Introduction to Computational Biology; An Evolutionary Approach: Gene Prediction

Bernhard Haubold & Thomas Wiehe

Outline

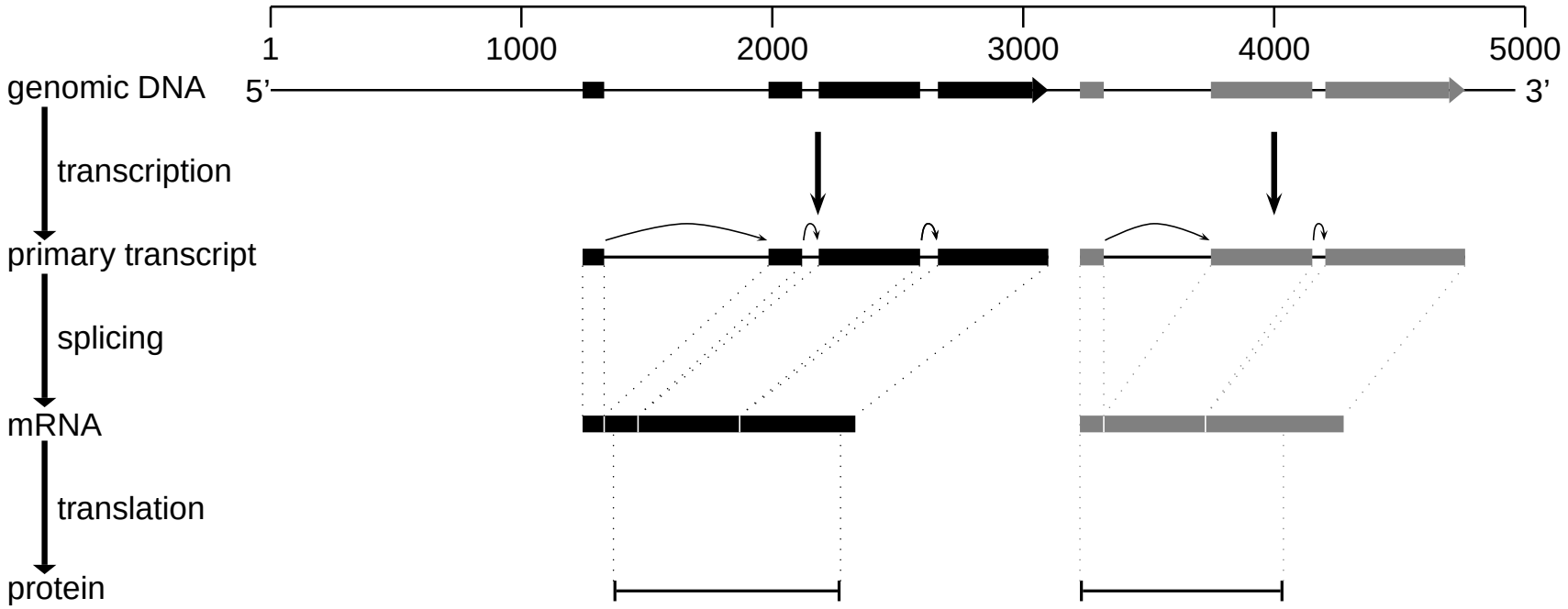
What is a Gene?

1. Mendel 1866
2. $\approx$ 1900: rediscovery of Mendel
3. 1913: first gene map
4. 1927: Muller: X-rays induce mutations
5. 1928: Griffith: transformation with heat-killed cells
6. 1944: Avery: Gene $\equiv$ DNA
7. 1953: Watson & Crick: B-form DNA
8. 1953: Sanger: sequence of insulin
9. 1960's: Nierenberg: Genetic code
10. 1960's/1970's: Arber/Nathans/Smith: discovery of restriction enzymes & gene cloning
11. 1977: Roberts & Sharp: split genes
12. 1977: Sanger / Maxam & Gilbert: DNA sequencing

