Ab initio gene prediction

Genome 559, Winter 2012

Review

- Comparing networks
- Node degree distributions
- Power law distribution $P(k) \propto k^{-c}$ for $k \neq 0, c > 1$
- Network motifs - over and under representation
- Randomizing networks while maintaining node degrees.

Ab initio gene prediction method

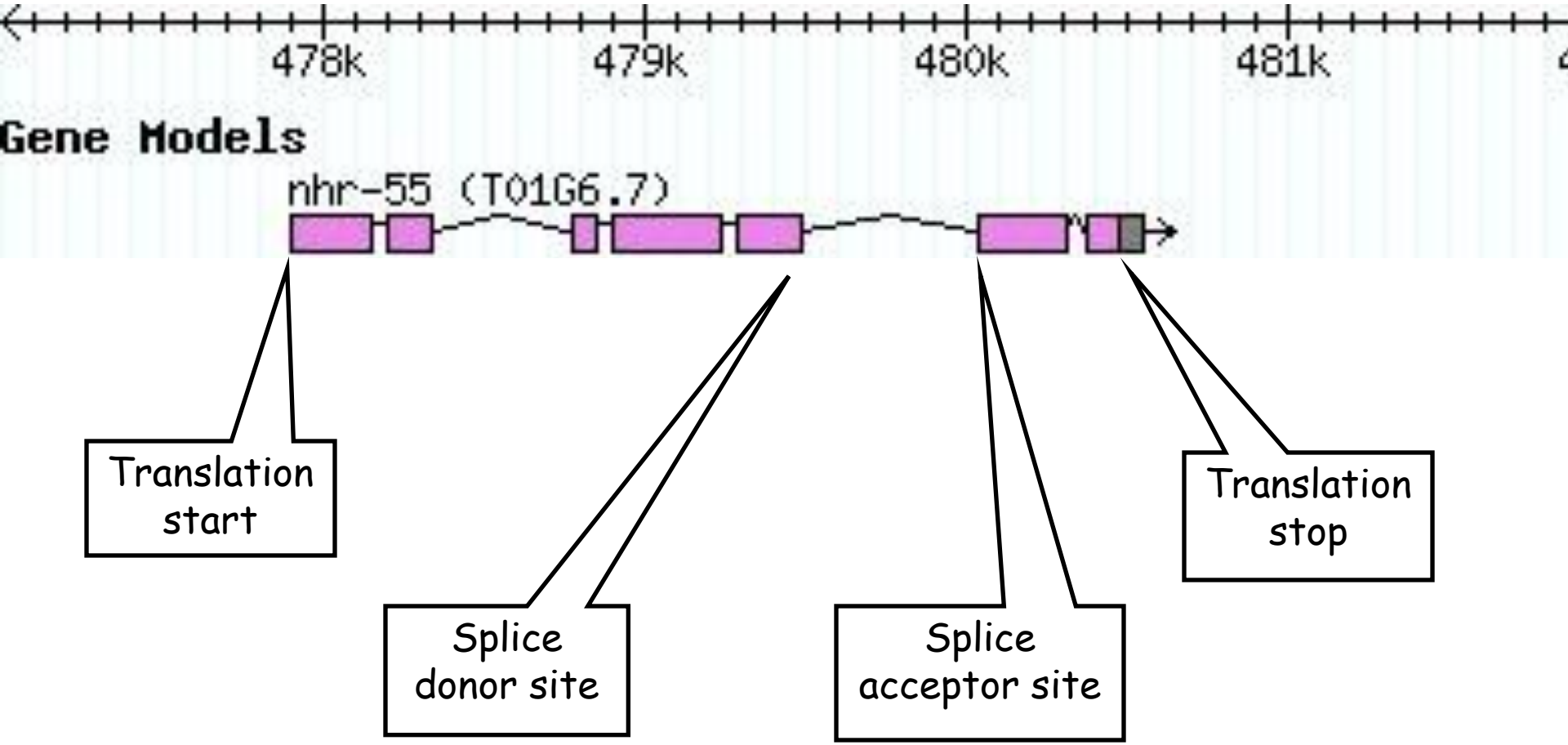
- Define parameters of real genes (based on experimental evidence):

1) Splice donor sequence model
2) Splice acceptor sequence model
3) Intron and exon length distribution
4) Open reading frame requirement in coding exons
5) Requirement that introns maintain reading frame
6) Transcription start and stop models (difficult to predict, often omitted).

- Use those parameters to obtain a best interpretation of genes from any region from genome sequence alone.

ab initio = "from the beginning" (i.e. without experimental evidence)

Sites we might want to predict



(some predictors only deal with coding exons; the 5' and 3' ends are harder to predict.)

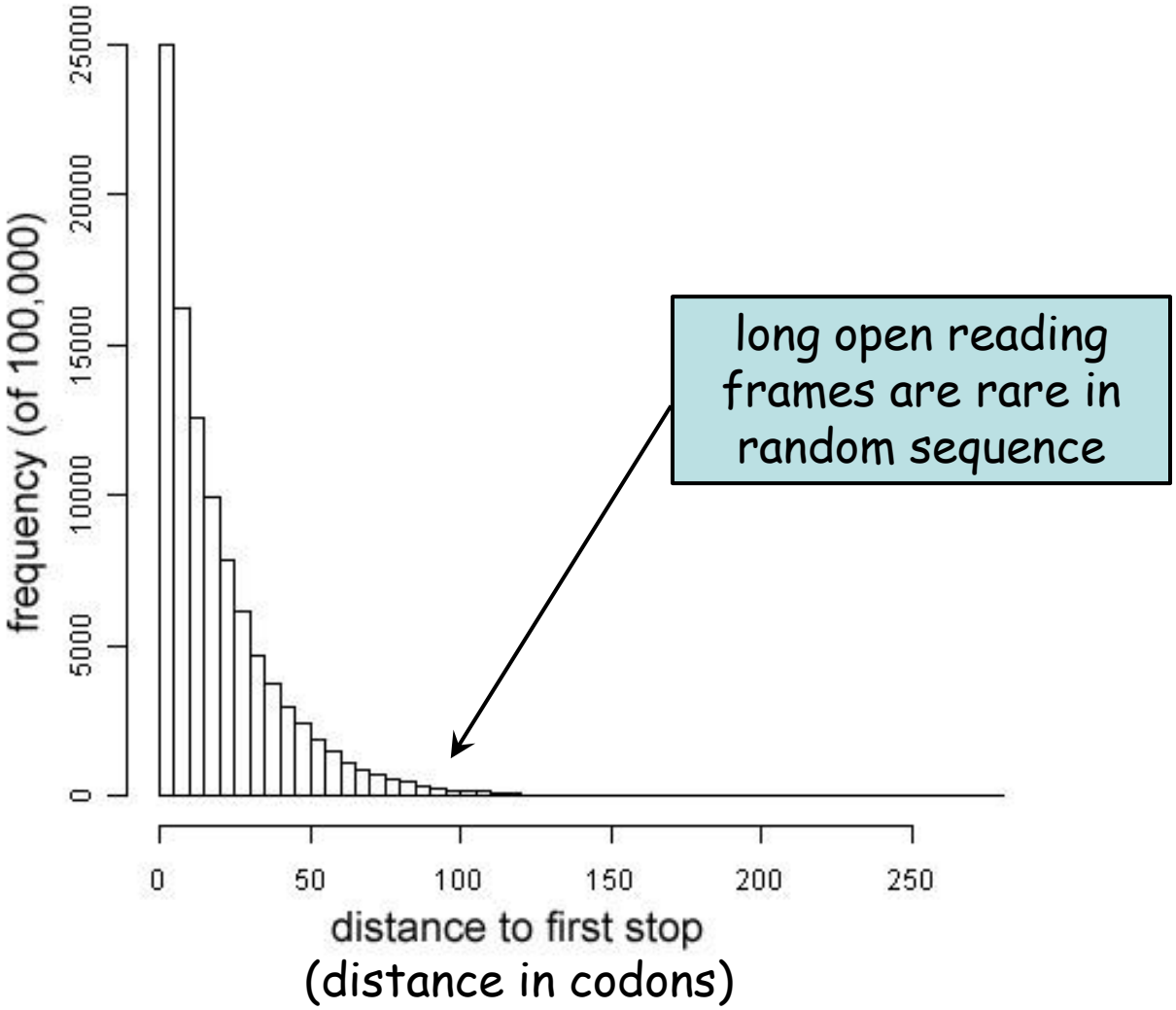
Open reading frames (random sequence)

