

Coevolutionary networks of splicing *cis*-regulatory elements

Xinshu Xiao, Zefeng Wang*, Minyoung Jang, and Christopher B. Burge†

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Philip P. Green, University of Washington School of Medicine, Seattle, WA, and approved October 4, 2007 (received for review August 3, 2007)

Accurate and efficient splicing of eukaryotic pre-mRNAs requires recognition by *trans*-acting factors of a complex array of *cis*-acting RNA elements. Here, we developed a generalized Bayesian network to model the coevolution of splicing *cis* elements in diverse eukaryotic taxa. Cross-exon but not cross-intron compensatory interactions between the 5' splice site (5'ss) and 3' splice site (3'ss) were observed in human/mouse, indicating that the exon is the primary evolutionary unit in mammals. Studied plants, fungi, and invertebrates exhibited exclusively cross-intron interactions, suggesting that intron definition drives evolution in these organisms. In mammals, 5'ss strength and the strength of several classes of exonic splicing silencers (ESSs) evolved in a correlated way, whereas specific exonic splicing enhancers (ESEs), including motifs associated with hTra2, SRp55, and SRp20, evolved in a compensatory manner relative to the 5'ss and 3'ss. Interactions between specific ESS or ESE motifs were not observed, suggesting that elements bound by different factors are not commonly interchangeable. Thus, the splicing elements defining exons coevolve in a way that preserves overall exon strength, allowing specific elements to substitute for loss or weakening of others.

exon definition | intron definition | motif | regulatory network | RNA splicing

Choice of splice sites in nuclear pre-mRNA splicing involves a complex set of recognition events. Motifs at the 5' splice site (5'ss), the polypyrimidine tract (PPT) and 3' splice site (3'ss), and the branch point sequence (BPS) are required for splicing, but lack sufficient information content to define exon locations in most organisms (1). Auxiliary splicing regulatory elements (SREs), known as exonic splicing enhancers (ESEs), intronic splicing enhancers (ISEs), exonic splicing silencers (ESSs), and intronic splicing silencers (ISSs), are defined by their effects on adjacent splice sites, e.g., ESEs tend to promote inclusion and ESSs promote exclusion of the exons they reside in. SREs control splice site choice by recruitment of specific *trans*-acting factors to the pre-mRNA that function to either activate or repress splicing by interaction with other regulatory factors or core spliceosome components (reviewed in refs. 2 and 3).

A prominent feature of splicing regulation is the elaborate network of interactions that occur among and between core and auxiliary splicing factors (4). Two general models have been proposed for early spliceosome assembly (5). In the intron definition model, intron-spanning interactions between factors recognizing the 5'ss and the downstream 3'ss occur initially; this model is supported for transcripts with large exons and small introns (6, 7). In the exon definition model, exon-spanning interactions between factors that recognize the 5'ss and factors bound at the upstream 3'ss lead to formation of exon definition complexes before rearrangements that yield intron-spanning spliceosomes. Exon definition is thought to predominate in splicing of mammalian internal exons, which are typically moderately sized (≈ 50 –250 bp) and often flanked by much larger introns (5, 8). Other interactions of general importance for splicing, including contacts between small nuclear ribonucleoprotein (snRNP) components bound to splice sites and proteins of the serine-arginine-rich (SR) protein family, most of which

bind ESEs (9–11). Conversely, several heterogeneous nuclear ribonucleoproteins (hnRNPs) bind to ESSs and inhibit exon recognition or alter splice site choice by antagonizing the activity or interactions of SR protein and/or snRNP factors (12–14). These interactions among splicing factors appear vital to the robustness, precision, and adaptability of the splicing reaction.

Sequences involved in the splicing of constitutive exons are expected to be under selection to optimize the network of splicing factor interactions for accurate and efficient splicing. To understand how such selection influences gene evolution, we studied the coevolution of splicing *cis*-regulatory elements in orthologous exons and introns. For example, a mutation that eliminated a critical ESE in an exon might be compensated by a second mutation strengthening a splice site to restore efficient splicing (Fig. 1A). Because several distinct *cis* elements are involved in splicing of a typical constitutive exon (15, 16), pairwise analysis of elements is inadequate, and the system is better modeled as a network of interacting nodes, each representing a particular *cis* element.

