

Supplemental Material – Zhang, M., and W. Gish (2005)

Three analysis methods in EXALIN

EXALIN could be used in three basic ways that differed in their accuracy and speed.

Method 1: Full dynamic programming with splice site information (EXALIN-DPS)

EXALIN-DPS was chosen to be the default method used by EXALIN, as it made minimal use of heuristics and was guaranteed to provide an optimal answer (including the complete optimal alignment), its behavior was governed by the fewest number of parameters, and only two user-supplied inputs were required: a transcript sequence and a genomic sequence. The practical weakness of this approach was that it was the slowest method and its memory requirements could be prohibitive. The initialization and recursion equations are provided in Methods and included a comparison of $M(i,j)$ to 0. The traceback procedure to produce an alignment was initiated from the highest-scoring point in state space, which yielded locally optimal alignments. For a single-pass approach to the alignment, the memory in bytes required for traceback data was equal to the product of the lengths of the two sequences.

When free memory was determined in advance by the program to be unable to accommodate traceback information for the entire search space, a three-pass procedure was automatically employed. The first and second passes identified the start and end points of best-scoring alignments by full DP in the forward and reverse directions, with no voluminous traceback information recorded at this time. The memory required for these two passes was merely a linear function of the lengths of the two sequences. The third pass then performed DP restricted to the region defined by these start and end points, this time with traceback information maintained. When long genomic contig sequences were involved, the three-pass procedure generally decreased memory requirements below 512 MB for human ESTs aligned to human genes and did not severely degrade speed. If an approximate location for an mRNA was known in advance, perhaps through a preliminary BLASTN search, the dynamic programming could be optionally constrained to the relevant segment of the genomic sequence. Reducing the area explored by the first two passes then increased speed proportionately.

Method 2: BLASTN-guided dynamic programming (EXALIN-BLAST)

Three user-supplied inputs were required when the optional EXALIN-BLAST method was invoked. In addition to the usual genomic sequence and transcript sequence, an output file from BLASTN was also utilized. For this method, BLASTN would be executed in a prior step to locate similar regions between the transcript and genomic sequences. To obtain the precise input format for BLAST data to be read by EXALIN, a BLAST parser (written in PERL) reformatted the BLASTN output. The parsed result provided the start and end coordinates of HSPs (high-scoring segment pairs; Altschul, et al., 1990) in both the genomic sequence and spliced sequences, which were used to guide the dynamic programming by EXALIN.

Initiation and termination (“anchor”) points for DP were localized a distance g (default value 40 and user-modifiable with the $-g$ parameter) from the start and end of each HSP. If an HSP was shorter than g , the distal end was used as the anchor. Except for the 5'- and 3'-terminal HSPs, the method used a global approach to DP, in which $M(i,j)$ was not compared to 0 in the recursion (Needleman and Wunsch, 1970). For most regions of DP, traceback began at the DP termination point. For the 3'-terminal HSP, traceback began at the highest scoring point in state space. EXALIN-BLAST only reported alignment information in the vicinity of its splice junction predictions, which could limit its utility. Drawbacks to the approach were that the path of the MLL spliced alignment identified by EXALIN-DPS might not be accurately traversed here; and the overall alignment score was inaccurate because it only examined regions local to splice junctions and could double count some portion of the similarity score for exons shorter than $2g$. For same-species alignments, EXALIN-BLAST identified splice junctions only marginally less accurately than EXALIN-DPS (data not shown). Besides its speed, an advantage of the method was that the multiple “topcombo” group output of BLASTN could be utilized: when multiple non-intersecting sets of consistent HSPs were reported by BLASTN, as a potential consequence of

paralogous gene duplications or pseudogenes being present in the genomic template, EXALIN-BLAST could provide a spliced alignment for each group (Fig. 3).

Two principal classes of outcomes were possible in the BLASTN search results, with different dynamic programming strategies taken to address each one. First, several nucleotides of the transcript might be missed between two adjacent HSPs or the ends of two adjacent HSPs might overlap. In this case, only the region encompassing the gap or overlap was explored by DP (Figure 2A). Three subclasses of outcomes could result from this: dynamic programming might coalesce the two HSPs into one unspliced alignment (not shown); a splice site might be found in the originally missed nucleotides (Figure 2.A.a) or in the overlap (Figures 2.A.b); or additional exon(s) might be found (Figure 2.A.c). If nucleotides from the transcript could not be accurately mapped to the genomic sequence, no satisfactory high-scoring alignment was likely to be found by EXALIN-BLAST, in which case the program would emit an error message and return an easily testable, non-zero exit status. In such cases, increasing the bounds of the dynamic programming (EXALIN-BLAST command line options $-f$ and $-g$) might allow a satisfactory alignment to be found, but the likelihood of this was not evaluated.

In the second class of outcomes from BLASTN, one or both terminal segments of the transcript sequence might not be aligned. Terminal exons might have been missed completely in a heuristic BLASTN search if their identity was too low, or they might have gone unreported by the program (even if found) if their score was below the cutoff. In such cases, the strategy implemented in EXALIN-BLAST was to establish an anchor point at a pair of matching residues located several nucleotides internal to the nearest reported HSP and extend the dynamic programming from this point to well beyond the end of the HSP in the genome. Many different outcomes were possible from the extended DP. Sometimes the original HSP would not be extended beyond its original boundaries (Figure 2.B.a). When the alignment could be extended, though, it might have extended the original HSP (Figure 2.B.b) or been ascribed to one (or more) new exons (Figure 2.B.c).

The distance, d , extended by EXALIN-BLAST along the genomic sequence in our study was a function of u , the length of the unmapped terminal segment of the transcript, where $d = \min(50000, 0.01 * 4^u)$. The minimum upper limit of 50,000 seemed a reasonable compromise between speed and sensitivity for sequences of human origin. For other species, a lower limit may be more appropriate and would lead to shorter execution times. The factor of 0.01 would roughly ensure that any alignment of length u would not be expected by chance in the extended region. EXALIN-BLAST could fail to map terminal segments of transcripts due to extensions not proceeding far enough, due to significant polymorphic differences or rearrangements existing between the haplotypes represented by the transcript and genomic sequences, or due to cloning, sequencing or assembly errors. When one or both transcript termini were not found, EXALIN-BLAST nevertheless reported the portions it was able to align, which often included multiple exons.

An advantage of the EXALIN-BLAST method was that it could utilize multiple group output: multiple alignments of a transcript to different loci on the same genomic sequence, as a potential consequence of paralogous gene duplications or pseudogenes being present in the genomic template (Figure 3). For such situations, BLASTN can optionally order the HSPs by strand and position, and then cluster the HSPs into internally consistent and non-intersecting sets (so-called “topcombo groups”). When more than one group was reported by BLASTN, EXALIN-BLAST processed each group separately and reported an optimized alignment for each. Use of groups beyond the first may be desirable for reducing the risk of missing the native alignment when the BLASTN search or its clustering heuristics may have faltered. In the case of an unfinished genomic sequence template, for example, the native locus for the transcript might have been partial and ranked lower in BLASTN output than a paralogous gene copy that was more complete or than a pseudogene that lacked any introns at all. By post-processing EXALIN-BLAST output for exons or complete spliced alignments that exhibited a minimum level of identity, the native locus might be more accurately identified as a lower scoring alignment. When BLASTN reported multiple groups, we saw some evidence that the EXALIN-BLAST result for the first group was more likely to be partially in error. We suggest that EXALIN-DPS should perhaps be

used instead in these cases. While EXALIN-DPS was potentially vulnerable to the same occurrences of paralogs and pseudogenes, this full dynamic programming method was guaranteed to find shorter or less-conserved exons and, consequently, did not seem as easily sidetracked by similarity (spurious or otherwise) to non-native regions.

Method 3: Full dynamic programming with a simplified splice site model (EXALIN-SSM)

A third method was implemented in which splice sites were scored similarly to EST_GENOME. A small penalty was applied for canonical GT..AG splice sites and a severe penalty was applied for non-canonical sites.

Additional Features

Most EXALIN parameters governing scoring and heuristics were user-modifiable, allowing the program to be targeted to a wide variety of conditions, including different sequencing error rates, mutation rates and expected gene lengths. While the program by default used the simple (+5,-11,-11,-11) scoring system, and each of the component scoring parameters could be altered on the command line, the program would also accept a fully-specified scoring matrix read from a file—a feature perhaps most advantageous in cross-species comparisons.