- 61 of 64 codons are not stop codons (0.953 assuming equal nucleotide frequencies).

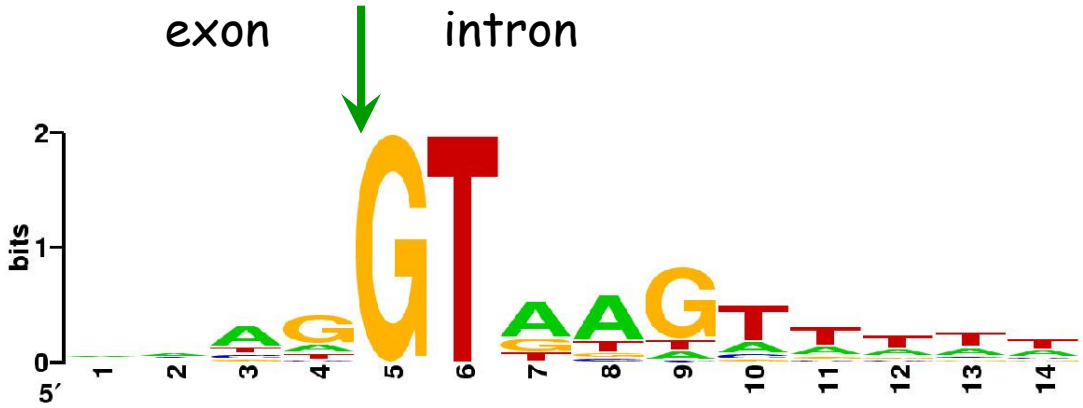
- Probability of not having a stop codon in a particular reading frame along a length L of DNA is a geometric distribution that decays rapidly with L.

- There are 3 reading frames on each DNA strand.

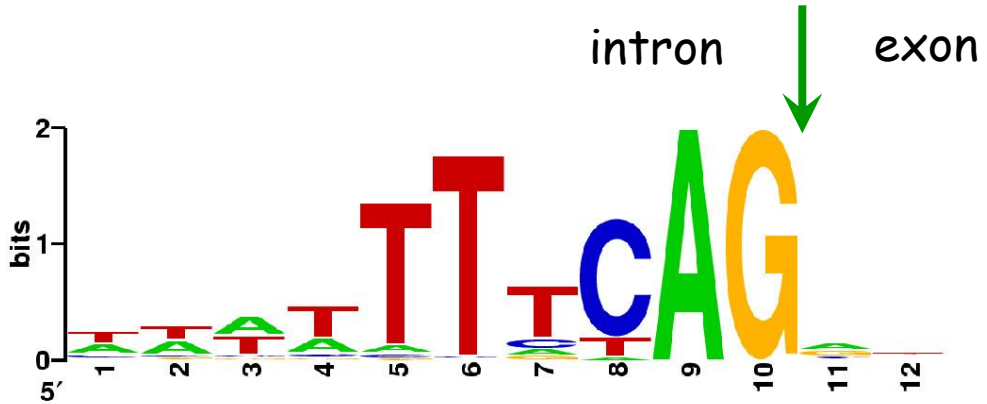
Geometric distribution in random sequence of distance to first stop codon (p=3/64)



Splice donor and acceptor information



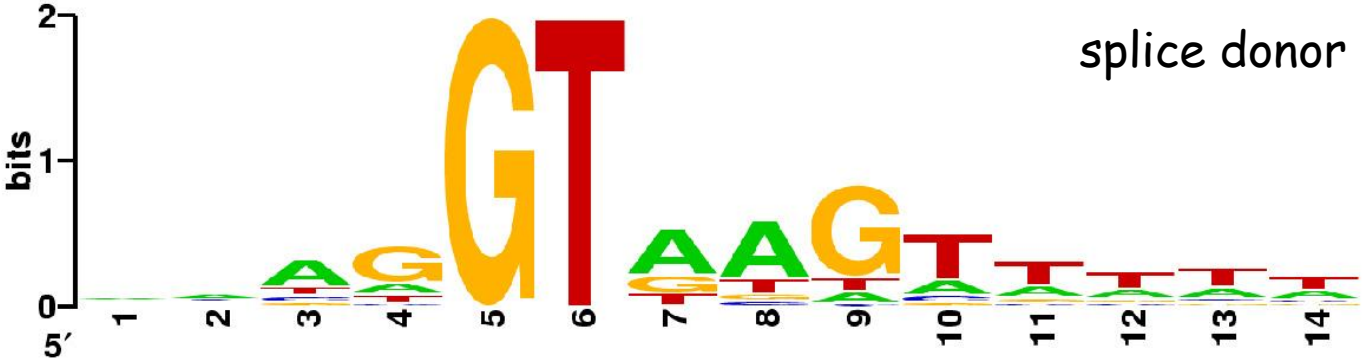
donor, *C. elegans*
(sums to ~8 bits)



acceptor, *C. elegans*
(sums to ~9 bits)

Note – these show a log-odds measure of information content compared to background nucleotide frequencies. Similar to BLOSUM matrix log-odds.