Network formalisms are widely used for modeling sets of physical and functional interactions between proteins (e.g., refs. 17 and 18). Here, we developed a network formalism for modeling evolutionary relationships among *cis* elements, which we call principal variation-based subset analysis (PVSA). Applied to large sets of orthologous exon pairs from related organisms, PVSA identified major trends in splicing element coevolution that differ between mammals and other eukaryotic taxa and uncovered unexpected complexity in the interplay between splice sites and various types of SREs.

Results

Modeling Coevolution of *cis* Elements. Splicing of a typical constitutive mammalian internal exon depends on a fairly broad set of *cis* elements, including the 5'ss, 3'ss/PPT, BPS, and multiple ESEs, and is often modulated by the presence of ESSs, ISEs and ISSs (3, 15, 16). Mutation of any of these elements may disrupt splicing, most often causing partial skipping of the mutated exon, thereby reducing the quantity of full-length mRNA and protein produced (19, 20). Substantial reduction in the level of a gene product is often deleterious, sometimes resulting in disease (reviewed in refs. 21 and 22). If only mildly deleterious, the mutant allele may survive for sufficient time to acquire secondary mutation(s) that suppress the disruptive effects of the original mutation, e.g., as illustrated in Fig. 1A.

Our goal was to understand the evolutionary relationships

Author contributions: X.X. and C.B.B. designed research; X.X., Z.W., and M.J. performed research; X.X. contributed new reagents/analytic tools; X.X. and Z.W. analyzed data; and X.X., Z.W., and C.B.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*Present address: Department of Pharmacology, University of North Carolina, Chapel Hill, NC 27599.

†To whom correspondence should be addressed. E-mail: cburge@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0707349104/DC1.