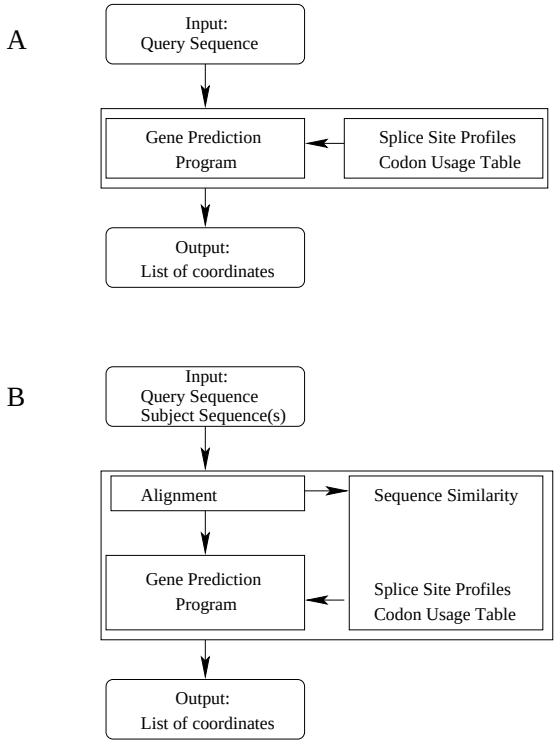
GenBank Entry for *Adh/Adh_dup*

```
LOCUS       DMADH                   4761 bp    DNA     linear   INV 09-MAY-1997
DEFINITION  D.melanogaster Adh and Adh-dup genes.
ACCESSION   X78384
VERSION     X78384.1  GI:483718
KEYWORDS    ADH gene; Adh-dup gene; polymorphism.
SOURCE      Drosophila melanogaster.
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 4761)
  AUTHORS   Kreitman,M. and Hudson,R.R.
  TITLE     Inferring the evolutionary histories of the Adh and Adh-dup loci in
            Drosophila melanogaster from patterns of polymorphism and
            divergence
  JOURNAL   Genetics 127 (3), 565-582 (1991)
  MEDLINE   91200630
   PUBMED   1673107
COMMENT     See paper for sequence comparison to D. simulans.
FEATURES             Location/Qualifiers
     source          1..4761
                     /organism="Drosophila melanogaster"
                     /db_xref="taxon:7227"
                     /chromosome="2"
                     /note="for sequence comparison with D.simulans see paper"
     gene            join(1244..1330,1985..2119,2185..2589,2660..3100)
                     /gene="Adh"
     mRNA            join(1244..1330,1985..2119,2185..2589,2660..3100)
                     /gene="Adh"
                     /note="distal transcript
                     for sequence comparison with D.simulans see paper"
     exon            1244..1330
                     /gene="Adh"
                     /number=1
     intron          1331..1950
                     /gene="Adh"
                     /number=1
     exon            1985..2119
                     /gene="Adh"
                     /number=2
     CDS             join(2021..2119,2185..2589,2660..2926)
                     /gene="Adh"
                     /note="unnamed protein product"
                     /codon_start=1
                     /protein_id="CAA55151.1"
                     /db_xref="GI:2077989"
                     /db_xref="FLYBASE:FBgn0000055"
                     /db_xref="SWISS-PROT:P00334"
                     /translation="MSFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENP
                     AAIAELKAINPKVTVTFYPYDVTVPIAETTKLLKTIFAQLKTVDVLINGAGILDDHQI
                     ERTIAVNYTGLVNTTTAILDFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAV
                     VNFTSSLAKLAPITGVTAYTVNPGITRTTLVHKFNSWLDVEPQVAEKLLAHPTQPSLA
                     CAENFVKAIELNQNGAIWKLDLGTLEAIQWTKHWDSGI"
     intron          2120..2184
                     /gene="Adh"
                     /number=2
     exon            2185..2589
                     /gene="Adh"
                     /number=3
     intron          2590..2659
                     /gene="Adh"
                     /number=3
     exon            2660..3100
                     /gene="Adh"
                     /number=4
     gene            3226..4761
                     /gene="Adh-dup"
     mRNA            join(<3226..3321,3748..4152,4204..4761)
                     /gene="Adh-dup"
                     /note="for sequence comparison with D.simulans see paper"