Position Specific Score Matrix (PSSM)

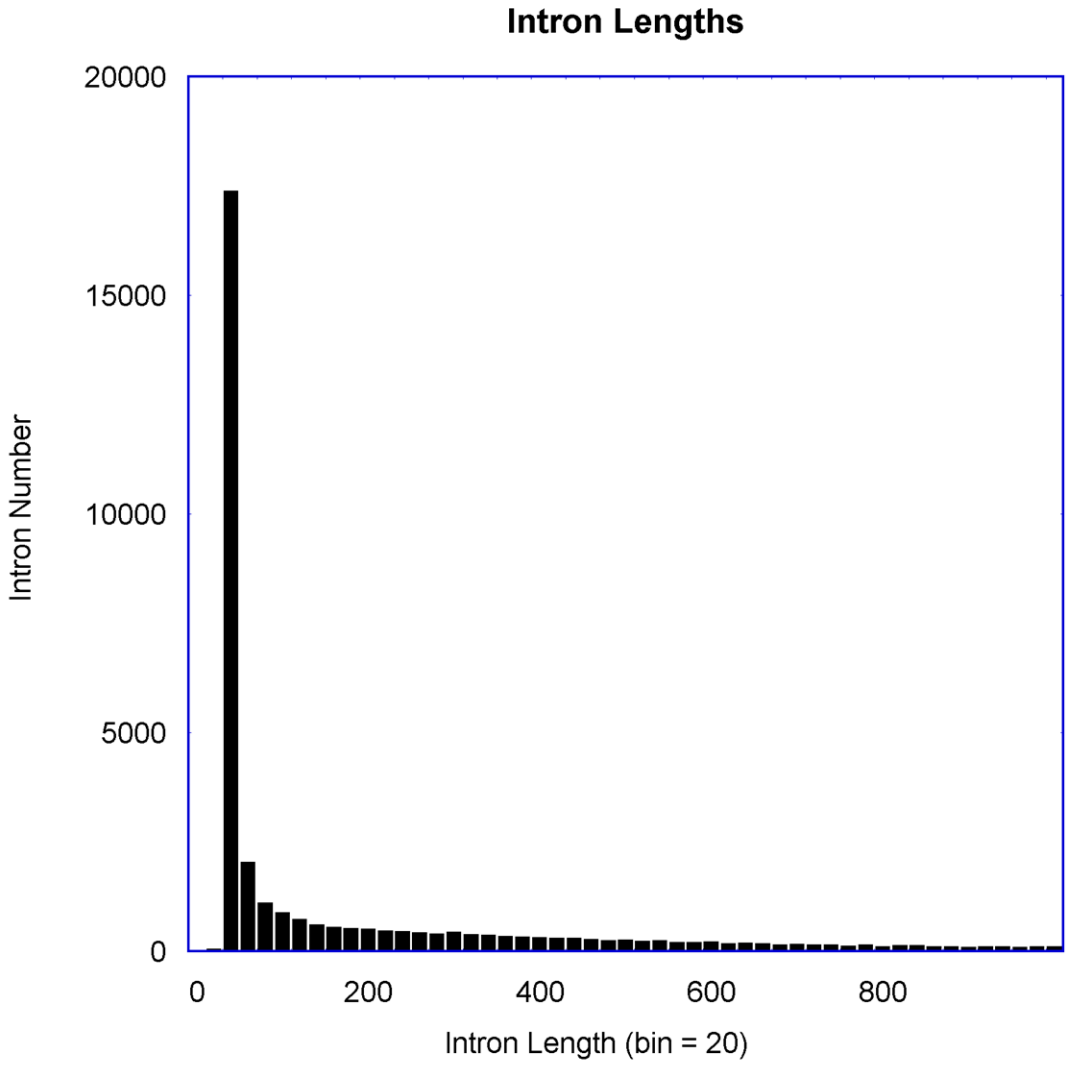


	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	1	1	2	1	0	0	2	2	0	1	1	1	1	1
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	2	8	0	1	0	3	0	0	0	0	0
T	0	0	0	0	0	8	1	1	1	1	2	1	1	1

Slide PSSM along DNA, computing a score at every position.

(this is a conceptual example, the real thing would be computed as log-odds values, similar to BLOSUM matrices)

Intron length distribution (*C. elegans*)



Note: intron length distributions in *Drosophila melanogaster* and *Homo sapiens* (and most other species) are longer and broader.

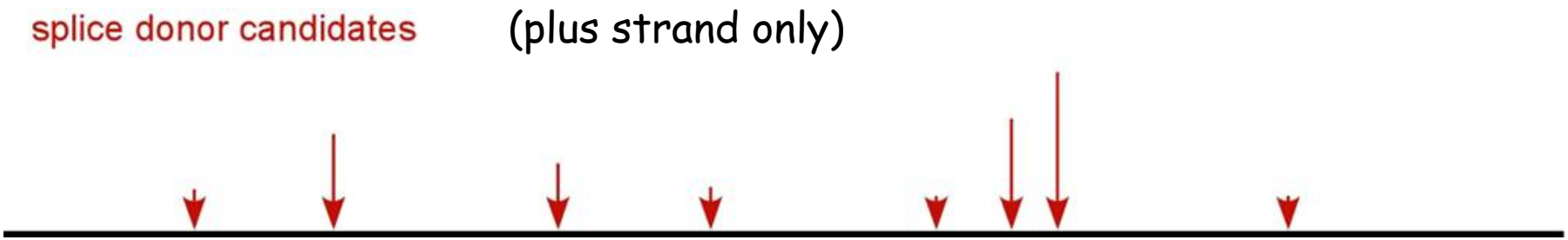
Other information that can be used

- Splice donor and acceptor must be paired and donor must be upstream of acceptor (duh).

- Introns in coding regions must maintain reading frame of the flanking exons.

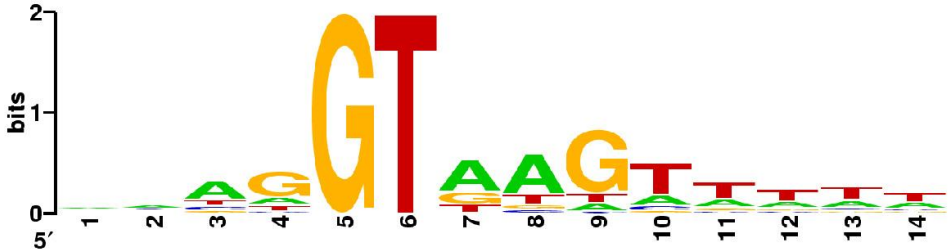
- Nucleotide content analysis (e.g. introns tend to be **AT** rich).

Simple conceptual example



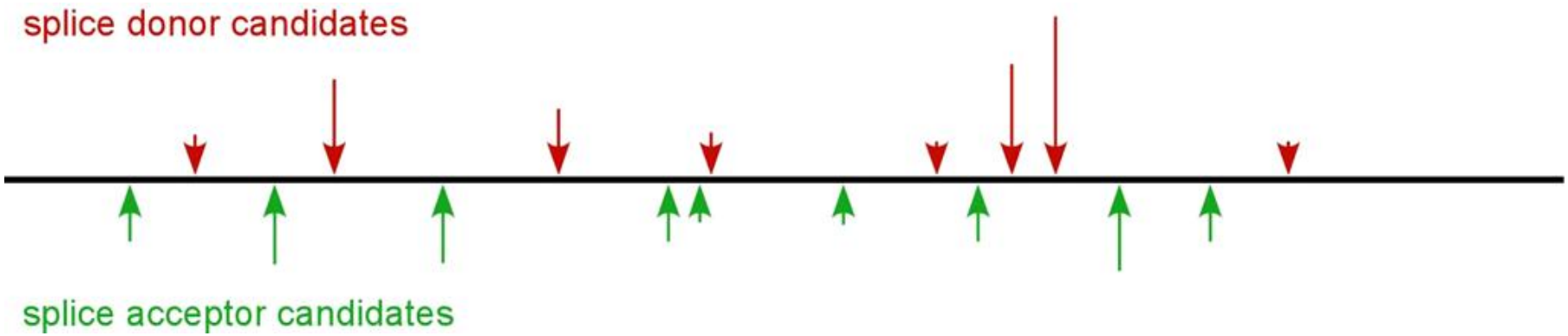
- Sites scored on basis of PSSM matches to known splice donor model (schematized below).