© 2007 by The National Academy of Sciences of the USA

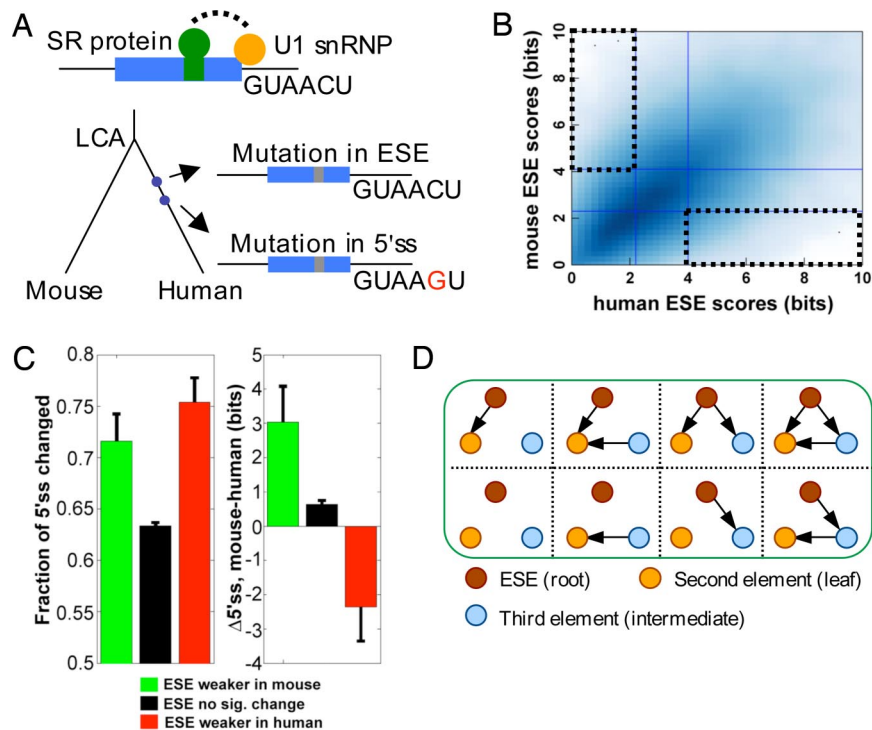


Fig. 1. The PVSA method for inference of coevolutionary networks. (A) Evolution of a hypothetical exon (blue box) is shown. After the last common ancestor (LCA) of human and mouse, two mutations occur on the human lineage, the first disrupting an ESE, the second strengthening the 5'ss. (B) The PVS for ESE (dotted boxes) is defined as the set of exons whose ESE strength is in the upper 33% of one genome but in the lower 33% of the other genome. (C) Frequency of changes to 5'ss sequences (*Left*) and changes in 5'ss score (*Right*) (mean \pm SE) in exons belonging to two subsets of the PVS-ESE (green, ESEs weaker in mouse; red, ESEs weaker in human) and to non-PVS-ESE exons (black). (D) Definition of subnetworks in the PVSA method. The subnetworks are rooted directed-acyclic graphs (e.g., with ESE as the root node). For simplicity, only subnetworks with a fixed leaf node (no outgoing edges) are listed.

among the major categories of splicing *cis* elements (5'ss, 3'ss, ESE, ESS, etc.). For example, two elements *X* and *Y* (e.g., the 5'ss and 3'ss of exons) might have a “compensatory” relationship, with *Y* strengthening when *X* weakens, and vice versa. Alternatively, *X* and *Y* might have a “correlated” relationship, with strengthening of *X* associated with strengthening of *Y*, or have no direct relationship if changes in *X* and *Y* occur independently of one another. The PVSA approach considers network rather than pairwise relationships among splicing *cis* elements, with each node representing a specific category of *cis* element (5'ss, 3'ss, ESE, ESS, etc.), and with interactions between elements represented by directed edges between nodes. In PVSA, a subset of data called the principal variation subset for *X* (PVS, or PVS-*X*) is defined for each *cis* element *X*, by identifying exons or introns in which the strength of *X* changed dramatically between orthologous exons or introns in the genomes under study (Fig. 1B), using log-odds scoring of *cis* elements [[supporting information \(SI\) Text](#)]. Such large changes can be thought of as perturbations to the splicing network, analogous to the experimental perturbations (such as administration of a specific kinase inhibitor) typically used to infer directed edges in signaling networks (18, 23).

When such a large change has occurred, changes in other elements are often observed. For example, in human/mouse changes to 5'ss sequences occurred more often in exons belonging to the PVS for ESEs than in other exons, with the 5'ss stronger on average in the organism in which ESEs were substantially weaker (Fig. 1C). This pattern suggested that 5'ss strengthening can and often does compensate for ESE-disrupting mutations. However, consideration of additional relevant variables in a network model is essential to avoid potential artifacts. For example, the scenario in which *Y* compensates *X*,

and *Z* compensates *Y* could lead a pairwise model to detect a spurious relationship between *X* and *Z* that could be avoided by inclusion of all three variables in a network. A simple pairwise analysis may also fail to detect significant relationships, e.g., in cases where one variable is compensated by several others.

In the PVSA method, patterns of change are analyzed by applying a Bayesian network inference procedure to the PVS for each element (see *Materials and Methods*). For each *cis* element *X*, the PVS-*X* is used to evaluate possible directed acyclic subnetworks that involve *X* as a root node (Fig. 1D and [SI Fig. 5](#)). Those subnetworks identified as significant (i.e., which have high likelihood of generating the observed patterns of association between the variables relative to controls) are then cascaded together to generate a final network that summarizes the relationships among the nodes (but which is not required to satisfy the acyclicity constraint of Bayesian networks). In this final network, an arrow $X \rightarrow Y$ indicates that large changes in element *X* were associated with significant changes in the strength of *Y*, with the color of the arrow indicating whether the changes tended to be in a compensatory (green) or correlated (red) direction (see Fig. 2 and *SI Text*).