     CDS             join(3226..3321,3748..4152,4204..4521)
                     /gene="Adh-dup"
                     /note="for sequence comparison with D.simulans see paper"
                     /codon_start=1
                     /protein_id="CAA55152.1"
                     /db_xref="GI:483719"
                     /db_xref="FLYBASE:FBgn0000056"
                     /db_xref="SWISS-PROT:P91615"
                     /translation="MFDLTGKHVCYVADCGGIALETSKVLMTKNIAKLAILQSTENPQ
                     AIAQLQSIKPSTQIFFWTYDVTMAREDMKKYFDEVMVQMDYIDVLINGATLCDENNID
                     ATINTNLTGMMNTVATVLPYMDRKIGGTGGLIVNVTSVIGLDPSPVFCAYSASKFGVI
                     GFTRSLADPLYYSQNGVAVMAVCCGPTRVFVDRELKAFLEYGQSFADRLRRAPCQSTS
                     VCGQNIVNAIERSENGQIWIADKGGLELVKLHWYWHMADQFVHYMQSNDEEDQD"
     exon            <3226..3321
                     /gene="Adh-dup"
                     /number=1
     intron          3322..3747
                     /gene="Adh-dup"
                     /number=1
     exon            3748..4152
                     /gene="Adh-dup"
                     /number=2
     intron          4153..4203
                     /gene="Adh-dup"
                     /number=2
     exon            4204..4761
                     /gene="Adh-dup"
                     /number=3
BASE COUNT     1417 a   1007 c    989 g   1348 t
ORIGIN
        1 tgtattttcc aattaggtga tagaacttgt gtgcacacac acatatagtt ctatatcaac
       61 aaacaggttt aagttttatg caaattgaaa gcttatttct tccgcatgct tatctctttc
      121 cttctcatca tttgtatgca aaaaatacat atgaatttgc agtagcctcc tcccacatca
      181 tatttaacgc cctatattca aaatttgctc aagaaaatat ttgaaccaaa ttgattttta
      241 gtcaattagt ttttaagtaa ttaagtggag taaacatata caatttttatt cttaccaaac
      301 acatatactc atatattttg aataaataaa taaacaaata tatataaaat ctacgaaatt
      361 ggcaaacaaa ttttaaagca ttatagtatt gccgatttaa ttaatataat taaataatat
      421 gtacatgtat taatcttgtg tgcgagcatg ggttaaatct agctgcattc gaaaccgcta
      ... .......... .......... .......... .......... .......... ..........
     2101 gctcaagcgc gatctgaagg taactatgcg atgcccacag gctccatgca gcgatggagg
     2161 ttaatctcgt gtattcaatc ctagaacctg gtgatcctcg accgcattga gaacccggct
     2221 gccattgccg agctgaaggc aatcaatcca aaggtgaccg tcaccttcta cccctatgat
     2281 gtgaccgtgc ccattgccga gaccaccaag ctgctgaaga ccatcttcgc ccagctgaag
     2341 accgtcgatg tcctgatcaa cggagctggt atcctggacg atcaccagat cgagcgcacc
     2401 attgccgtca actacactgg cctggtcaac accacgacgg ccattctgga cttctgggac
     2461 aagcgcaagg gcggtcccgg tggtatcatc tgcaacattg gatccgtcac tggattcaat
     2521 gccatctacc aggtgcccgt ctactccggc accaaggccg ccgtggtcaa cttcaccagc
     2581 tccctggcgg taagttgatc aaaggaaacg caaagttttc aagaaaaaac aaaactattt
     2641 gattttataa cacctttaga aactggcccc cattaccggc gtgaccgctt acaccgtgaa
     2701 ccccggcatc acccgcacca ccctggtgca caagttcaac tcctggttgg atgttgagcc
     2761 ccaggttgct gagaagctcc tggctcatcc cacccagcca tcgttggcct gcgccgagaa
     2821 gttgtcagtt gcagttcagc agacgggcta acgagtactt gcatctctc ..........
```