By default, EXALIN compared the transcript sequence on both + and - (reverse complemented) strands to the genomic sequence and reported the best result of all. Due to data handling errors, sequences may be provided in an unconventional 3'→5' orientation. To accommodate such reversed transcripts directly, EXALIN optionally (-s1) reversed the splice models before analyzing the sequences. For cases where the forward/reverse orientation of a transcript was unknown, EXALIN could optionally (-s2) utilize the splice model in *both* forward and reverse directions and either report the results from both orientations or just report the best result of all. To accelerate this more extensive search with both forward and reversed splice models, a heuristic employed by EST_GENOME was also implemented in EXALIN: only the strand of the transcript yielding the best scoring alignment with the forward splice model was re-aligned using the reversed splice model. Alignments to the conventional orientation were indicated in the output by “SPLICE-DIR: +”, whereas alignments using the reversed, non-complemented splice site model were indicated by “SPLICE-DIR: -”.

In reporting sequence coordinates, EXALIN assigned the same coordinate numbers to complementary residues from the plus and minus strands of each sequence. This is the convention used by BLAST, but it differs from Sim4 and EST_GENOME, two programs which re-number residues on the complementary strand starting from its 5' end.

Internally EXALIN used 4-byte integer values in its DP recursion but accepted fractional values for scoring parameters. To maintain precision, all scores (the similarity scoring system, gap penalties, and splice site model scores) were scaled up by a factor of 100 and rounded to the nearest integer prior to the DP steps. The maximal scores obtained by DP were then scaled down by the same factor prior to output.

Micro-exon Prediction

Unrelated to the data set of Volfovsky *et al.* (2003) is the example shown in Figure 4 of an EXALIN-produced alignment containing a 7 nucleotide micro-exon. The EST sequence in this case exhibits a common pattern of sparsely distributed ambiguity codes throughout most of its length, with a higher concentration of Ns near the 3' end. All intron predictions are flanked by canonical GT..AG splice sites. The micro-exon was not reported by EST_GENOME, Sim4 or Spidey. As for the 3'-terminal exon, Sim4 did not detect it at all. Spidey reported the last exon only to position 543 of the transcript, omitting the last 53 nucleotides. Using BLAT with our default parameters (see Methods), the micro-exon was not detected. If just a single seed match was required (-minMatch=1), BLAT then detected the first two exons, including the micro-exon, but truncated the last exon at position 543. When the -fine option was further added, the micro-exon was no longer detected and the alignment was truncated even sooner, at position 508.

Cross-species alignment

For a given ortholog pair, the mapping of sequences from the Makalowski and Boguski (1998) data set to contemporary RefSeq entries was deemed successful only if all of the following criteria were satisfied: 1) both members of the pair could be mapped to RefSeq entries by BLASTN; 2) both members aligned to just a single locus in their native genomic sequence; and 3) the native RefSeq-genome alignments exhibited >99% identity and >97% coverage of the transcripts. In all, 219 orthologous pairs of transcripts were successfully mapped to RefSeq entries. The locations of coding regions and exon boundaries were then extracted from the RefSeq annotation and donor/acceptor splice junction pairs were collected just from the introns located internal to coding regions. A total of 2,354 splice junction pairs were so obtained from the human data and 2,383 pairs from the mouse.

Exon counts and exon and intron length distributions

We compared the performance of the programs in other respects, as well. For the aforementioned set of 16,639 human ESTs aligned to the human genome, each of the programs predicted different numbers of exons and produced different length distributions for exons and introns. (Tables 8-10). This further illustrates that the introduction of sequencing errors, polymorphisms and incompleteness of sequences can lead to disparities between the different programs and, consequently, ambiguity in the results.

Figure 2. BLAST-guided Dynamic Programming. Dotted rectangles enclose the areas explored by DP with EXALIN-BLAST.

A. When BLASTN missed a small fragment of EST sequence or two HSPs overlapped, part of two flanking HSPs in EST sequence and the relevant portion of the genomic sequence were used in dynamic programming.

a. Missed part is distributed among two HSPs. Two thickest lines are original HSPs found by BLASTN.

b. Overlap part is assigned to two HSPs and splice site is found. The thick lines are the final result predicted by EXALIN.

c. Find another small HSP. Two thickest lines are HSPs found by BLASTN. The middle line is the new exon found by EXALIN.

B. When 5'- or 3'-terminal portions of the transcript sequence were not mapped by BLASTN, the missed part plus adjacent tens of nucleotides in first HSP were aligned with an extension portion of the genomic sequence.

a. Missed portion not found.

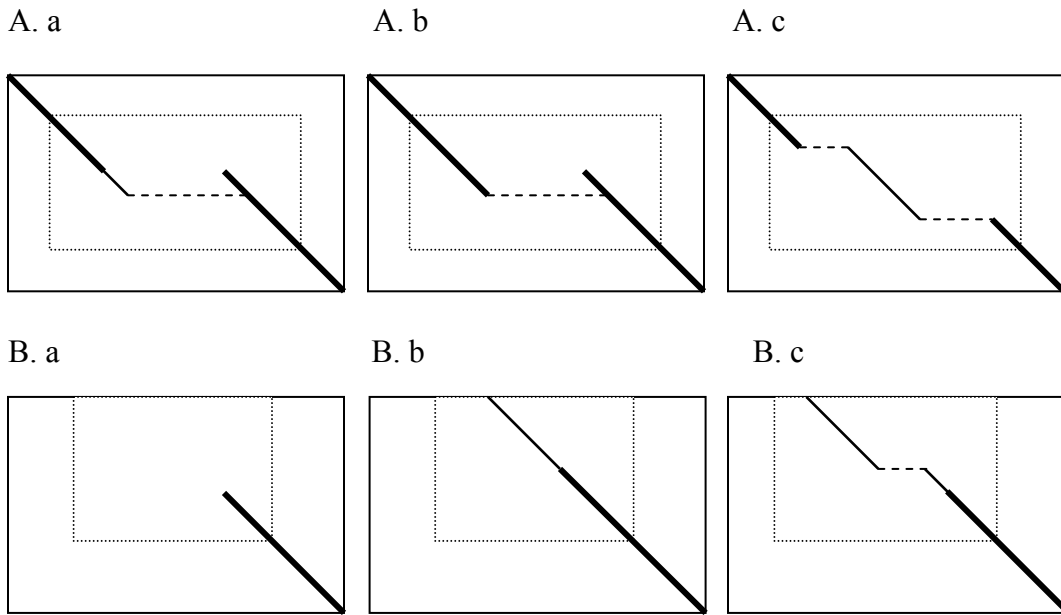
b. Unaligned segment is just an extension of the original HSP.

c. Another exon is found.

Figure 3. BLAST alignment group output. The same transcript aligns well to two genomic loci. HSPs in the same “topcombo” group from BLASTN, which are consistently oriented along both sequences, would be processed separately by EXALIN-BLAST to produce two optimal spliced alignments.

Figure 4. EXALIN-DPS output for an example containing a micro-exon and all canonical GT..AG splice sites.

Figure 2.