Widespread Constraints on Splicing Element Evolution Associated with Exon and/or Intron Definition. Initially, the coevolution of the *cis* elements involved in splicing of constitutive exons and introns was studied in an evolutionarily diverse group of eukaryotes ([SI Table 1](#)). The 3'ss and 5'ss of pairs of consecutive constitutive exons (i.e., a total of four splice site variables) were considered. The length of the intervening intron was considered as a fifth variable (except in plants, where the sample size was insufficient), because intron length is known to influence splicing in a number of organisms (24, 25); exon length did not vary suffi-

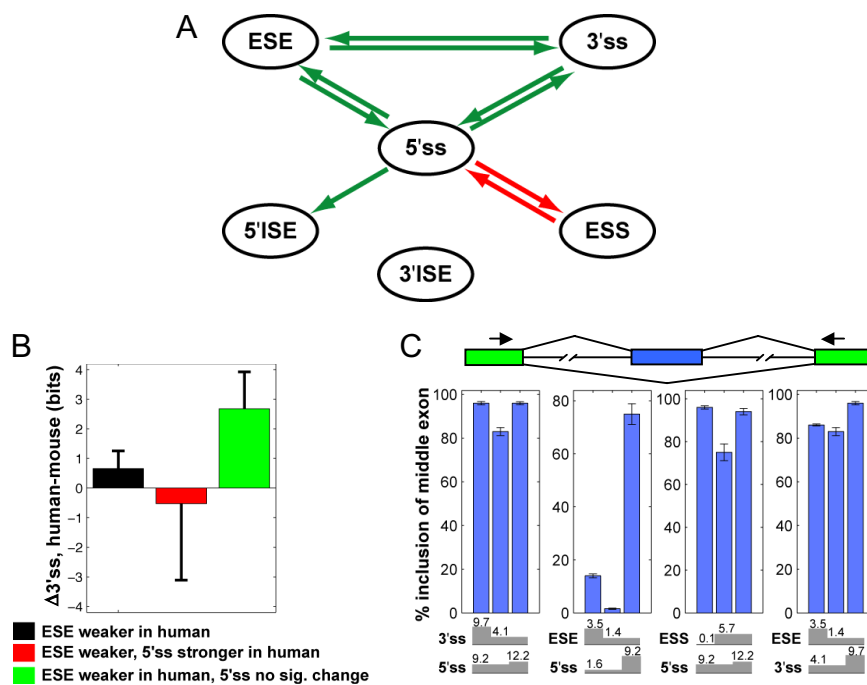


Fig. 3. SRE coevolutionary network for human and mouse. (A) The network identified by using the PVSA method based on six classes of splicing *cis* elements is shown. For all edges $P < 0.001$. Arrows are colored as in Fig. 2. (B) Change in 3'ss strength (mean \pm SE) for the set of PVS-ESE exons with weaker ESEs in human than mouse (black), and for subsets created by intersecting this set with the subset of the PVS-5'ss having weaker 5'ss in mouse (red) or the non-PVS-5'ss set (green). (C) Tests of bidirectional compensatory relationships detected in A used constructs containing combinations of stronger and weaker 5'ss, 3'ss, ESE, and ESS (with log-odds scores indicated) in the second exon. Inclusion of the second exon was assayed by body-labeled RT-PCR of total RNA from 293T cells transfected with each reporter by using primers targeted to the first and third exons. Blue bars indicate mean, and error bars indicate range of duplicated experiments.

containing a number of highly significant relationships between *cis* element classes, including all of the analyzed elements except 3'ISEs (Fig. 3A; P values listed in SI Table 3).

Several features of the identified network are notable. First, the 5'ss appears to play a particularly central role, with compensatory or correlated relationships detected to four other elements in the network, whereas other elements had significant interactions with at most two other elements. This observation suggests that splicing of constitutive exons may be particularly sensitive to 5'ss strength, perhaps because 5'ss recognition by U1 snRNP occurs before other steps in spliceosome assembly (33, 34). Notably, all of the interactions detected are in the direction expected to result from selection acting to maintain a minimum exon "strength." Compensatory interactions (Fig. 3A, green) occur between the 5'ss and the three other positively acting elements (3'ss, ESE, and 5'ISE), whereas a correlated interaction (Fig. 3A, red) was detected between the 5'ss and negatively acting ESSs.