- Arrow length reflects quality of match (worse matches not shown).

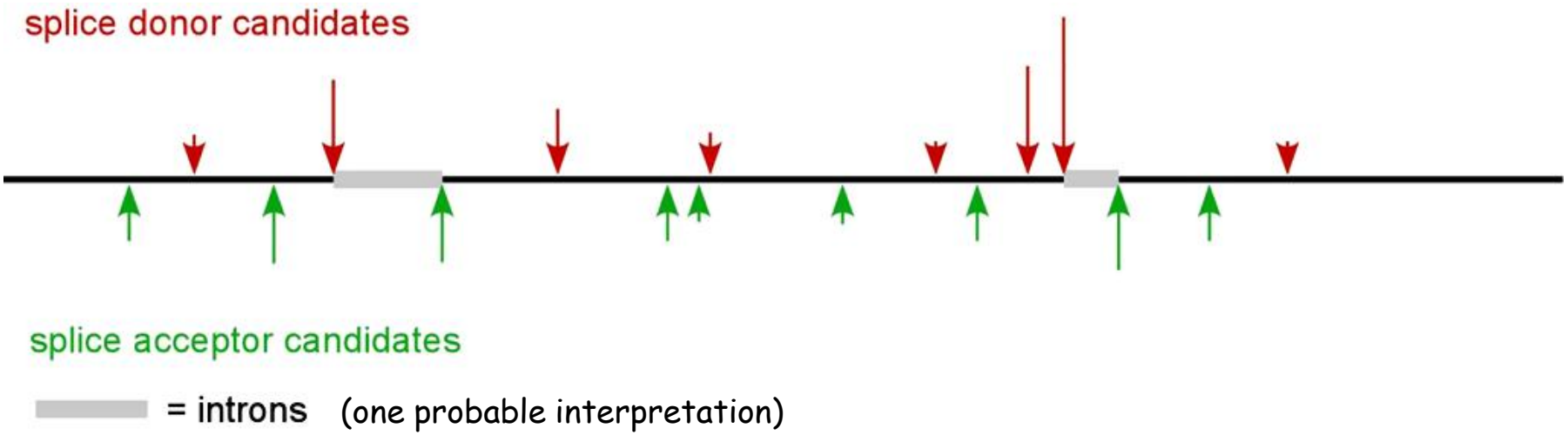


(example cont.)