Dashed rectangle: Area of exploration by dynamic programming
Thick black line: HSPs originally identified by BLASTN
Thin black line: Exon or part of exon identified by dynamic programming
Dashed line: Intron

Fig. 3

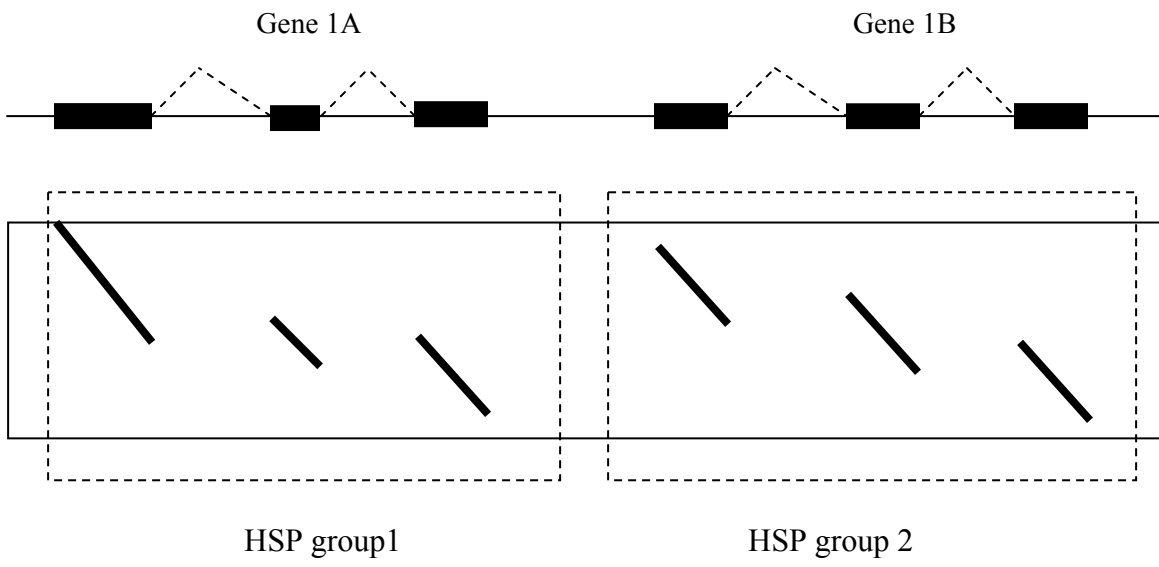


Fig. 4

EXALIN [28-Mar-2005]
QUERY: gi|1766096|gb|AA182927.1|AA182927 zp36c07.r1 Stratagene muscle 937209
Homo sapiens cDNA clone IMAGE:611532 5' similar to gb:J04760 TROPONIN I, SLOW
SKELETAL MUSCLE (HUMAN);, mRNA sequence.
SBJCT: H_KvLQT1 : renamed from Hs chr11.(669054-1948475) length=1279422
SCORE: 2060.0
SPLICE-DIR: GT-AG
QUERY-LENGTH: 596
SBJCT-LENGTH: 1279422
STRAND: +

EXON: 1 - 21 (76027 - 76047)	105.0	100.00%
EXON: 22 - 28 (76192 - 76198)	35.0	100.00%
EXON: 29 - 70 (76746 - 76787)	156.0	90.91%
EXON: 71 - 196 (76871 - 76999)	495.0	92.37%
EXON: 197 - 285 (77162 - 77251)	434.0	98.89%
EXON: 286 - 461 (77374 - 77550)	751.0	94.97%
EXON: 462 - 596 (77799 - 77941)	261.0	80.00%

Q: TCAGGACCTCAGGATGGGAGA..... TGAGGAGAAGCGG	34
>>>>...>>>> >>>>...>>>>	
G: TCAGGACCTCAGGATGGGAGAGTAAG...GACAG TGAGGAG GTAAG...TGCAGAAGCGG	76751
Q: AACAGGGCCATCACGG--CGCNANGGCAGCACCTGAAG.....AGTGTGATG	79
>>>>...>>>>	
G: AACAGGGCCATCACGGCCCGC-A-GGCAGCACCTGAAGGTAGG...CGCAGAGTGTGATG	76879
Q: CTGCAGATAGCGCCACGGAGCTGGAGAAGGAGGAGAGCCCGCTGANG-A-NGAAGCAG	137
>>>>...>>>>	
G: CTGCAGATAGCGCCACGGAGCTGGAGAAGGAGGAGAGCCCGCTGAGGCAGAGAAGCAG	76939
Q: AACTACCTGGCGGNA-C-NTGCCCG-CGCNTGCATATCCCGGGCTCCATGTCTGAAGTGC	194
>>>>...>>>>	
G: AACTACCTGGCGG-AGCACTGCCCGCCGC-TGCATATCCCGGGCTCCATGTCTGAAGTGC	76997
Q: AG.....GAGCTCT-CAAACAGCTGCACGCCAAGATCGATGCGGCTGAAGAG	240
>>>>...>>>> >>>>...>>>>	
G: AGGTACC...CCTAGGAGCTTGCAAACAGCTGCACGCCAAGATCGATGCGGCTGAAGAG	77206
Q: GAGAAGTACGACATGGAGGTGAGGGTGCAGAAGACCAGCAAGGAG.....CT	287
>>>>...>>>> >>>>...>>>>	
G: GAGAAGTACGACATGGAGGTGAGGGTGCAGAAGACCAGCAAGGAGGTGAG...CACAGCT	77375
Q: GGAGGACATGAACCAGAAGCTATTTGATCTGCGGGCAAGTTCAAGCGGCCCCACTGCG	347
>>>>...>>>> >>>>...>>>>	
G: GGAGGACATGAACCAGAAGCTATTTGATCTGCGGGCAAGTTCAAGCGGCCCCACTGCG	77435
Q: GAGGGTGCATGTGCGACCGATGCCATGCTCAAGGCCCTGCT--GCTTCGAAGCACAAGG	405
>>>>...>>>> >>>>...>>>>	
G: GAGGGTGCATGTGCGCCGATGCCATGCTCAAGGCCCTGCTGGGC-TCGAAGCACAAGG	77494
Q: TGTTTCATGGACCTGANGG-CAACCTGAAGCAGGTTCAAGAAGGANGACACAGAGAAG...	461
>>>>...>>>> >>>>...>>>>	
G: TGTGCATGGACCTGAGGGCCAACCTGAAGCAGG-TCAAGAAGGAGGACACAGAGAAGGTG	77550
Q:GAGCCGGGACCTNCGAGACGTTGGTNACTTGAAGAAGAACATCGAGGGAG	511
>>...>>>> >>>>...>>>>	
G: CG...CACAGGAG-CGGGACCTGCGAGACGTTGGTNACTTGAAGAAGAACATCGA-GGAG	77846
Q: AAGTNTGGCAT-GAAGGCCNGAAGAAGATNTTT TANTTCGAGT-CTA-GCCA-TNGCTG-	566
G: AAGTCTGGCATGGAGGGCCGAAGAAGATNTTT GAGTCCGAGTCTAGGCCACTCGCTGC	77906
Q: CCCTA-ANCTG-CCC-GT-TCCGG-TTCCAGCAGA	596
G: CCCTACGCCTGCCCGGTGCCCGGCTCCAGCAGA	77941

Table 6. Bootstrap Data: Human Donor and Acceptor Splice Site Models

A. Donor site nucleotide counts

Position	-2	-1	+1	+2	+3	+4	+5
A	1048	155	0	0	873	1256	116
C	233	56	0	0	47	129	88
G	240	1407	1754	0	790	214	1467
T	233	136	0	1754	44	155	83

B. Donor site log-odds scores (in nats)

Position	-2	-1	+1	+2	+3	+4	+5
A	0.865504	-1.02199	-4.61512	-4.61512	0.683636	1.045862	-1.30262
C	-0.62363	-1.99257	-4.61512	-4.61512	-2.15396	-1.20007	-1.56735
G	-0.59457	1.159017	1.378841	-4.61512	0.584258	-0.70705	1.200649
T	-0.62363	-1.14892	-4.61512	1.378841	-2.21412	-1.02199	-1.62299

C. Acceptor site nucleotide counts

Position	-6	-5	-4	-3	-2	-1	+1
A	117	132	394	72	1754	0	415
C	790	691	594	1287	0	0	233
G	109	100	376	3	0	1754	938
T	738	831	390	392	0	0	168

D. Acceptor site log-odds scores (in nats)

Position	-6	-5	-4	-3	-2	-1	+1
A	-1.29434	-1.17783	-0.10589	-1.75752	1.378841	-4.61512	-0.05452
C	0.584258	0.451156	0.300924	1.07016	-4.61512	-4.61512	-0.62363
G	-1.36252	-1.44522	-0.15213	-4.09386	-4.61512	1.378841	0.755104
T	0.516558	0.634583	-0.11598	-0.11092	-4.61512	-4.61512	-0.94358

Raw counts were from a data set composed entirely of canonical GT..AG splice junctions from human provided by Ian Korf (personal communication).