Another prominent feature of the network is the set of mutually compensatory relationships among all combinations of 5'ss, 3'ss, and ESEs. The nature of these three-way interactions and the interpretation of the final PVSA network is illustrated in Fig. 3B. No significant change in 3'ss strength was observed for the set of exons in which ESEs were substantially weaker in human than mouse. However, a large increase in 3'ss strength was observed in the subset of this set for which 5'ss strength changed little or not at all, and a modest decrease in 3'ss strength was seen in the subset where 5'ss strength increased. Thus, the compensatory relationship between ESEs and the 3'ss is only readily observed when 5'ss changes are controlled, underlining the complex and multivariate nature of *cis* element relationships in mammals.

Three of the four interactions involving the 5'ss, as well as the compensatory relationship between ESEs and the 3'ss, were bidirectional, suggesting "functional symmetry," with each ele-

ment in the pair commonly able to rescue weakness in the other. This conclusion was supported by splicing reporter assays. For each of the four pairs of elements with bidirectional compensation relationships identified in Fig. 3A, three constructs were tested, representing a reference exon subjected to two successive mutations, weakening one element, then strengthening the other (Fig. 3C and SI Fig. 8). The results confirmed that, in each case, exon skipping resulting from weakening of one element could be rescued completely or partially by strengthening of the second element for each of these four pairs. The magnitudes of the changes in exon inclusion tended to be larger when the changes in log-odds scores of the affected elements were larger, supporting the relevance of these scores to splicing activity. Some of these compensatory relationships have also been supported by previous studies (35, 36).

The unidirectional arrow between 5'ss and 5'ISE in Fig. 3A indicated that large decreases in 5'ss strength between human and mouse were associated with increased ISE density in the downstream intron, but not vice versa. This asymmetry could indicate that large changes in 5'ss strength are more likely to impact splicing than are large changes in 5'ISEs and/or that the activity of these ISEs is highly context-dependent.

Evolutionary Relationships of Individual SREs. In the network analysis of Fig. 3A, all ESE and ESS hexanucleotides derived from large screens were analyzed together. This approach therefore captures the average behavior of these classes of SREs, but could obscure any differences that may exist between ESE and ESS motifs recognized by different factors. Therefore, we also analyzed the evolutionary relationships among individual ESE and ESS motifs (Fig. 4, SI Fig. 9, and SI Tables 4–6). ESE motifs associated with specific SR proteins were obtained primarily from SELEX or functional SELEX experiments (10, 16, 37), and the ESS motifs were derived based on clustering of ESS oligo-

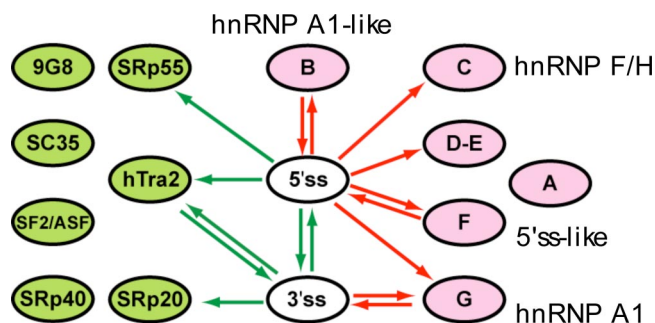


Fig. 4. Evolutionary relationship among individual groups of ESE and ESS motifs. The splicing interaction network among ESE and ESS motifs resembling binding sites of SR proteins and hnRNPs is shown. All edges $P < 0.001$. Arrows are colored as in Fig. 2.

nucleotides identified in a cell fluorescence-based screen (38), some of which correspond to known binding motifs of specific hnRNP proteins. To obtain the network shown in Fig. 4, subnetworks were first identified based on four nodes [the two splice sites and two regulatory element motifs (i.e., two ESE groups, two ESS groups or one ESE and one ESS group)]. Given the seven ESE groups and six ESS groups, a total of 78 four-node networks were analyzed and combined together to form the final network. Analyzing no more than two ESE or ESS groups at a time was essential to reduce the search space; this approach implicitly assumes that indirect interactions of two SRE motifs through a third SRE motif are negligible. A P value cutoff of 0.001 was used, yielding an expectation of less than one false positive network edge in this analysis.