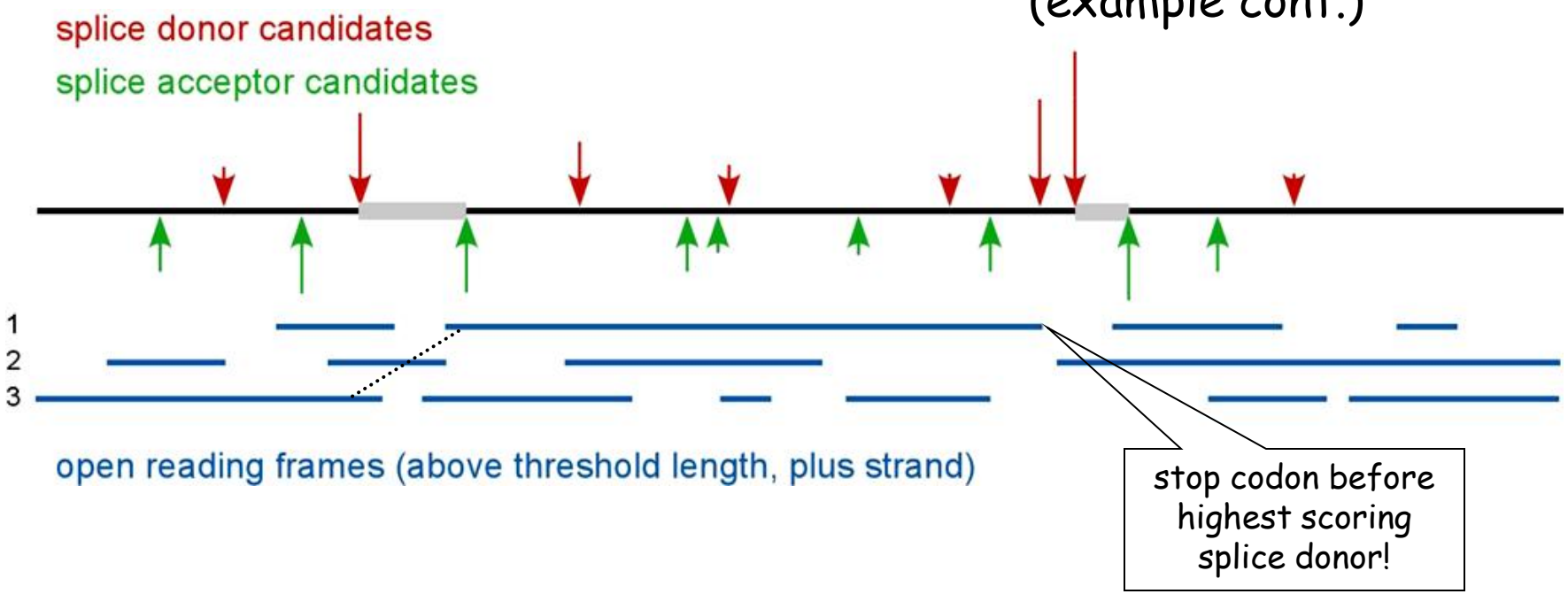
Add splice acceptor information



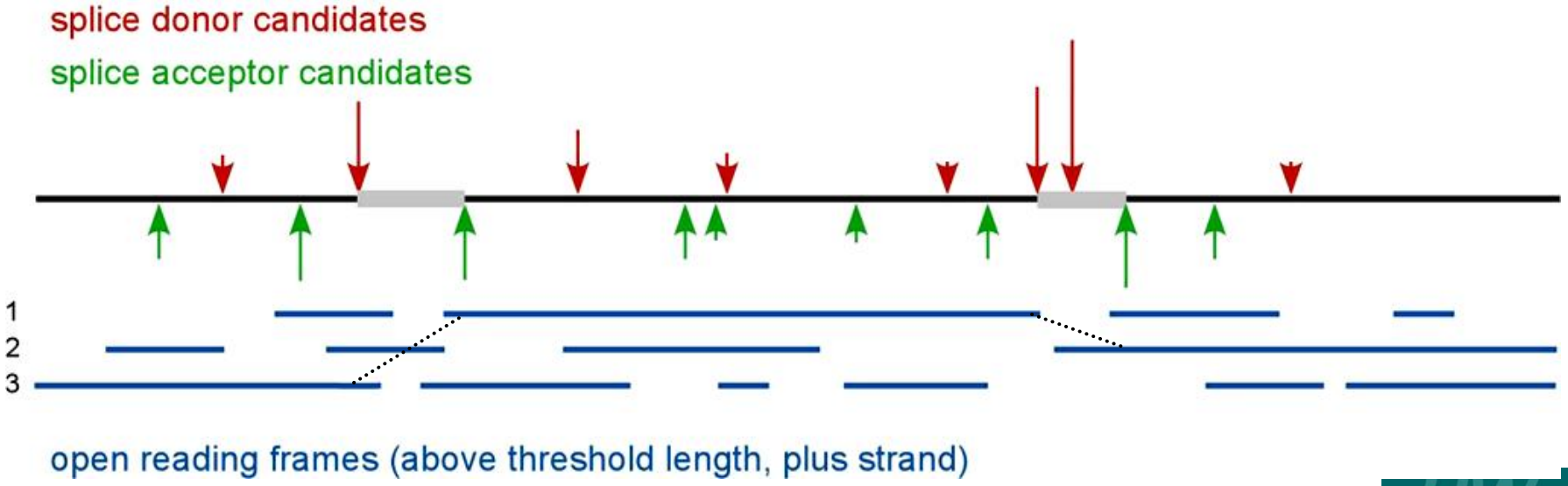
Where would you infer introns?



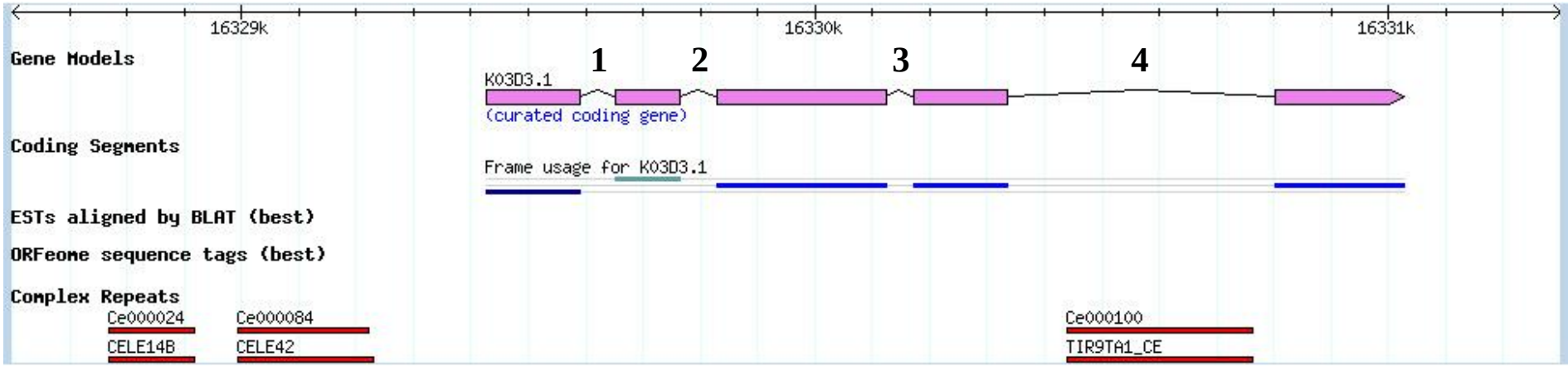
(example cont.)



reinterpreted (avoids stop codon by using lower scoring splice donor):



Real example (end result)



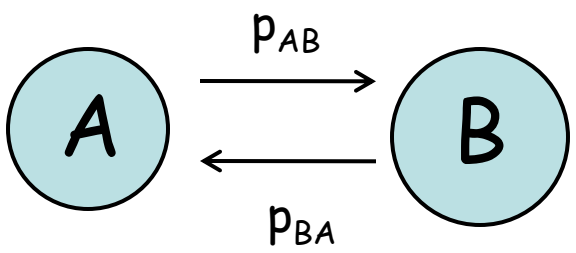
Note that this gene has no mRNA sequences (EST and ORFeome tracks empty). This is a pure *ab initio* prediction.

Hidden Markov Model (HMM)

Markov chain - a linear series of states in which each state is dependent only on the previous state.

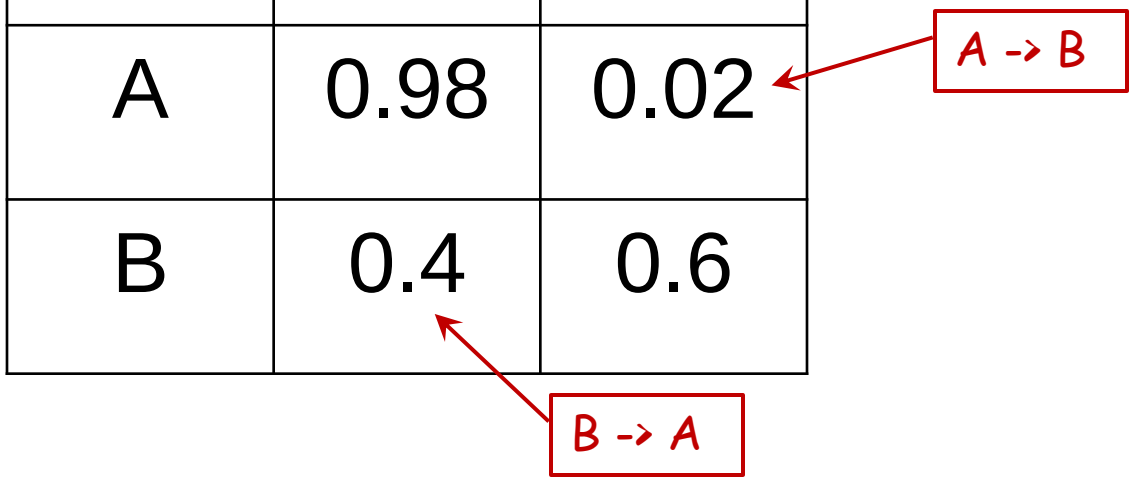
HMM - a model that uses a Markov chain to infer the most likely states in data with unknown states ("hidden" states).

A Markov chain has states and transition probabilities:



(implicitly the probability of staying in state A is 1-p_{AB} and the probability of staying in state B is 1-p_{BA})

	A	B
A	0.98	0.02
B	0.4	0.6



What will the series of states look like (roughly) for this Markov chain?

It will have long stretches of A states, interspersed with short stretches of B states.