Table 7. G-test statistics between donor and acceptor splice site data gathered for human and mouse

HUMAN DONOR										G		T									
A	66086	125302	19550	339	255	118063	137374	16906	34522												
C	71228	21716	5647	272	1744	5208	14406	11113	29566												
G	35730	22182	158103	195626	174	67616	22747	153686	37807												
T	23342	27186	13086	149	194213	5499	21859	14681	94491												
MOUSE DONOR										G		T									
A	40914	78020	11898	212	161	74462	86598	9120	20488												
C	44513	13557	3454	191	1069	3395	9115	7104	19455												
G	22564	13907	98847	122164	104	41432	13550	98155	22911												
T	14689	17196	8481	103	121327	3366	13387	8263	59784												
G_{adj}	4.16	2.26	12.99	2.15	0.38	18.34	27.81	218.88	84.63												
$P(worse)$	0.245	0.52	0.005	0.542	0.944	0.0004	4.E-06	4.E-47	3.E-18												
HUMAN ACCEPTOR																		A		G	
A	31170	28416	26068	23747	21813	19917	17859	16407	16609	18937	20709	21663	16520	17115	46888	11064	195835	240	50194		
C	54167	54334	54756	56421	55872	55489	55875	51675	56327	58598	64753	66992	68393	59169	54105	127878	184	329	27561		
G	28052	27339	26743	25500	24215	22707	20912	20127	21471	22598	20651	17887	12269	12079	40456	557	168	195658	96558		
T	82997	86296	88819	90718	94486	98273	101740	108177	101979	96253	90273	89844	99204	108023	54937	56887	199	159	22073		
MOUSE ACCEPTOR																		A		G	
A	19497	16958	16004	14236	13137	11938	10975	9666	9855	10915	12679	12992	10311	10115	30800	6541	122335	149	31298		
C	33647	34069	34427	35434	35241	34886	34321	32755	34547	36736	41130	41868	43364	37108	32845	79907	116	199	17194		
G	18525	18180	17279	16601	15700	15062	13685	13138	13803	14399	13254	11316	7625	7360	25112	317	97	122224	60440		
T	50909	53377	54876	56318	58517	60711	63619	67045	64402	60560	55555	56448	61331	68056	33889	35887	115	99	13748		
G_{adj}	45.48	68.24	19.86	34.16	28.81	50.76	25.69	43.11	35.21	50.76	23.74	15.79	10.07	25.82	67.52	16.42	0.83	0.13	0.31		
$P(worse)$	7.E-10	1.E-14	2.E-04	2.E-07	2.E-06	6.E-11	1.E-05	2.E-09	1.E-07	6.E-11	3.E-05	0.001	0.018	1.E-05	1.E-14	0.001	0.842	0.988	0.958		

G_{adj} = G-test statistic from $R \times C$ test of independence, after applying Williams's correction. In all cases, the Williams correction, q , was less than 1.002; see *Biometry* third edition by Sokal and Rohlf. $P(worse)$ = the probability of obtaining a worse correspondence between the human and mouse data at each position, if the data were sampled from the same underlying distribution, given $(2-1)(4-1) = 3$ degrees of freedom. Lower values for $P(worse)$ imply a higher likelihood that the underlying distribution at each position differs between the two species.

Table 8. Exon Number Predictions in EST Data

<i>Method</i>	<i>Single-Exon ESTs</i>	<i>Multiple-Exon ESTs</i>	<i># Exons in Multiple-Exon ESTs, Total</i>	<i># Splice Junctions</i>	<i># Exons in Multiple-Exon ESTs, Average</i>	<i>Total Exons (Single- + Multiple-Exon ESTs)</i>
Spidey	6480	8743	37628	28885	4.30	44108
Sim4	6653	8570	36161	27591	4.21	42814
BLAT	7638	7585	31569	23984	4.16	39208
EST_GENOME*	7390	7833	33970	26137	4.34	41360
EXALIN	7170	8053	35397	27339	4.40	42567

Table 9. Exon Length Distributions**A. External exons**

Exon length	1-5	5-10	11-15	15-20	20-25	25-30	30-60	60-90	90-120	120-160	160-600	600-1K	1K-2K	2K-5K
Spidey	0	5	150	525	495	503	2836	3994	3753	1986	1133	4386	3022	1086
Sim4	0	4	111	508	636	460	2745	4015	3696	2257	1206	4263	2881	956
BLAT	0	242	141	96	225	312	2293	3731	3567	2058	1254	4507	3150	1183
EST_GENOME*	0	0	0	0	40	389	2501	4104	3845	2196	1157	4362	3218	1214
EXALIN	2	10	235	222	298	479	2599	3933	3589	2963	1091	4533	3161	1222

B. Internal exons

Exon length	1-5	5-10	11-15	15-20	20-25	25-30	30-60	60-90	90-120	120-160	160-600	600-1K	1K-2K	2K-5K
Spidey	55	464	398	239	111	64	1266	6182	4522	3961	2250	571	137	6
Sim4	0	44	81	53	88	76	1092	6097	4637	3967	2175	579	143	9
BLAT	88	70	7	9	5	7	926	4683	4138	3768	2031	587	104	5
EST_GENOME*	0	0	0	1	4	6	1038	6018	4377	3991	2241	510	118	4
EXALIN	0	16	3	5	7	8	1107	6280	4711	4170	2316	530	133	5

Table 10. Intron Length Distributions

Intron Length	<5	5-9	10-19	20-29	30-39	40-59	60-79	80-100	100-120	120-160	160-200	200-4k	4k-10k	10k-20k	>20k
Spidey	67	27	240	280	120	145	171	773	439	1395	13687	6481	2461	726	1916
Sim4	26	1	11	8	27	107	186	745	465	1546	12992	6326	2362	567	2232
BLAT	0	0	16	25	41	91	140	726	400	1361	12403	5744	2106	513	432
EST_GENOME*	0	0	14	7	8	47	113	720	408	1340	12778	6254	2245	578	1627
EXALIN	0	0	26	12	8	46	119	759	430	1381	13431	6457	2404	591	1625

Table 11. Discrepancy counts and overall nucleotide counts in 7 non-overlapping, 5-nucleotide windows immediately upstream of donors and downstream of acceptors after attempted alignment of 16,639 human ESTs to the human genome by the indicated programs.

EST_GENOME							
Donor up	Overall	Acceptor	Overall				
2037	131445	1958	131445				
1530	131445	1435	131445				
1451	131445	1468	131443				
1338	131441	1495	131439				
1225	131330	1568	131267	Sim4			
1380	130574	1600	130337	Donor up	Overall	Acceptor	Overall
1683	129588	1883	128746	3247	135455	3478	135455
				1905	135207	1646	135422
				1587	134567	1605	134853
				1450	133470	2057	134860
BLAT				1246	131952	1504	133640
Donor up	Overall	Acceptor	Overall				
861	119421	946	119421	1444	130622	1552	129562
2007	118868	1980	119381	1675	129042	1745	127922
2027	117804	1927	118490				
2024	117277	2116	117660				
1414	116682	1633	117635	EXALIN			
1543	115308	1675	116934	Donor up	Overall	Acceptor	Overall
2210	114441	2421	114707	2098	136990	2002	136998
				1968	136943	1893	136952
				1826	136245	1812	136291
				1836	135762	1772	135585
SPIDEY				1634	134893	1597	135110
Donor up	Overall	Acceptor	Overall				
5574	136364	5550	136364	1711	133601	1804	134181
2882	135351	2717	134518	1988	131966	2111	133018
2615	133501	2180	132542				
2648	131115	2383	130596				
2356	128695	2183	128541				
2309	127496	1981	125556				
2387	125362	1996	123332				


```

T: GCTTCAGCTGTGCTTCCAGGGCCCCACAGCACTGAAAGCCCAGCAAGCACCACAAGAAC 407
|
G: GCTTCAGCTGTGCTTCCAGGGCCCCACAGCACTGAAAGCCCAGCAAGCACCACAAGAAC 1214838

T: C.....TTGATATTTCACTTCAGTAACACACCCTTCTCTCTCCAGTCCA 362
|<<<<<<...<<<<<<|
G: CCTAAAA...ACTCACTTGATATTTCACTTCAGTAACACACCCTTCTCTCTCCAGTCCA 1217875

T: CAGAATCAGGCAATATCCGATTAGGGTTTGACTTATATGTGATATTTCTCTGCCACTGGC 302
|
G: CAGAATCAGGCAATATCCGATTAGGGTTTGACTTATATGTGATATTTCTCTGCCACTGGC 1217935

T: TGGGAACTCTCAGGGAACCTCGTCAAAGACATCACTTCTTCACTGGT..... 256
|
G: TGGGAACTCTCAGGGAACCTCATCAAAGACATCACTTCTTCACTGGTCTACAA...ACCTA 1220688

T: .CATGTCTCCCAGGTGGTTCATGCCAGATCGTATGAGTGCATTCCCATTGAATGCTCCA 197
<|
G: CCATGTCTCCCAGGTGGTTCATGCCAGATCGTATGAGTGCATTCCCATTGAATGCTCCA 1220747

T: GGTGTGAAGCATCACAACCTTTAGATTCTTTCCAGATGAGACGTCGACTGCTTCTT 137
|
G: GGTGTGAAGCATCACAACCTTTAGATTCTTTCCAGATGAGACGTCGACTGCTTCTT 1220807

T: CATT.....CTTTCCCTTGATTGTTTGCCATAGGTTTTCTTCCAGAGAT 92
|||<<<<<<...<<<<<<|
G: CATTCTAAAA...ATCTACCTTTTCCCTTGATTGTTTGCCATAGGTTTTCTTCCAGAGAT 1227509

T: GCCAGTGGTGATCCAGGGTAGGATCTTTATGCAACTGTGCCACTGCAGAGGAGCACACCA 32
|
G: GCCAGTGGTGATCCAGGGTAGGATCTTTATGCAACTGTGCCACTGCAGAGGAGCACACCA 1227569

T: AGAGCACACAAACCAGCCGTTTCAT 7
|
G: AGAGCACACAAACCAGCCGTTTCAT 1227594