The results of this analysis yielded a very clear pattern, with extensive compensatory relationships between three ESE motifs (corresponding to hTra2, SRp20, and SRp55) and the 5'ss and/or 3'ss, and correlated relationships between ESS motifs and splice sites, principally the 5'ss (Fig. 4). These results reinforce those seen in Fig. 3A, suggesting that gain/loss, respectively, of several different ESE/ESS motifs can compensate for mutations that weaken splice sites. However, it was notable that no significant compensatory or correlated relationships were observed between different ESE or ESS motifs. Previously, evidence of selection against changes from one ESE sequence to another in recent primate evolution was obtained from analyses of single-nucleotide polymorphism data (39), and selection to conserve "neighborhood-inference" score (reflecting high similarity to other known ESEs) was observed over longer periods of mammalian evolution (40). The analysis here differs in that overlapping SREs were excluded to avoid artifacts in network inference (*SI Text* and *SI Fig. 10*), whereas the previous studies considered only overlapping SREs. The absence of compensatory interactions between ESE motifs observed here suggests that loss/gain of specific ESEs is more often compensated by changes in splice site strength than by gain/loss of another ESE bound by a different factor. This observation suggests that the *in vivo* activities of different mammalian ESE motifs are not commonly interchangeable.

The ESS motifs analyzed showed more consistent behavior than the ESEs, with almost all having significant correlated relationships to the 5'ss. These observations suggested that there is a major evolutionary dynamic between the 5'ss and ESSs, with emergence of new ESSs in constitutive exons (e.g., arising by drift or driven by selection at the protein level for specific codons that happen to create ESSs) very often requiring strengthening of the 5'ss to restore efficient splicing. Conversely, exons weakened or disabled by 5'ss mutations are likely under strong selection to eliminate ESSs. For one ESS motif, group G, which contains the core hnRNP A1 consensus motif UAGG, a corre-

lated relationship with the 3'ss was also observed, which might indicate that A1 can commonly inhibit recognition of both splice sites (41). However, similar 3'ss correlations were not detected for any other ESS motifs, suggesting that most ESSs act primarily to inhibit productive association of factors with the 5'ss (consistent with Fig. 3A). Among ESEs, only the hTra2 motif had a compensatory relationship to both the 5'ss and 3'ss, suggesting that this element more commonly functions to enhance both splice sites than other ESEs.

Discussion

Splicing of constitutive exons represents a common evolutionary dynamic in which natural selection likely acts on a particular outcome (production of adequate amounts of accurately spliced mRNA) with little concern for how that outcome is achieved. Because splicing of a mammalian exon depends on recognition of several distinct *cis* elements, selection for efficient and accurate splicing can lead to complex dynamics in the evolution of splicing *cis* elements. Here, we have introduced a method, PVSA, that provides a framework for understanding coevolutionary networks of *cis*-regulatory elements. Coevolutionary networks inferred with PVSA are analogous to other network representations of biological systems such as protein interaction, signaling, and transcriptional regulatory networks (18, 42), in that they reflect functional relationships among proteins and/or nucleic acids, but PVSA requires no direct experimental data for network inference, instead using comparative genomic data. The PVSA method could, without difficulty, be applied to study the coevolution of transcriptional elements in core promoters or other evolutionary processes involving coordinated activity of diverse *cis*-regulatory elements.

The coevolutionary networks inferred for splicing in diverse eukaryotic taxa suggest that intron definition and exon definition models, sometimes regarded as abstractions, play fundamental roles in shaping gene evolution (5, 8). Incorporation of SREs into the splicing networks in mammals yielded a set of relationships describing which elements tend to compensate for mutations that weaken other elements. These rules may aid in determining which regions should be preferentially targeted in repairing mutations that contribute to disease through disruption of splicing (21). For example, the results of Fig. 4 suggest that exon skipping resulting from 5'ss mutations (a very common type of human disease mutation) in exons containing any of several ESS motifs may be commonly rescuable by inhibition of ESS recognition or other activities of specific splicing repressors.

