Supplementary Material

Part 1

Table A The number of genes, exons and introns for four species

	gene	exon	intron
C.elegans	185	1138	953
A.thaliana	749	4282	3533
D. melanogaster	1196	3722	2526
Human	1231	6835	5604

The table gives the number of genes, exons and introns for four species used in the splice site prediction in 10-fold cross validation.

Table B Numbers of genes, true and false splice sites in the data sets for four species

	subset	gene	True donor	True acceptor	False donor	False acceptor
C.elegans	1	60	319	319	17144	11940
	2	60	328	328	19410	14118
	3	65	306	306	15070	10584
A.thaliana	1	246	1251	1251	47315	30206
	2	252	1201	1201	49494	32373
	3	251	1081	1081	45041	28946
D. melanogaster	1	388	831	831	73279	38529
	2	397	830	830	78395	39928
	3	411	865	865	77670	40174
Human	1	399	1843	1843	233681	159645
	2	408	1841	1841	295501	196136
	3	424	1920	1920	236109	155552

The table gives the numbers of genes, true and false splice sites in the data sets for four species used in the acceptor and donor site detection on three disjoint subsets.

Part 2

How to make splice site prediction and exon/intron identification?

The procedure in training set includes:

(1.1) Calculate ξ for each potential splice site following Eq.(9).

(1.2) For each species set a value for ξ_D and find all candidate donors with $\xi > \xi_D$, and set a value for ξ_A and find all candidate acceptors with $\xi > \xi_A$.

(1.3) In each gene label each candidate splice site *D* (GT) or *A* (AG) along the sequence by its ξ value.

(1.4) In a gene of *n* candidate sites we divide them into several successive regions that each region (possibly except the first and the last one) initiates from a *D* (say D_j) and ends by a *A* (say A_k). A region is called irreducible if it cannot be further divided. Evidently, only one irreducible division exists for any *n*-symbol sequence written by *D* and *A*. For a given region we predict the first *D* in sequence with positive ξ (if existed) or the *D* with the largest ξ （if all *D's* in the region have ξ smaller than or equal 0) as the donor, and predict the *A* with the largest ξ as the acceptor. (see following examples)

(1.5) Identify all introns and exons based on the assignment of donors and acceptors. Note that the first exon of a gene initiates from ATG and the last exon terminates at stop codon. The initiator and terminator have been given in database (see following examples).

(1.6) Find *Sn*, *Sp*, *Ac*(e), *Ac*(o) and *Ac*(all) through the comparison between predicted and true introns and exons following Eq.(10) and (11).

(1.7) Change the setting of ξ_{D} and ξ_{A} and repeat the steps from (1.2) to (1.6), then find the best-fit ξ_{D} and ξ_{A} through the maximization of *Ac*(all) and (*Sn*+*Sp*)/2 respectively.

(1.8) List all best-fit ξ_{D}'s and ξ_{A}'s and the corresponding values of *Ac*(all) or (*Sn*+*Sp*)/2 in ten computations in 10-fold cross validation. (namely, Table C1 to C4).

(1.9) Compare the results of 10 computations and find the optimal ξ_{D} and ξ_{A}.

The procedure in test set includes:

(2.1) The same as (1.1).

(2.2) By use of the optimal ξ_{D} and ξ_{A} find all candidate donors with $\xi > \xi_{\mathrm{D}}$, and all candidate acceptors with $\xi > \xi_{\mathrm{A}}$.

(2.3) The same as (1.3).

(2.4) The same as (1.4).

(2.5) The same as (1.5).

(2.6) The same as (1.6).

(2.7) Average accuracy parameters *Sn*, *Sp*, *Ac*(e), *Ac*(o) and *Ac*(all) over 10 computations (in 10-fold cross validation) and list the result in a table (namely, Table 2 in text).

(2.8) Change the setting of ξ_{D} and ξ_{A} and repeat the steps from (1.2) to (1.6), and study how the accuracy parameters change with ξ_{D} and ξ_{A}.

(2.9) List the variation of accuracy parameters in some given range of ξ_{D} and ξ_{A}. (Table D)

Table C1 The best fit values of parameters ζ_D and ζ_A in 10 computations and corresponding accuracies of prediction for splice sites in training set (*C. elegans*)

	1	2	3	4	5	6	7	8	9	10
ζ_D	-10	-10	-9	-10	-10	-10	-9	-10	-10	-10
ζ_A	-4	-4	-3	-3	-3	-4	-3	-4	-2	-4
Ac(all)	97.6	98.0	98.2	97.9	97.6	98.3	97.8	97.9	98.3	98.0

	1	2	3	4	5	6	7	8	9	10
ζ_D	-10	-10	-9	-9	-10	-10	-9	-10	-10	-10
ζ_A	-4	-4	-3	-5	-5	-4	-4	-4	-4	-4
(*Sn*+*Sp*)/2	97.2	97.3	97.4	97.5	97.0	97.3	97.0	97.0	97.4	97.4

Table C2 The best fit values of parameters ζ_D and ζ_A in 10 computations and corresponding accuracies of prediction for splice sites in training set (*A. thaliana*)

	1	2	3	4	5	6	7	8	9	10
ζ_D	-5	-5	-6	-5	-5	-5	-4	-4	-5	-5
ζ_A	-4	-5	-4	-4	-5	-5	-5	-4	-5	-5
Ac(all)	98.2	98.2	98.1	98.2	98.3	98.3	98.2	98.1	98.2	98.2