```

EXIT 0

EXALIN [02-Apr-2005]
 ARGS: -a 1 -i 2 -p /home/mzhang/package/HumanSPM.par -A 17720436 -B 17807145
 /am/tmp/ucsc/contig_may04/4/NT_022792/NT_022792.mask 33966
 T-SEQUENCE: gi|33966|gb|X15949.1|HSIRF2 Human mRNA for interferon regulatory factor-2 (IRF-2).
 T-LENGTH: 2144
 G-SEQUENCE: NT_022792
 G-LENGTH: 23437590
 SCORE: 2509.6 nats (3620.6 bits)
 SPLICE-DIR: +
 T-STRAND: -

EXON: 2144 - 840 (17720436 - 17721741)	1527.8	97.79%
EXON: 839 - 793 (17723378 - 17723424)	58.8	100.00%
EXON: 792 - 628 (17731590 - 17731754)	206.2	100.00%
EXON: 627 - 510 (17740833 - 17740950)	147.5	100.00%
EXON: 509 - 463 (17750842 - 17750888)	58.8	100.00%
EXON: 462 - 286 (17751207 - 17751383)	221.2	100.00%
EXON: 285 - 186 (17752144 - 17752243)	121.0	99.00%
EXON: 185 - 93 (17761653 - 17761745)	116.2	100.00%
EXON: 92 - 1 (17807054 - 17807145)	111.0	98.91%

T: TCCACAGGAAAATCTGATTGCTAGATGAGCTCATAAAAGCTTTTTTCACCTTTACAAAAT	2085
G: TCCACAGGAAAATCTGATTGCTACATGAGCTCATAAAAGCTTTTTTCACCTTTACAAAAT	17720495
T: T--AAAAAAAATGTGCCACAAGGATGTAATAACAACCTGCTGATAAACAAGAAAAGGAAT	2027
G: TAAAAAAAATGTGCCACAAGGATGTAATAACAACCTGCTGATAAACAAGAAAAGGAAT	17720555
T: CTTAAAATTATAAATTTGCTGTAGAAAAGATAAAAAACAATTATATTTTATTTAGAATTT	1967
G: CTTAAAATTATAAATTTGCTGTAGAAAAGATAAAAAACAATTATATTTTATTTAGAATTT	17720615
T: TACCTTGAATAAAAATATCCCCTGTATAAAAAATAAAAAAGCTTGCTCTGGTTAGAATTA	1907
G: TACCTTGAATAAAAATATCCCCTGTATAAAAAATAAAAAAGCTTGCTCTGGTTAGAATTA	17720675
T: GAGTATTTTGTCTTCAAATCTGGGAATTTGCATAATATTTCCATGATACTTTTTCCTTT	1847
G: GAGTATTTTGTCTTCAAATCTGGGAATTTGCATAATATTTCCATGATACTTTTTCCTTT	17720735
T: GTACCGCGTGGCATTCAAGCATAGCAGATTAGAAGGATTTTTTTTAAAGCAGTCTGAAAA	1787
G: GTACCGCGTGGCATTCAAGCATAGCAGATTAGAAGGATTTTTTTTAAAGCAGTCTGAAAA	17720795
T: TGGGACATCTGTAGAGAAATTCATTTCTTCTTCTCCTCCGGATGTGGAATGGAAGCTTT	1727
G: TGGGACATCTGTAGAGAAATTCATTTCTTCTTCTCCTCCGGATGTGGAATGGAAGCTTT	17720855
T: GAGGGAAGGAAAAGTAGGAAAAGAGC-GGGATG--G-GATGGGATGGGA-TGGGA-T--G	1675
G: GAGGGAAGGAAAAGTAGGAAAAGAGCAGTGA-GCCGAGAT--CATGCCACT-GCACTCCA	17720911
T: GGATGGGATGGGATAGGAAGAGAGGCTGGGGAATGGGCAGAGAAGGGGGTGCTGAGTGTG	1615
G: GCCTGGG-T--GACA-GAAGAGAGGCTGGGGAATGGGCAGAGAAGGGGGTGCTGAGTGTG	17720967
T: CTGTGAGATAGAGCAAGATCACAAGAAGGCCTATCTGTAAGTGCTTTAAGATAAGGTGCA	1555
G: CTGTGAGATAGAGCAAGATCACAAGAAGGCCTATCTGTAAGTGCTTTAAGATAAGGTGCA	17721027
T: GAAGCAGACTGCAATGTCGCTAGTGTGCTGCTATGTACATATTCACAAGAATAAAAAATTC	1495
G: GAAGCAGACTGCAATGTCGCTAGTGTGCTGCTATGTACATATTCACAAGAATAAAAAATTC	17721087
T: CAGCTAGTTCACATTATCTCGTCCGTTCTGAGGCATCCACTCCACCTTGCTGGCCTTGAC	1435
G: CAGCTAGTTCACATTATCTCGTCCGTTCTGAGGCATCCACTCCACCTTGCTGGCCTTGAC	17721147
T: CGCAGGCAATGCAGCACCTCCACTGCCCTTGTTGGTCCCTGAACTTGAGTGAGTAGCCCT	1375
G: CGCAGGCAATGCAGCACCTCCACTGCCCTTGTTGGTCCCTGAACTTGAGTGAGTAGCCCT	17721207
T: GGGAGATCCAGCAGGCTCTAGAAAACACAGTCTACCAATGGGCTGGAGTCCCTGAGTTAAA	1315


```

|||||<<<<<...<<<<<|||||
G: TATCTACTCCTGGTTGATGCTTTCCTAACA...TAGTACCTGTATGGATTGCCCAGTTTC 17752164

T: TAAAGAGTGGTGCATCTTTTCCACATCCCACCCATGTCTAGCCGCATGCATCCAGGGGA 205
|||||
G: TAAAGAGTGGTGCATCTTTTCCACATCCCACCCATGTCTAGCCGCATGCATCCAGGGGA 17752224

T: TCTGAAAAATCTTCTTTC.....CTTGTTAAGCCACTTGAGCCCCGGGA 160
|||||<<<<<...<<<<<|||||
G: TCTGAAAAATCTTCTTTCCTGAAA...ACTCACCTTGTTAAGCCACTTGAGCCCCGGGA 17761678

T: TCGTGTGGAGTTTATCTGCTCCTCCAGCCACGGGGCGCATGCGCATCCTTCCACCGGCA 100
|||||
G: TCGTGTGGAGTTTATCTGCTCCTCCAGCCACGGGGCGCATGCGCATCCTTCCACCGGCA 17761738

T: TGGTGCC.....CTCTCAGTGTGCTTTTTTACGCTACCAATACAATTC 55
|||||<<<<<...<<<<<|||||
G: TGGTGCCCTTGAG...TGTTACCTCTCAGTGTGCTTTTTTACGCTACCAATACAATTC 17807091

T: GCAAGGTATAAGTGTGCTAGGGTGTGTAATGGAAATGAAAGCCCGTCAGTT 1
|||||
G: GCAAGGTATAAGTGTGCTAGGGTATGTGAAATGGAAATGAAAGCCCGTCAGTT 17807145