Hidden Markov Model

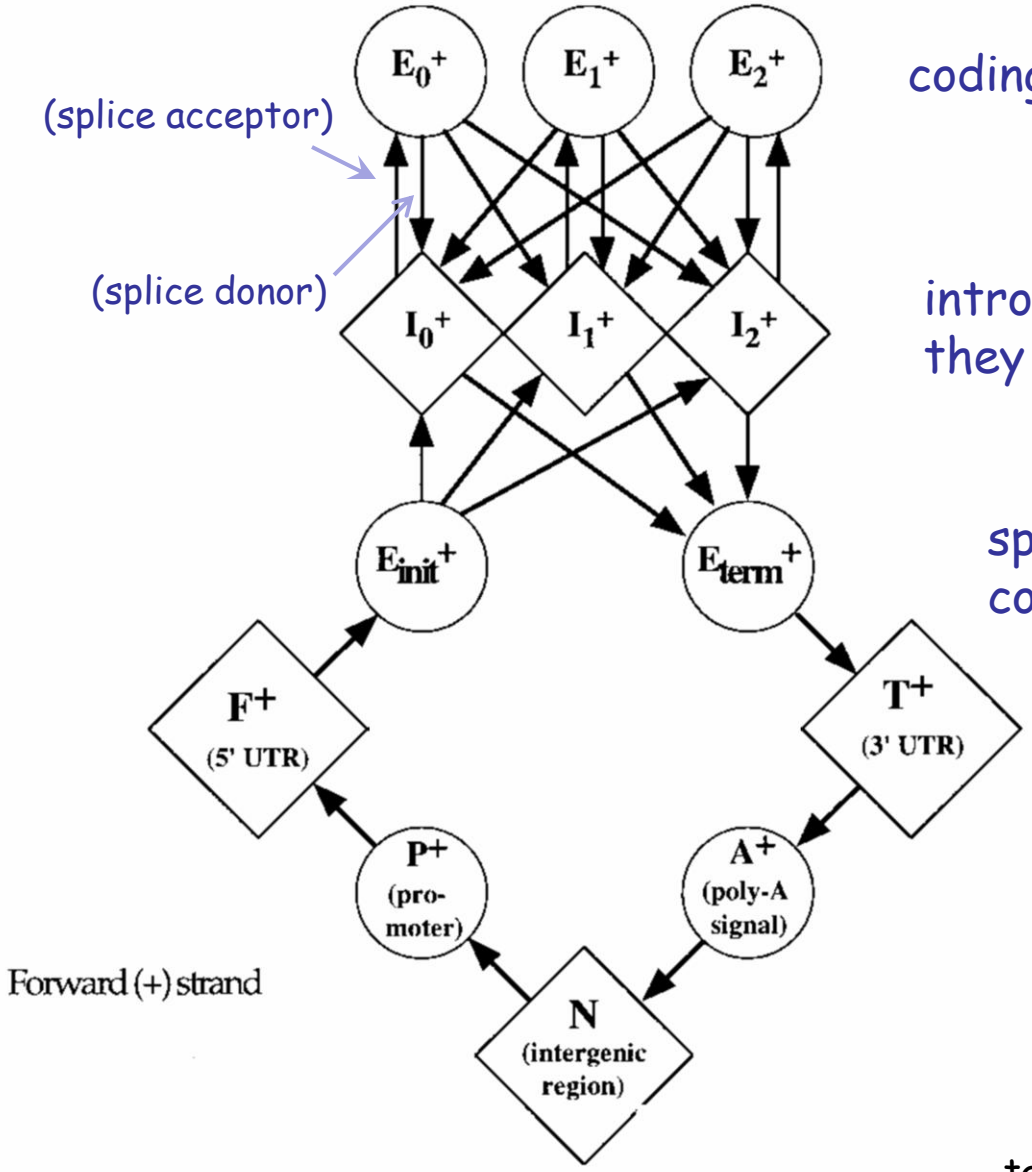
We have a Markov chain with appropriate states and known transition probabilities (e.g. inferred from experimentally known genes).

We have a DNA sequence with unknown states.

Find the series of Markov chain states with the maximum likelihood for the DNA sequence.

Solved with the Viterbi algorithm (we won't cover this, but it is another dynamic programming algorithm). See http://en.wikipedia.org/wiki/Viterbi_algorithm

Gene Prediction HMM States



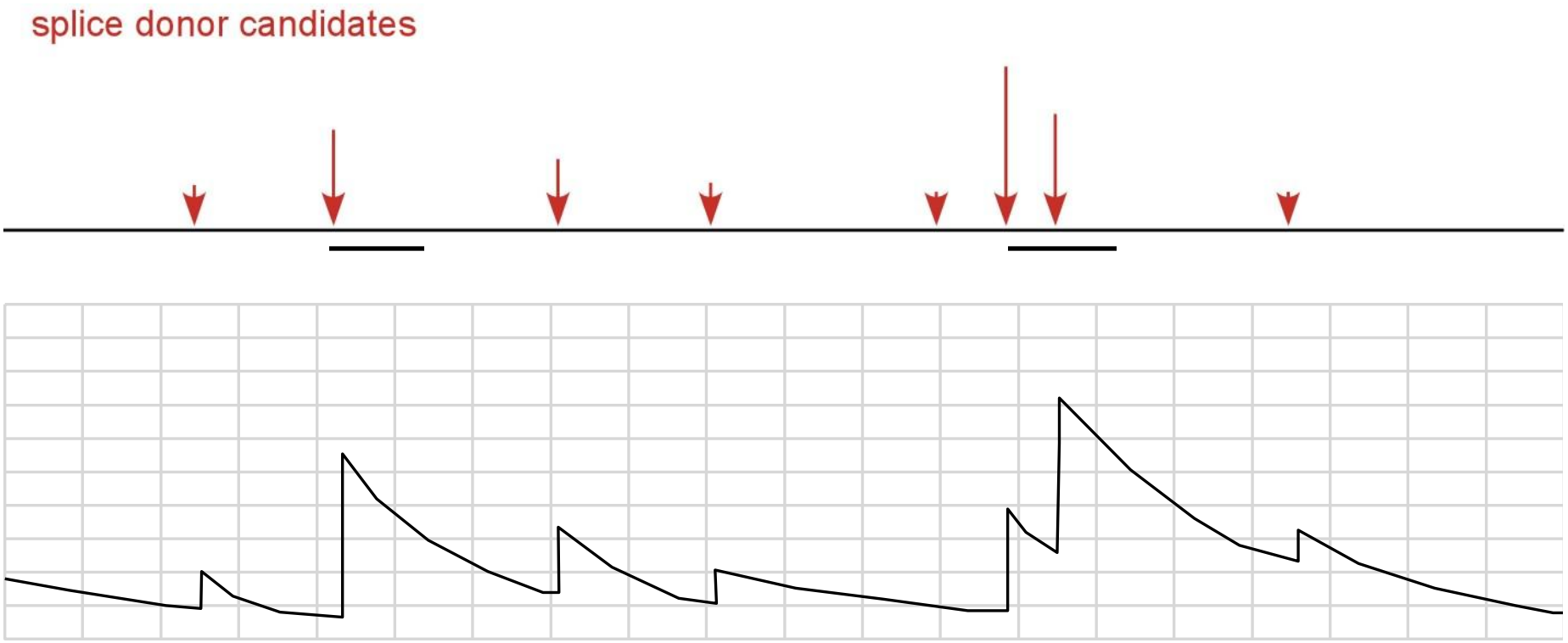
coding exon states (three frames)