The observation of common evolutionary compensation between ESEs and splice sites in mammals is also consistent with a recently proposed model for splicing evolution in which the emergence of SR proteins allowed evolutionary diversification of splicing signals across organisms and within organisms. This diversification, in turn, allowed versatility and flexibility in splice site recognition, enabling alternative splicing, and perhaps also increased robustness of constitutive exons to mutations (43, 44). The frequent covariation between exonic SREs and splice sites suggests that the constraints of accurate and efficient splicing may pose an effective speed limit on the rate of amino acid divergence in exons, with evolution of certain codon changes requiring prior or immediately subsequent splice site strengthening.

Materials and Methods

Exon and Intron Data Sets. Constitutive human exons and introns were obtained from the HOLLYWOOD database (30). Orthologous mouse exons were identified by using multigenome alignment provided by the University of California Santa Cruz genome browser (45). Further details of data set construction for mammals and other organisms are given in *SI Text*.

The PVSA Method. For all classes of *cis* elements, we evaluated their strength in the exons or introns of the two genomes under study (e.g., human and mouse). Specifically, 5'ss and 3'ss strength was scored by using a maximum entropy model (46). ESE, ESS and ISE elements were assigned log-odds scores reflecting their relative enrichment in exons and introns (*SI Text*). When included, the length of the intron was used directly as the “score” associated with the “intron length” node. For each *cis* element, the PVS consisted of human–mouse orthologous exons in which the score was in the lowest 33% of all exons in one genome and in the upper 33% in the other genome (Fig. 1*B*). The 33% cutoff was chosen by considering the tradeoff between the size of PVSs and the extent of variation in the *cis* elements. Use of different cutoffs within a reasonable range (e.g., 25%) yielded the same network structures.

To identify the subnetwork associated with the PVS for element *X*, all possible network structures with *X* as the root node were constructed (see Fig. 1*D* for examples of three-node rooted subnetworks with a fixed leaf). In this study, only network structures with a maximum depth of three were considered (i.e., one root node, arbitrary numbers of intermediate and leaf nodes, with no path longer than two edges). These subnetworks were then scored by using standard Bayesian inference methods (23), which calculate the likelihood of generating the observed data under a given Bayesian network structure. The high scored subnetworks (2% of subnetworks with highest likelihood) were averaged together to calculate a score for each edge between the

root and the leaf nodes. The edge score was defined as the fraction of times that this edge is included in the high-scored networks.

Subsequently, a shuffling procedure was performed to obtain a *P* value for the edges in the final subnetwork. For each *cis* element in each exon (or intron), a control exon (or intron) was chosen randomly from among those whose differences in G+C content between human and mouse orthologs were similar to that of the exon (or intron) under study. The scores of the *cis* element in the control exon (or intron) in human and mouse genomes were substituted for the scores of the test exon (or intron) in the two genomes, respectively. This procedure was carried out independently for each *cis* element. The same PVSA method was applied to the shuffled data set to derive an averaged control network. The *P* value of each edge in the final subnetwork was calculated as the fraction of times that this edge received a higher score in the shuffled data than in the real data. Further details are provided in *SI Text* and *SI Figs. 11 and 12*.

Cell Culture and Splicing Reporter Assays. 293T cells were cultured and transfected with splicing reporter constructs, and splicing was assayed by radio-labeled RT-PCR, as described in *SI Text*.

We thank D. Black and P. A. Sharp for comments on the manuscript and members of C.B.B.'s laboratory for helpful discussions. This work was supported by an Anna Fuller Fund postdoctoral fellowship (to X.X.) and grants from the National Institutes of Health and National Science Foundation (to C.B.B.).