```

EXIT 0

EXALIN [02-Apr-2005]
 ARGS: -a 1 -i 1 -p /home/mzhang/package/HumanSPM.par -A 4384173 -B 4415890
 /am/tmp/ucsc/contig_may04/1/NT_004671/NT_004671.mask 387656
 T-SEQUENCE: gi|387656|gb|M25077.1|HUMANTARNP Human SS-A/Ro ribonucleoprotein autoantigen 60 kd subunit mRNA,
 complete cds.
 T-LENGTH: 1851
 G-SEQUENCE: NT_004671
 G-LENGTH: 14767479
 SCORE: 2229.7 nats (3216.7 bits)
 SPLICE-DIR: +
 T-STRAND: +

EXON: 1 - 164 (4384173 - 4384338) 199.5 98.80%
 EXON: 165 - 764 (4393150 - 4393750) 747.2 99.83%
 EXON: 765 - 985 (4399936 - 4400156) 276.2 100.00%
 EXON: 986 - 1132 (4400617 - 4400763) 183.8 100.00%
 EXON: 1133 - 1270 (4401029 - 4401166) 172.5 100.00%
 EXON: 1271 - 1387 (4405480 - 4405596) 146.2 100.00%
 EXON: 1388 - 1501 (4406302 - 4406415) 142.5 100.00%
 EXON: 1502 - 1648 (4406678 - 4406824) 183.8 100.00%
 EXON: 1649 - 1727 (4408695 - 4408773) 98.8 100.00%
 EXON: 1728 - 1851 (4415767 - 4415890) 155.0 100.00%

T: CACAGGCCGACGTCGAGAGGGCCCTGCTTTACTCCTCCTTTCTCCTCCTTCTCCCGCGG 60
 |||
 G: CACAGGCCGACGTCGAGAGGGCCCTGCTTTACTCCTCCTTTCTCCTCCTTCTCCCGCGG 4384232
 T: CTTCTGCGC-GAGAGGCGTCG-CCGGGATCTGGGTTTTGGAAGAAGGATCTTTGTGGGAA 118
 |||
 G: CTTCTGCGCGGAGAGGCGTCGCCGGGATCTGGGTTTTGGAAGAAGGATCTTTGTGGGAA 4384292
 T: GACAGGGTGAATTTATCACAGAGGAATAACGAGGGAGAGGAGAAAAG..... 164
 |||>>>>>...>>>>>
 G: GACAGGGTGAATTTATCACAGAGGAATAACGAGGGAGAGGAGAAAAGTTTGT...TGTTA 4393149
 T: .GTTTCCTAAAGAC-AAAAAAAAATGGAGGAATCTGTAAACCAATGCAGCCACTGAATGA 222
 >|||
 G: GGTTCCTAAAGACAAAAAAAAATGGAGGAATCTGTAAACCAATGCAGCCACTGAATGA 4393208
 T: GAAGCAGATAGCCAATTCTCAGGATGGATATGTATGGCAAGTCACTGACATGAATCGACT 282
 |||
 G: GAAGCAGATAGCCAATTCTCAGGATGGATATGTATGGCAAGTCACTGACATGAATCGACT 4393268
 T: ACACCGGTTCTTATGTTTCGGTTCTGAAGGTGGGACTTATTATATCAAGAACAGAAGTT 342
 |||
 G: ACACCGGTTCTTATGTTTCGGTTCTGAAGGTGGGACTTATTATATCAAGAACAGAAGTT 4393328
 T: GGGCCTTGA AAAATGCTGAAGCTTTAATTAGATTGATTGAAGATGGCAGAGGATGTGAAGT 402
 |||
 G: GGGCCTTGA AAAATGCTGAAGCTTTAATTAGATTGATTGAAGATGGCAGAGGATGTGAAGT 4393388
 T: GATACAAGAAATAAAGTCATTTAGTCAAGAAGGCAGAACCACAAAGCAAGAGCCTATGCT 462
 |||
 G: GATACAAGAAATAAAGTCATTTAGTCAAGAAGGCAGAACCACAAAGCAAGAGCCTATGCT 4393448
 T: CTTTGCACTTGCCATTTGTTCCAGTGCTCCGACATAAGCACAAAACAAGCAGCATTTAA 522
 |||
 G: CTTTGCACTTGCCATTTGTTCCAGTGCTCCGACATAAGCACAAAACAAGCAGCATTTAA 4393508
 T: AGCTGTTTCTGAAGTTTGTGCGCATTCCTACCCATCTCTTACTTTTATCCAGTTAAGAA 582
 |||
 G: AGCTGTTTCTGAAGTTTGTGCGCATTCCTACCCATCTCTTACTTTTATCCAGTTAAGAA 4393568
 T: AGATCTGAAGGAAAGCATGAAATGTGGCATGTGGGGTCTGCGCCCTCCGGAAGGCTATAGC 642
 |||
 G: AGATCTGAAGGAAAGCATGAAATGTGGCATGTGGGGTCTGCGCCCTCCGGAAGGCTATAGC 4393628
 T: GGACTGGTACAATGAGAAAGGTGGCATGGCCCTTGCTCTGGCAGTTACAAAATATAAACA 702
 |||
 G: GGACTGGTACAATGAGAAAGGTGGCATGGCCCTTGCTCTGGCAGTTACAAAATATAAACA 4393688
 T: GAGAAATGGCTGGTCTCACAAAGATCTATTAAGATTGTACATCTTAAACCTTCCAGTGA 762
 |||
 G: GAGAAATGGCTGGTCTCACAAAGATCTATTAAGATTGTACATCTTAAACCTTCCAGTGA 4393748

EXALIN [02-Apr-2005]
 ARGS: -a 1 -i 2 -p /home/mzhang/package/HumanSPM.par -A 1148114.96 -B 1219189
 /am/tmp/ucsc/contig_may04/19/NT_077812/NT_077812.mask 1924939
 T-SEQUENCE: gi|1924939|gb|X98411.1|HSMYOSIE Homo sapiens partial mRNA for myosin-IF.
 T-LENGTH: 2711
 G-SEQUENCE: NT_077812
 G-LENGTH: 1291194
 SCORE: 3035.8 nats (4379.7 bits)
 SPLICE-DIR: +
 T-STRAND: -