intron states (three frames of codon they insert into)

special first (init) and last (term) coding exon states

taken from Stormo lab paper

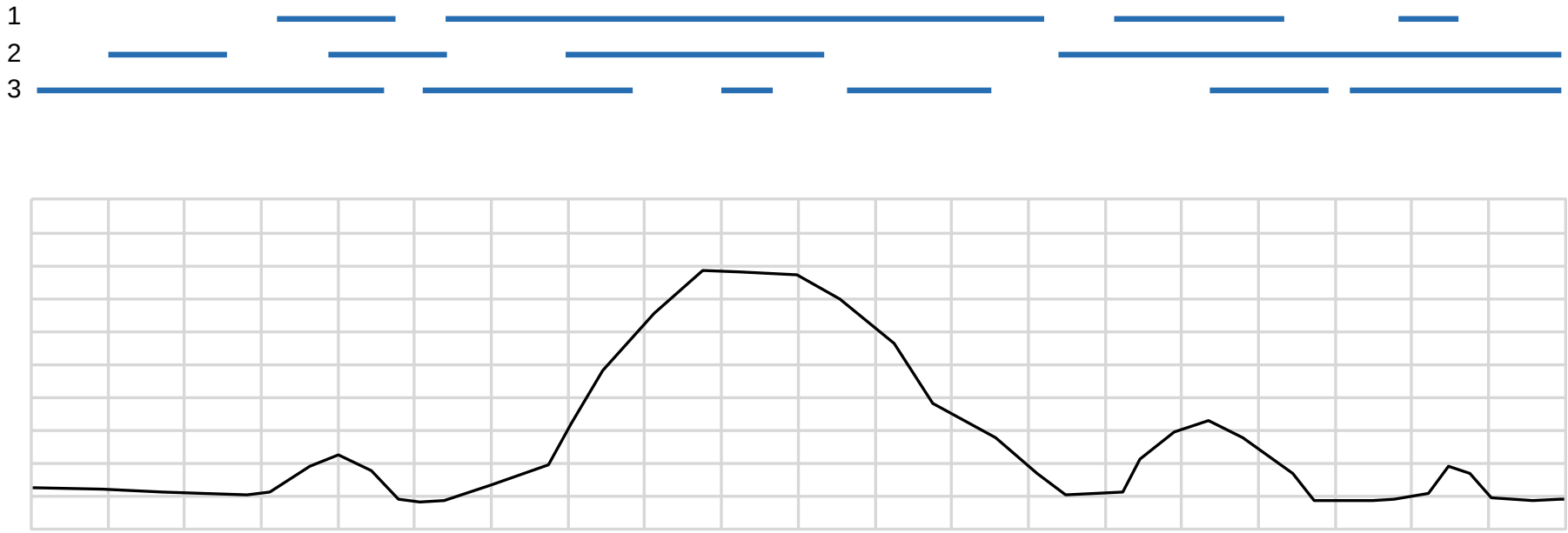
A way to connect the HMM formalism to specifics



probability of being in an intron “state” (based solely on donor sites)

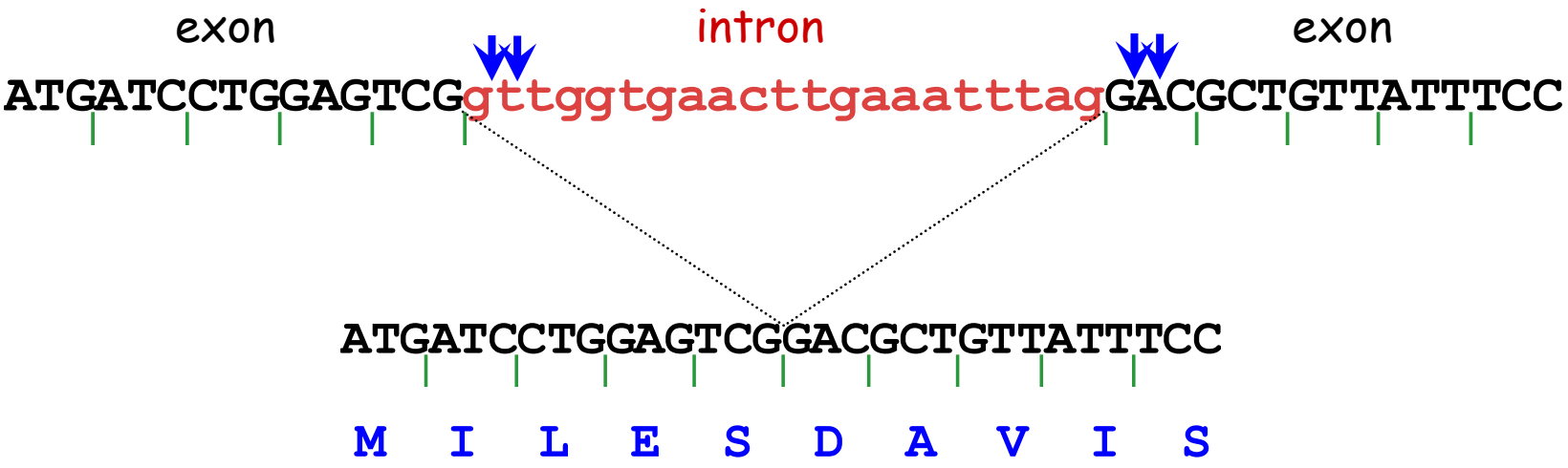
Note – these probabilities are qualitative and are intended only to portray the local trends.

