAs
  - Functional genomics (genome-wide RNAi)
  - Small molecule screens



Future issues

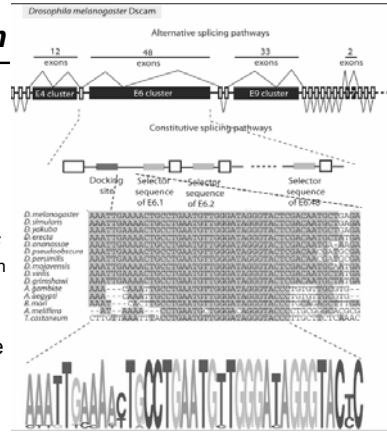
172



## Selector:docking mechanism in *Dscam*

*"The observation that [...] the sequences are complementary [...] occurred through a combination of **staring** at the sequences for months and sheer luck."* Brenton Graveley (2005)

While admirable, what do we need to be staring less and still be lucky to find mechanisms such as selector:docking sites?



Last slide - Stop staring

173

## The Presenters



Dirk Holste, PhD

Institute of Molecular Pathology (Vienna)

Dirk is a Group Leader in computational molecular biology with research interests in sequence analysis, transcriptome quantitative analysis, comparative genomics, evolution and mathematical-biological modeling. He lectured in mathematics, biophysics, and computational biology courses for undergraduate and graduate students. Since 2006, Dirk supervises the bioinformatics for the RNAi facility at the Center for Functional Genomics joint at the Institutes for Molecular Biology (IMP) and Molecular Biotechnology (IMBA).

Uwe Ohler, PhD

Duke University (Durham)

Uwe is an Assistant Professor in computational biology at Duke University, at the Institute for Genome Sciences and Policy. He has extensive research experience in experience in pattern recognition and machine learning, sequence analysis, comparative genomics, evolution and computational-biological modeling. He lectures in computer science and computational biology courses for undergraduate and graduate students. Since 2005, he has served as a member of the curriculum and student advisory committees of the Duke graduate program in Computational Biology and Bioinformatics.

Dirk and Uwe met while being postdoctoral scholars at MIT (Dept Biology, Chris Burge's lab)



174