	1	2	3	4	5	6	7	8	9	10
ζ_D	-5	-5	-7	-5	-5	-5	-4	-8	-5	-5
ζ_A	-5	-5	-4	-5	-5	-5	-5	-4	-5	-5
(*Sn*+*Sp*)/2	94.7	94.6	94.5	94.8	94.8	94.8	94.4	94.3	94.5	94.6

Table C3 The best fit values of parameters ζ_D and ζ_A in 10 computations and corresponding accuracies of prediction for splice sites in training set (*D. melanogaster*)

	1	2	3	4	5	6	7	8	9	10
ζ_D	-3	-5	-4	-4	-3	-4	-5	-4	-5	-5
ζ_A	-5	-3	-4	-5	-5	-5	-5	-5	-5	-4
Ac(all)	97.3	97.3	97.3	97.6	97.6	97.3	97.3	97.4	97.5	97.3

	1	2	3	4	5	6	7	8	9	10
ζ_D	-3	-3	-3	-3	-3	-4	-5	-3	-4	-3
ζ_A	-5	-5	-5	-5	-5	-5	-4	-5	-5	-5
(*Sn*+*Sp*)/2	97.0	97.2	97.2	97.1	97.2	97.1	97.1	97.2	97.2	97.2

Table C4 The best fit values of parameters ζ_D and ζ_A in 10 computations and corresponding accuracies of prediction for splice sites in training set (human)

	1	2	3	4	5	6	7	8	9	10
ζ_D	-2	-2	-2	-4	-2	-6	-4	-2	-4	-2
ζ_A	-1	0	0	0	0	0	0	0	0	0
Ac(all)	94.3	94.1	94.5	94.4	94.3	94.3	94.3	94.5	94.1	94.2

	1	2	3	4	5	6	7	8	9	10
ζ_D	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4
ζ_A	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
(*Sn*+*Sp*)/2	88.8	88.9	89.0	88.7	88.9	88.8	88.6	88.9	88.6	88.8

In above Table C1 to C4 the optimal ζ_D and ζ_A with respect to *Ac*(all) and with respect to the average of *Sn* and *Sp* , (*Sn*+*Sp*)/2, in 10 computations are listed.

Table D The variation of prediction accuracy for splice sites in test set with parameter ζ_D in a range from –10 to 0 and ζ_A from –5 to 0

	range of *Sn* (%)	range of *Sp* (%)	range of *Ac*(all) (%)
C. elegans	(88.5, 94.8)	(96.3, 97.7)	(95.2, 97.1)
A. thaliana	(86.3, 93.6)	(93.1, 96.0)	(94.5, 97.7)
D. melanogaster	(92.3, 96.4)	(95.4, 98.1)	(94.6, 96.9)
Human (1)	(82.4, 89.8)	(71.7, 91.9)	(88.0, 93.9)
Human (2)	(82.4, 89.1)	(80.6, 91.9)	(91.6, 93.9)

For *C. elegans* , *A. thaliana, D. melanogaster* and Human (1) ζ_D changes between (-10,0) and ζ_A changes between (-5,0); for Human (2) ζ_D changes between (-10,0) but ζ_A changes between (-3,0).

Example 1: > 9084_CELGES1B protein_id:AAA28057.1; Caenorhabditis elegans gut esterase (ges1) gene, complete cds.

candidate donor		candidate acceptor		predicted intron sites	true intron sites
site	ζ_D	site	ζ_A		
		100	-4.44	163 — 228	163 — 228
163	6.46				
168	0.21				
		204	-1.91		
		228	-0.62		
524	6.81			524 — 1552	524 — 1542
528	3.01				
623	-9.69				
639	-3.97				
988	0.08				
996	-8.81				
		1542	1.52		
		1552	4.78		
		1902	-2.23		
2045	-4.21			2056 — 2719	2056 — 2719
2048	-1.01				
2056	12.58				
2074	-5.55				
2077	-0.79				
2458	-4.45				
2672	-1.99				
		2719	23.46		
2953	9.23			2953 — 3072	2953 — 3072
3031	-7.56				
		3072	12.18		
3215	4.66			3215 — 3299	3215 — 3254
		3299	-3.22		
3365	-2.66			3388 — 4156	3388 — 4156
3388	9.09				
3395	-8.67				
3523	-7.85				
3547	1.82				
3889	4.71				
3902	-5.67				
		4156	26.17		
4255	-0.56			4267 — 4330	4267 — 4330
4267	8.55				
		4330	14.33		

Example2: > 2372_CECED4A protein_id:CAA48781.1; C.elegans ced-4 gene

candidate donor		**candidate acceptor**		**predicted intron sites**	**true intron sites**
site	ζ_D	site	ζ_A		
		91	-2.40		
443	-4.57			452 — 502	452 — 502
452	6.41				
		502	8.63		
597	6.34			597 — 640	597 — 640
604	-3.14				
		640	7.37		
731	2.85			731 — 1033	731 — 916
735	-1.49				984 — 1033
750	-7.89				
792	-0.71				
984	11.54				
		1033	9.98		
		1238	-0.17		
1372	6.76			1372 — 2134	1372 — 1928
1650	2.35				
		2134	5.95		
2205	4.25			2205 — 2252	2205 — 2252
		2252	20.76		2359 — 2404
		2270	3.29		
		2404	20.44		
		2462	-3.40		
		2513	-2.60		