EXON: 2700 - 2273 (1190058 - 1190484)	491.0	97.43%
EXON: 2272 - 2106 (1191257 - 1191426)	200.5	98.24%
EXON: 2105 - 1904 (1191514 - 1191709)	176.0	89.90%
EXON: 1903 - 1820 (1194359 - 1194442)	83.5	93.02%
EXON: 1819 - 1671 (1195333 - 1195481)	182.2	99.33%
EXON: 1670 - 1524 (1195669 - 1195815)	183.8	100.00%
EXON: 1523 - 1378 (1196218 - 1196363)	153.0	94.59%
EXON: 1377 - 1208 (1199076 - 1199245)	208.5	99.41%
EXON: 1207 - 1093 (1199339 - 1199453)	143.8	100.00%
EXON: 1092 - 948 (1205132 - 1205276)	181.2	100.00%
EXON: 947 - 849 (1205379 - 1205477)	115.8	97.98%
EXON: 848 - 742 (1205829 - 1205935)	123.0	97.22%
EXON: 741 - 660 (1208827 - 1208908)	102.5	100.00%
EXON: 659 - 574 (1210786 - 1210871)	99.5	97.67%
EXON: 573 - 406 (1213177 - 1213344)	210.0	100.00%
EXON: 405 - 319 (1214530 - 1214616)	108.8	100.00%
EXON: 318 - 232 (1216916 - 1217002)	108.8	100.00%
EXON: 231 - 151 (1217117 - 1217197)	101.2	100.00%
EXON: 150 - 1 (1219040 - 1219189)	187.5	100.00%

```

T: CCCCCCTCCCAACTGTGCCTGGAACTTTGCCAACAGCACAGGACTCAAGACCCATTTG 2641
   |||
G: CCCCCCTCCCAACTGTGCCTGGCACTTTGCCAACAGCACAGGACTCAAGACCCATTTG 1190117

T: TTAATCACATGGCAGTTGGGAGGTAAATTCTGCCTGTTAGTCCCCTTTGACACACACAGTT 2581
   |||
G: TTAATCACATGGCAGTTGGGAGGTAGATTCTGCCTGTTAGTCCCCTTTGACACACACAGAA 1190177

T: TTGGACAATGGGCAGCAGGTGTGATGTTGGAGGAAAACCAGGGCCCAGGGGCGGGGGCTG 2521
   |||
G: ATGGAGAATGGGCAGCAGGTGTGATGTTGGAGGAAGACCAGGGCCCAGGGGCGGGGGCTG 1190237

T: CACCAGTGACCTGGTGACCCAGGGCCTGGGCAAGGACTGGAGGCCAAAGGATTGGACTTT 2461
   ||
G: CAGCAGTGACCTGGTGACCCAGGGCCTGGGC-AGGACTGGAGGCCAAAGGACTGGACTTT 1190296

T: TAGGCTATTGCAGCCCAGGTAAACGAGGCTTTTCATTGGCAGGGCCTGGCTCCCCACCAGG 2401
   |||
G: TAGGCTATTGCAGCCCAGGTAAACGAGGCTCTCATTGGCAGGGCCTGGCTCCCCACCAGG 1190356

T: CCGGCAGGCAGATAGGCGGGCGAAAGAGAAGGCAGTATCCCAGGGCCCAGCTCAGATCTT 2341
   |||
G: CCGGCAGGCAGATAGGCGGGCGAAAGAGAAGGCAGTATCCCAGGGCCCAGCTCAGATCTT 1190416

T: CTCCACGTAGTTTCTGGGAAAAGGCCCTCTGGCCGTGAAGCCGGCCCTTCCACCAGCC 2281
   |||
G: CTCCACGTAGTTTCTGGGAAAAGGCCCTCTGGCCGTGAAGCCGGCCCTTCCACCAGCC 1190476

T: CGAGGGAT.....CTTCCATGAGGATCTCAATGACCTCGTTCACGTTGAA 2236
   |||<<<<<<...<<<<<<|||
G: CGAGGGATCTGTGG...ACACACCTTCCATGAGGATCTCAATGACCTCGTTCACGTTGAA 1191293

T: GCTCAGCTCGTCCACATCTTGGCCCACGTACTGGTATAGGGCCCGGCACCTGGGACCATG 2176
   |||
G: GCTCAGCTCGTCCACATCTTGGCCCACGTACTGGTATAGGGCCCGGCACCTGGGACCATG 1191353

T: TGTCCGAGGCTGGGGCTTGGGTCCG-CC-CA-CAGGCACTGGCCGTTGCCCCACGCTGCG 2119
   |||
G: TGTCCGAGGCTGGGGCTTGGGTCCGCCCACACAGGCACTGGCCGTTGCCCCACGCTGCG 1191413

T: CTTCTCTGCATG.....CCGCCATCCCCTGGTCAGGCACGTTGAGGAA 2074
   |||<<<<<<...<<<<<<|||
G: CTTCTCTGCATGCTGTGG...ACCTACCCGGCCATGCCCTGGTCAGGCACGTTGAGGAA 1191545

```

T: TTCTGTGTGTGTGCTCTGAGGGCGGACGTGCCCGGGGTCGTCTGCTGGCATCCCAGGGATG 2014
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: TTCTGTGTGTGTGCTCTGAGGGCGGACGTGCCCGGGGTCGTCTGCTGGC-TCCCAGGGATG 1191604

T: TGGACGGAGGG--CCGGGGAGGCCTGTGGGTGCCCCCTGGGCCA-A-ATGAATTTCTCCA 1958
 ||||||||||||||||||||||||||||||||||||||||||||| ||| | ||| | |||||
 G: TGGACGGAGGGCCCCGGGGAGGCCTGTGGGTGCCCCCT---CCAGACATG-A--TCTCCA 1191658

T: GGGGACAGGGGGCCCCCCTGTGGGCAGAGGGGG--CCCCATTTTCGATCCATGC... 1904
 ||||||||||||||||||||||||||||||||||||||||||||| ||||||| |||||||<<<<<<
 G: -GGGACAGGGGG---CCCCCTCT-GGCAGAGGGGGGCACCCCATTCGATCCATGCCTGT 1191709

T:CT-TGGGGGGCGCCAGGGGGCCGCCGGGTAGGGGCTTGGGACGACCTCC 1856
 <<...<<<<<<| |||||||||| |||||||||| |||||||||| |||||||||| |||||||
 G: GG...CCTCACCTTGGGGGGCG-CAGGGGGCCGCCGGGTAGGGGCTTGGGACGACCTCC 1194406

T: GAGGTTTACCCTTGGCCAAATCCCTTTCCG-GTAGGCT.....TGGAGTTC 1812
 ||||||| ||||||| ||||||| ||||| |||||||<<<<<<...<<<<<<||| |||
 G: GAGGTTTTCCTTGGCCATTCCT-TTCCGCTAGGCTCTGAAA...ACTCACTGGAGCTC 1195340

T: TTGGGCAGCCCATCGCCACGCTGACCGTGAGGGTCCGACCGCCAACCTTGAGCACTGCC 1752
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: TTGGGCAGCCCATCGCCACGCTGACCGTGAGGGTCCGACCGCCAACCTTGAGCACTGCC 1195400

T: AAGTCGCCGAAGCCGCGGGAGAAGGTGACGCTGCGGGTGCCGCCACCGCCCAGCCCTCC 1692
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: AAGTCGCCGAAGCCGCGGGAGAAGGTGACGCTGCGGGTGCCGCCACCGCCCAGCCCTCC 1195460

T: TTCTTCACCCGAAACTGTAGT.....GTGTGCTGAAGGTGAGGGGCAGG 1647
 |||||||||||||||||||<<<<<<...<<<<<<||| ||||||| ||||||| |||||||
 G: TTCTTCACCCGAAACTGTAGTCTATGG...CCGTACGTGTGCTGAAGGTGAGGGGCAGG 1195692

T: GGCCTCCGCGTTCGCTCCTCGAAGCGCTTGACACAGAAGGCTGACAAACTCGGTCTTGAAG 1587
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: GGCCTCCGCGTTCGCTCCTCGAAGCGCTTGACACAGAAGGCTGACAAACTCGGTCTTGAAG 1195752

T: ACGCTCTCCAGGAAGCTGTGCGGGCATCCTCTTGGAGGATGAAGAAGTCGCTCCTGTGCG 1527
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: ACGCTCTCCAGGAAGCTGTGCGGGCATCCTCTTGGAGGATGAAGAAGTCGCTCCTGTGCG 1195812

T: GTG.....CTGAGGGAGACCCCCGAGAGCCTGGATGTCCAATTTCTTC 1482
 ||<<<<<<...<<<<<<||| ||||||| ||||||| ||||||| ||||||| |||||||
 G: GTGCTGGG...CTCACCTGAGGGAGACTCCCCGAGAGCCTGGATGTCCAATTTCTTC 1196259

T: TTCAAGATTCACACA-CTGGACCTTCTCAGGTCCTTCTTCAFC-TTCTCTCGCCAA 1424
 ||||||| ||||||| ||||| || ||| ||||||| ||||||| ||| ||||||| |||||||
 G: TTCAAGACTTCACACACCTGG-CCCTTCTCAGGTCCTTCTTCA-CTTCTCTCGCCAA 1196317

T: TCACATACACACTTGGGCGTCAGGATCAAGTCCCCTTGATGGG..... 1378
 |||||||||||||||||||||||||||||||||||||||||||||<<<<<<...<<<<<<
 G: TCACATACACACTTGGGCGTCAGGATCAAGTCCCCTTGATGGGCTGTGG...CCTCA 1199075