Long open reading frames favor exon state

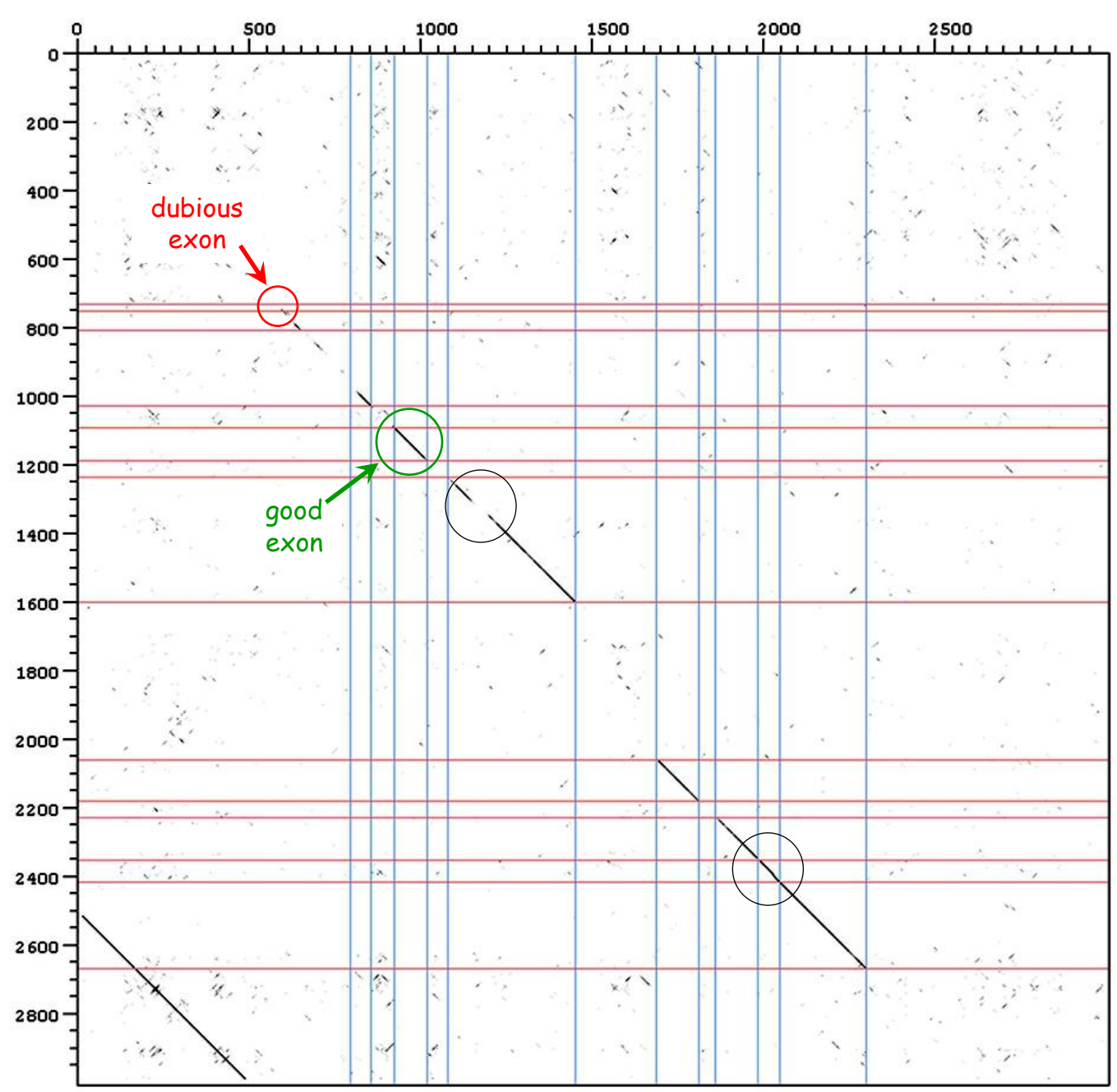


probability of being in an exon “state” (based only on frame 1 ORF)

Intron positions and reading frame



- The intron can be any length and still produce the same exons
- This particular splice is between two codons (0-shifting)
- The splice position can move and maintain coding frame as long as both positions move coordinately.
- If <u>one</u> splice endpoint moves it may change reading frame

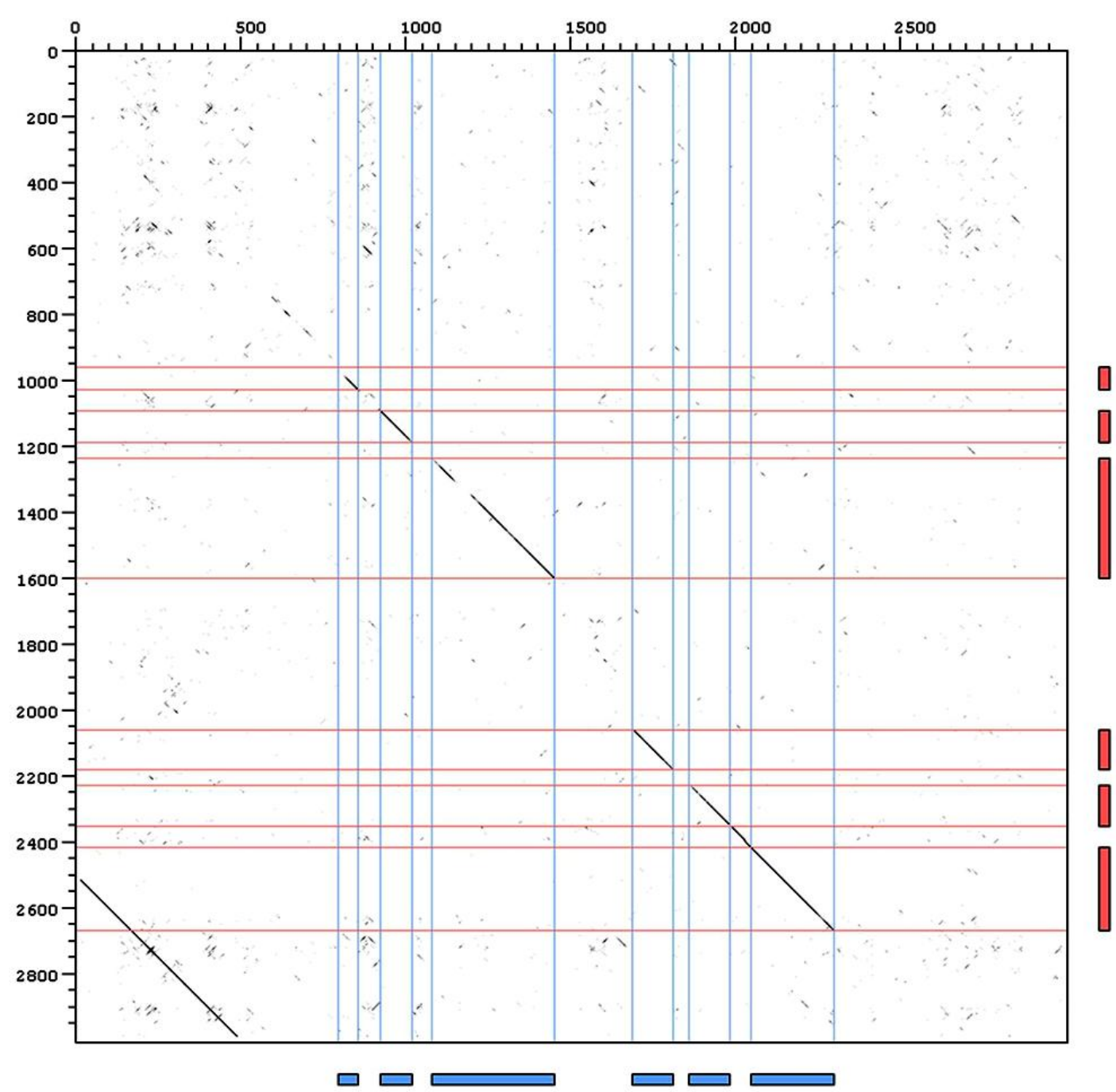


DNA dot matrix comparison of two ab initio gene predictions in related genomes

Gene A (*ab initio* model)

other possible corrections?

Gene B (*ab initio* model)



After correction of exons 1 and 2