1. Lim LP, Burge CB (2001) *Proc Natl Acad Sci USA* 98:11193–11198.
2. Black DL (2003) *Annu Rev Biochem* 72:291–336.
3. Matlin AJ, Clark F, Smith CW (2005) *Nat Rev Mol Cell Biol* 6:386–398.
4. Blencowe BJ (2000) *Trends Biochem Sci* 25:106–110.
5. Berget SM (1995) *J Biol Chem* 270:2411–2414.
6. Talerico M, Berget SM (1994) *Mol Cell Biol* 14:3434–3445.
7. Sterner DA, Carlo T, Berget SM (1996) *Proc Natl Acad Sci USA* 93:15081–15085.
8. Robberson BL, Cote GJ, Berget SM (1990) *Mol Cell Biol* 10:84–94.
9. Fu XD (1995) *RNA* 1:663–680.
10. Liu HX, Zhang M, Krainer AR (1998) *Genes Dev* 12:1998–2012.
11. Tacke R, Manley JL (1999) *Curr Opin Cell Biol* 11:358–362.
12. Eperon IC, Makarova OV, Mayeda A, Munroe SH, Caceres JF, Hayward DG, Krainer AR (2000) *Mol Cell Biol* 20:8303–8318.
13. Mayeda A, Krainer AR (1992) *Cell* 68:365–375.
14. Wang Z, Xiao X, Van Nostrand E, Burge CB (2006) *Mol Cell* 23:61–70.
15. Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) *Science* 297:1007–1013.
16. Schaal TD, Maniatis T (1999) *Mol Cell Biol* 19:1705–1719.
17. Friedman N, Linial M, Nachman I, Pe'er D (2000) *J Comput Biol* 7:601–620.
18. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) *Science* 308:523–529.
19. Cartegni L, Chew SL, Krainer AR (2002) *Nat Rev Genet* 3:285–298.
20. Pagani F, Baralle FE (2004) *Nat Rev Genet* 5:389–396.
21. Buratti E, Baralle M, Baralle FE (2006) *Nucleic Acids Res* 34:3494–3510.
22. Faustino NA, Cooper TA (2003) *Genes Dev* 17:419–437.
23. Pearl J (2000) *Causality: Models, Reasoning, and Inference* (Cambridge Univ Press, Cambridge, UK).
24. Bell MV, Cowper AE, Lefranc MP, Bell JL, Sreaton GR (1998) *Mol Cell Biol* 18:5930–5941.
25. Fox-Walsh KL, Dou Y, Lam BJ, Hung SP, Baldi PF, Hertel KJ (2005) *Proc Natl Acad Sci USA* 102:16176–16181.
26. Dewey CN, Rogozin IB, Koonin EV (2006) *BMC Genomics* 7:311–319.
27. Yeo G, Hoon S, Venkatesh B, Burge CB (2004) *Proc Natl Acad Sci USA* 101:15700–15705.
28. Kupfer DM, Drabentstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW (2004) *Eukaryot Cell* 3:1088–1100.
29. McCullough AJ, Baynton CE, Schuler MA (1996) *Plant Cell* 8:2295–2307.
30. Holste D, Huo G, Tung V, Burge CB (2006) *Nucleic Acids Res* 34:D56–D62.
31. Wang BB, Brendel V (2006) *Proc Natl Acad Sci USA* 103:7175–7180.
32. McCullough AJ, Berget SM (1997) *Mol Cell Biol* 17:4562–4571.
33. Das R, Dufu K, Romney B, Feldt M, Elenko M, Reed R (2006) *Genes Dev* 20:1100–1109.
34. Lacadie SA, Rosbash M (2005) *Mol Cell* 19:65–75.
35. Carothers AM, Urlaub G, Grunberger D, Chasin LA (1993) *Mol Cell Biol* 13:5085–5098.
36. Singh NN, Androphy EJ, Singh RN (2004) *RNA* 10:1291–1305.
37. Tacke R, Tohyama M, Ogawa S, Manley JL (1998) *Cell* 93:139–148.
38. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB (2004) *Cell* 119:831–845.
39. Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) *PLoS Biol* 2:E268.
40. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB (2006) *PLoS Genet* 2:e191.
41. Zhu J, Mayeda A, Krainer AR (2001) *Mol Cell* 8:1351–1361.
42. Wyrick JJ, Young RA (2002) *Curr Opin Genet Dev* 12:130–136.
43. Izquierdo JM, Valcarcel J (2006) *Genes Dev* 20:1679–1684.
44. Shen H, Green MR (2006) *Genes Dev* 20:1755–1765.
45. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. (2003) *Nucleic Acids Res* 31:51–54.
46. Yeo G, Burge CB (2004) *J Comput Biol* 11:377–394.