Supplementary Materials: Alt-Splice Gene Predictor Using Multitrack-Clique Analysis: Verification of Statistical Support for Modelling in Genomes of Multicellular Eukaryotes

Stephen Winters-Hilt and Andrew J. Lewis

1. Comparative Performance of the Meta-State HMM

The top performing results from the evaluations performed in [14] and [16] are included in Supplementary Table S1 and S2.

The meta-state HMM's (metaHMMs) performance on the ALLSEQ dataset (Supplementary Table S1) clearly exceeds that of the top performing program, GeneID+, cited in [16], by substantial margins, 6.5% and 17%, at the base- and exon-levels, respectively. GeneID+ also uses *extrinsic* information via "amino acid similarity searches" in the process of forming its prediction, whereas the meta-state HMM in this effort uses only the *intrinsic* information contained in the DNA sequence data alone. The Genie, Genescan, and HMMgene programs (shown with an asterisk in Supplementary Table S1) have scores shown on a more challenging dataset than ALLSEQ known as HMR195, but it is still a curated dataset of transcripts each containing one gene per transcript only. Unlike the other methods, the meta-state HMM can also operate on raw genome directly as shown in Supplementary Table S2 (not a specially curated set of gene sequences, one gene per sequence) with similar performance as with the ALLSEQ data (since the HMM hasn't been optimized to the ALLSEQ of HMR195 dataset situation of individual transcripts). For the full genome analysis there aren't repeated statistical measurements of something (the individual sequence prediction), so values are not computed as an expectation on single-gene single transcript sensitivity (sn) and specificity (sp), but simply as the sn and sp for multiple-genes on a single (genome) transcript. Also shown in Supplementary Table S2 are the M and F values (described in Figure S2) that define the footprint state that gave rise to the optimal performance.

Table S1. Gene-predictor results using curated gene sets, including the ALLSEQ dataset and the HMR195 dataset (denoted with *). FGENEH and GeneID+ are the top two performers in [19]. FGENEH and metaHMM, use only intrinsic statitistics (e.g., the ALLSEQ sequences and their annotation and nothing else), while GeneID+, Genie, Genscan, and HMMgene use extrinsic statistics, with the latter three the top performers in [17] using HMR195.

Software	Nucleotide level			Full Exon Level		
Name	E[sn]	E[sp]	E[(sn + sp)/2]	E[SN]	E[SP]	E[(SN + SP)/2]
FGENEH	0.77	0.88	0.825	0.61	0.64	0.64
GeneID+	0.91	0.91	0.91	0.73	0.70	0.71
Genie *	0.91	0.90	0.905	0.71	0.70	0.71
Genscan *	0.95	0.90	0.925	0.70	0.70	0.70
HMMgene *	0.93	0.93	0.93	0.76	0.77	0.76
metaHMM	0.978	0.954	0.966	0.919	0.803	0.861

Table S2. Maximum accuracy of meta-state HMM for the parameter values tested.

Data Set	Nucleotide Level					Full Exon Level				
Name	sn	sp	(sn + sp)/2	M	F	SN	SP	(SN + SP)/2	M	F
ALLSEQ	0.978	0.954	0.966	8	4	0.919	0.803	0.861	8	12
Chr. I–V	0.938	0.864	0.901	5	12	0.775	0.711	0.743	2	20

2. 33-Dimer State Model, and General Footprint Model

Using a single-pass forward/reverse coding Gene Predictor, and using the previously mentioned refinements involving intron frame-pass and end-of-coding stop codon validation states, we now work directly with both forward and reverse states. For reverse states we have exon and intron primitives {A, B, C} and I, respectively, with the 3 possible intron framings for the reverse encoding shown below, where the read direction for 'follows' is reversed (right to left):

jj...je$_C$e$_B$e$_A$…e$_B$i$_A$i$_A$…i$_A$e$_C$…ece$_B$e$_A$jj…j (intron follows exon base with frame C)
jj...je$_C$e$_B$e$_A$…ec$_B$i$_B$…i$_B$e$_A$…ece$_B$e$_A$jj…j (intron follows exon base with frame A)
jj...je$_C$e$_B$e$_A$…e$_A$ic$_C$…ice$_B$…ece$_B$e$_A$jj…j (intron follows exon base with frame B)

There are 16 reverse encoding state transitions in direct correspondence with the 16 non-jj state transitions for the forward read. The jj transition couples the forward and reverse reads in that a forward encoding can 'end', i.e., transition to a region of junk, then eventually transition to a reverse encoded gene. The total number of state-transition (dimer states) in our single pass model is, thus, 33:

13 xx-type (homogeneous) dimers

a. 6 Intron-intron—i$_0$i$_0$,i$_1$i$_1$,i$_2$i$_2$,i$_A$i$_A$,i$_B$i$_B$,i$_C$i$_C$
b. 6 Exon-exon—e$_0$e$_1$,e$_1$e$_2$,e$_2$e$_0$,e$_A$e$_B$,e$_B$e$_C$,e$_C$e$_A$
c. 1 Junk-junk—jj

20 eij-type (heterogeneous) dimers

d. 6 Exon-intron—e$_0$i$_0$,e$_1$i$_1$,e$_2$i$_2$,i$_A$e$_A$,i$_B$e$_B$,i$_C$e$_C$ (with reverse read on reverse encoding states)
e. 6 Intron-exon—i$_0$e$_1$,i$_1$e$_2$,i$_2$e$_0$,e$_B$i$_A$,e$_C$i$_B$,e$_A$i$_C$
f. 6 Exon-junk—(e$_2$j)$_{TAA}$,(e$_2$j)$_{TAG}$,(e$_2$j)$_{TGA}$,(je$_C$)$_{TAA}$,(je$_C$)$_{TAG}$,(je$_C$)$_{TGA}$
g. 2 Junk-exon—(je$_0$),(e$_A$j)

When working with footprint states that are constrained to only have one 'major' eij-transition in a footprint we find that the number of meta-states grows linearly as:

meta-states = # xx-states + # eij-states x (n−1), where n is the size of the footprint (where n = 2 for dimer states shown) [11].

3. Alt-Splice Overlap Encoding Scenarios

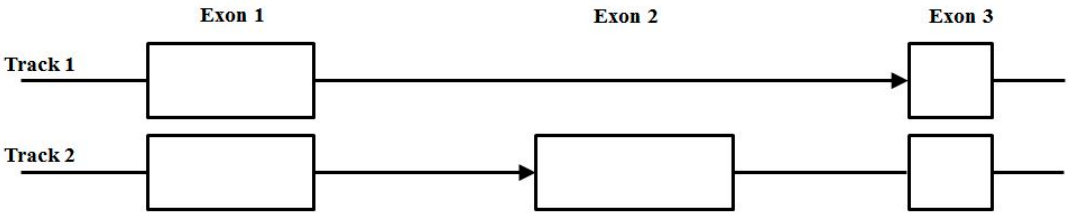
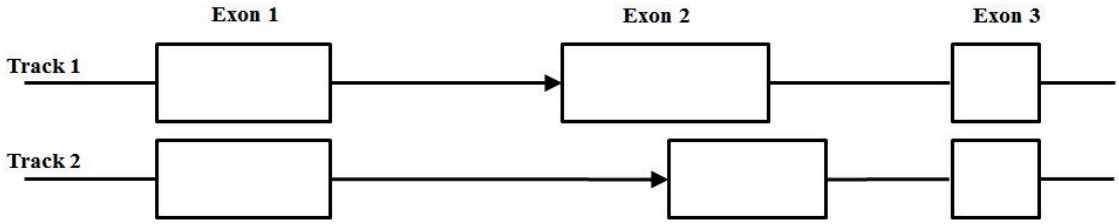


Figure S1. Alt-splice overlap encoding with alternative exon on track 2. The first track describes a transcript with an exon spliced-out with respect to the encoding on track 2.



jjjjjjjjjj0120120120$i_0i_0i_0i_0i_0i_0i_0i_0$120120120120$i_0i_0i_0i_0i_0i_0$012012jjj

jjjjjjjjjj0120120120$i_0i_0i_0i_0i_0i_0i_0i_0$iiiiii120120120$i_0i_0i_0i_0$012012jjj

Figure S2. Alt-splicing involving different exons. V-transitions: $\binom{01}{i1}$… '01i1' shown. The 3-prime splice-site, intron-exon (ie), overlap with exon (ee) transitions, for both track placements and read directions, denoted (3′|e), are: 01i1,12i2, 20i0, i020, i101, i202, AIAC, BIBA, CICB, BABI, CBCI, ACAI. Counts on (3′|e) group to follow. V-transitions: $\binom{0i}{01}$… '0i01' also shown. The 5-prime splice-site, exon-intron (ei), overlap with exon (ee) transitions, for both track placements and read directions, denoted (5′|e), are: 0i01, 1i12, 2i20, 010i, 121i, 202i, IABA, IBCB, ICAC, BAIA, CBIB, ACIC. Counts on (5′|e) group to follow.