T: .CTTGAAGCGGGTTCGTAAGTGGTGACCGAATCGGCGAAGTCCACCCGCTCCTTCTTGC 1319
 <||| ||||||| ||||||| ||||||| ||||||| ||||||| ||||||| ||||||| |||||||
 G: CTTGAAGCGGGTTCGTAAGTGGTGACCGAATCGGCGAAGTCCACCCGCTCCTTCTTGC 1199134

T: CCAGGAAGTACGACGCTCGGGCCGCTCCTCCAGCCCCAGGTAGTCCCCGACGAAGTTC 1259
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: CCAGGAAGTACGACGCTCGGGCCGCTCCTCCAGCCCCAGGTAGTCCCCGACGAAGTTC 1199194

T: GATTGATGCTGTGCGCCTCCGCTCCTTCTTGTTCAGCAGGATGTTGGAAG..... 1208
 |||||||||||||||||||||||||||||||||||||||||||||<<<<<<...
 G: GATTGATGCTGTGCGCCTCCGCTCCTTCTTGTTCAGCAGGATGTTGGAAGTGC GG... 1199246

T:CTTCTCCCGCATCTCCTCGTACTTCCGGACAGCCACGTGGCGCCGCCAGGCCT 1154
 <<<<<<||| ||||||| ||||||| ||||||| ||||||| ||||||| ||||||| |||||||
 G: TCTCACCTTCTCCCGCATCTCCTCGTACTTCCGGACAGCCACGTGGCGCCGCCAGGCCT 1199392

T: TCTGGATGGTTTCGGGCAAAGCCATCGAACTTTCGCTCTCGCACCTCCTCCAGGAGGAAAA 1094
 |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 G: TCTGGATGGTTTCGGGCAAAGCCATCGAACTTTCGCTCTCGCACCTCCTCCAGGAGGAAAA 1199452

T: G.....CGACTCTGGGTTCTTGACAAAGACCTTGGTGTCTCCCATCTGGT 1049
 |<<<<<<...<<<<<<||| ||||||| ||||||| ||||||| ||||||| |||||||
 G: GCTGGG...ACTCACGACTCTGGGTTCTTGACAAAGACCTTGGTGTCTCCCATCTGGT 1205175

G: GCATCACGGGTGTAGGCTGCCTGCTCCACGTTGAGGGTCACATTGATGGACTCGCTGCGC 1219145

T: CCGCCCCAGCGGCTGTCCATCTTGCGGCTGGTCAGCTTCTCCTG 1
|||||

G: CCGCCCCAGCGGCTGTCCATCTTGCGGCTGGTCAGCTTCTCCTG 1219189

EXIT 0

EXIT 0

EXALIN [02-Apr-2005]
 ARGS: -a 1 -i 2 -p /home/mzhang/package/HumanSPM.par -A 19225406 -B 19310898
 /am/tmp/ucsc/contig_may04/15/NT_010194/NT_010194.mask 5410335
 T-SEQUENCE: gi|5410335|gb|AF106685.1|AF106685 Homo sapiens myelin gene expression factor 2 mRNA, complete cds.
 T-LENGTH: 1983
 G-SEQUENCE: NT_010194
 G-LENGTH: 53619965
 SCORE: 2108.9 nats (3042.5 bits)
 SPLICE-DIR: +
 T-STRAND: -

EXON: 1983 - 1561 (19225406 - 19225825)	504.8	98.58%
EXON: 1560 - 1509 (19231786 - 19231837)	65.0	100.00%
EXON: 1508 - 1300 (19231917 - 19232125)	261.2	100.00%
EXON: 1299 - 1228 (19233854 - 19233925)	82.0	97.22%
EXON: 1227 - 1129 (19234227 - 19234325)	95.8	92.93%
EXON: 1128 - 1060 (19234628 - 19234696)	86.2	100.00%
EXON: 1059 - 1009 (19234988 - 19235038)	63.8	100.00%
EXON: 1008 - 907 (19236546 - 19236647)	127.5	100.00%
EXON: 906 - 843 (19240747 - 19240810)	76.0	98.44%
EXON: 842 - 793 (19240929 - 19240978)	62.5	100.00%
EXON: 792 - 639 (19241523 - 19241676)	173.8	96.77%
EXON: 638 - 447 (19242367 - 19242558)	169.2	90.67%
EXON: 446 - 353 (19248687 - 19248780)	109.5	97.87%
EXON: 352 - 345 (19248875 - 19248882)	10.0	100.00%
EXON: 344 - 292 (19250103 - 19250155)	66.2	100.00%
EXON: 291 - 83 (19251385 - 19251593)	257.2	99.52%
EXON: 82 - 15 (19260831 - 19260898)	78.2	97.10%

T: AAAATGGCTTATTTTCATTAATAAACAGTATACCCATTCATTTAAACTGAATGACCAGACT 1924
 |||
 G: AAAATGGCTTATTTTCATTAATAAAA-CAGTATAACCATTCATTTAAACTGAATGACCAGACT 19225464

T: TGCTGTCTTTAAAAACCCAAACTTGAGATTACCAAAAATTTACAGTATATTTTTACCATT 1864
 |||
 G: TGCTGTCTTTAAAAACCCAAACTTGAGATTAACAAAAA-TTACAGTATATTTTTAACATT 19225523

T: ATACCTGTTAAAAGCTGGTGGGAGTTTTAAAAGTTCATTTTTACAGCTTTTGTAAAGCATA 1804
 |||
 G: ATA-CTGTTAAAAGCTGGTGGGAGTTTTAAAAGTTCATTTTTACAGCTTTTGTAAAGCATA 19225582

T: CAATATTACTTTAAAAAATGACTTTTACTAGGAGATTACGAAAACAGATGTAGGAATG 1744
 |||
 G: CAATATTACTTTAAAAAATGACTTTTACTAGGAGATTACGAAAACAGATGTAGGAATG 19225642

T: TTCCAACCATGGCTTGAAATTATGCATTACGATCCAAGCGAACATCAATTTCTCTGCCAC 1684
 |||
 G: TTCCAACCATGGCTTGAAATTATGCATTACGATCCAAGCGAACATCAATTTCTCTGCCAC 19225702

T: TGATTTTTATGCCATTCAATTTCTGCAGGCTTTTTCAGCTGATTTCTGGGGAGTCAAATC 1624
 |||
 G: TGATTTTTATGCCATTCAATTTCTGCAGGCTTTTTCAGCTGATTTCTGGGGAGTCAAATC 19225762

T: TGACTGTTCACAGCCTTTTGACTTTCCATTCTCCATTTTATTTCTGCAAACATTACAT 1564
 |||
 G: TGACTGTTCACAGCCTTTTGACTTTCCATTCTCCATTTTATTTCTGCAAACATTACAT 19225822

T: GAC.....CACACTGACTGAATTTCTCTTTTAGTTTCTGCCAAGTCAAGT 1519
 |||<<<<<<...<<<<<<|
 G: GACCTGTAA...ACTTACCACACTGACTGAATTTCTCTTTTAGTTTCTGCCAAGTCAAGT 19231827

T: CAAAAGGTAG.....ATTTCTGACAAATATCTGGTTGCCTTTGGAGCCTA 1474
 |||<<<<<<...<<<<<<|
 G: CAAAAGGTAGCTAGAA...GCTTACATTTCTGACAAATATCTGGTTGCCTTTGGAGCCTA 19231951

T: TTCTCTCTCTCATTCCGCTTCCCATTTGGACCCGATAAAAAATCCTCGATCCATATCGATGC 1414
 |||
 G: TTCTCTCTCTCATTCCGCTTCCCATTTGGACCCGATAAAAAATCCTCGATCCATATCGATGC 19232011

T: TCCTTTCCAGTATAGCTCCTATACCTGGTCCCATTTCTATCAAAGCTGGAACTCATCCGGT 1354
 |||
 G: TCCTTTCCAGTATAGCTCCTATACCTGGTCCCATTTCTATCAAAGCTGGAACTCATCCGGT 19232071

T: CCAGTCCCATCCCCATTCCTCCAGTCACACTGTTTCATGTACCCATTCACCAC..... 1300