Structure of *Adh/Adh_dup*

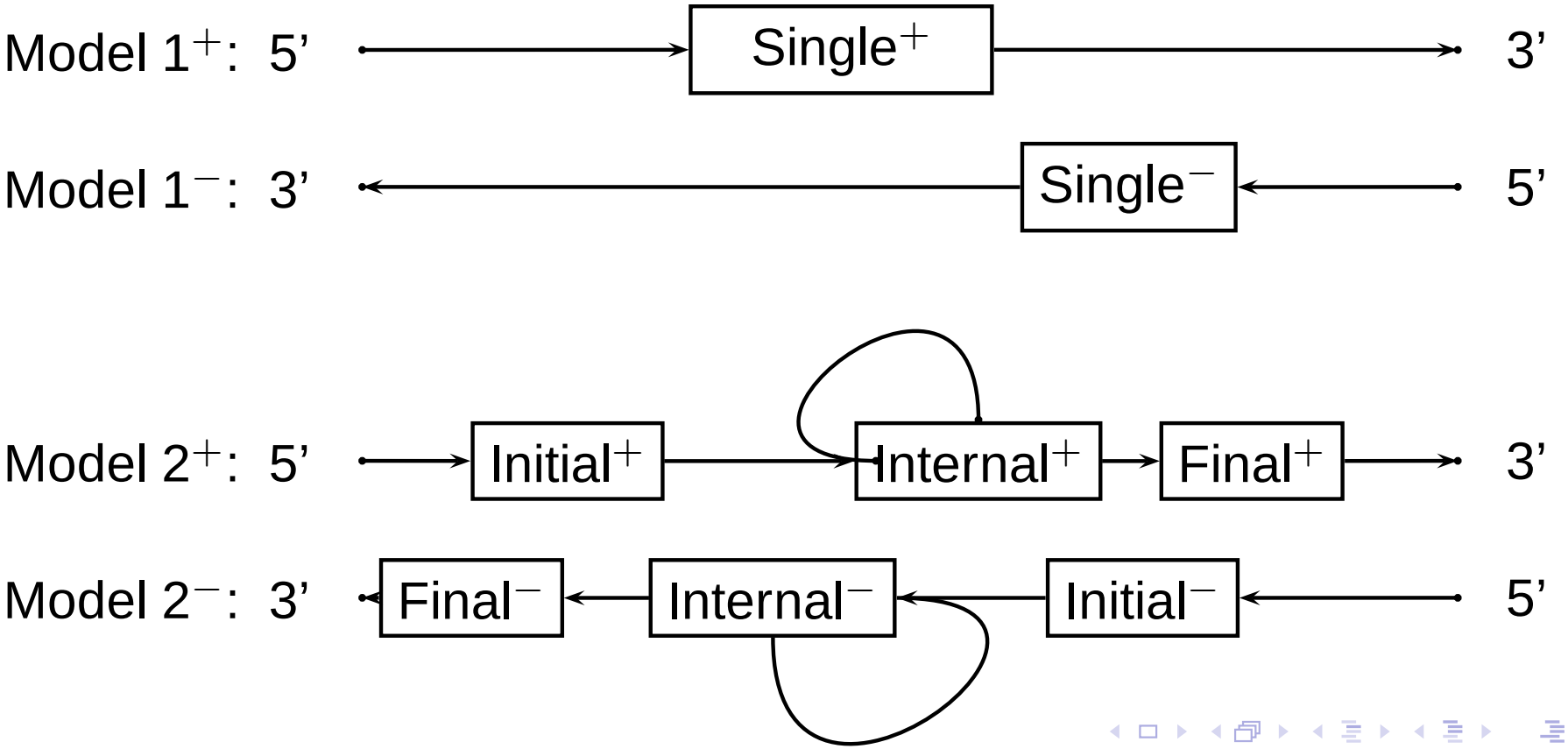


Gene Prediction Algorithms



A: *Ab initio*; B: comparative

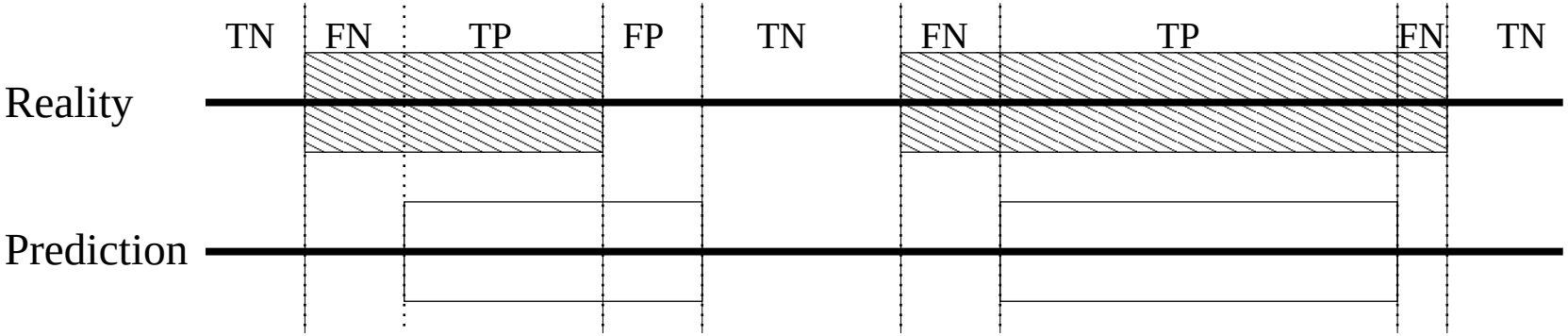
Gene Models


Model 1+: 5'
Single+
3'
Model 1−: 3'
Single−
5'
Model 2+: 5'
Initial+
Internal+
Final+
3'
Model 2−: 3'
Final−
Internal−
Initial−
5'

Sensitivity & Specificity

- *sensitivity* (S_n): proportion of real elements (coding nucleotides, exons or genes) that have been correctly predicted
- *specificity* (S_p): proportion of predicted elements which are correct

Accuracy at Nucleotide Level



- TP: true positive; TN: true negative; FP: false positive; FN: false negative
- Sensitivity: $S_n(N) = \frac{\mathrm{TP}(N)}{\mathrm{TP}(N)+\mathrm{FN}(N)}$
- Specificity: $S_p(N) = \frac{\mathrm{TP}(N)}{\mathrm{TP}(N)+\mathrm{FP}(N)}$

Sensitivity & Specificity

1. $0 \leq S_p \leq 1; 0 \leq S_n \leq 1$
2. Ideal prediction program: $S_n = S_p = 1$
3. Summary measure: Correlation coefficient:

$$\mathrm{CC} = \frac{\mathrm{TP}(N) \times \mathrm{TN}(N) - \mathrm{FP}(N) \times \mathrm{FN}(N)}{\sqrt{(\mathrm{TP}(N) + \mathrm{FP}(N))(\mathrm{TN}(N) + \mathrm{FN}(N))(\mathrm{TN}(N) + \mathrm{FP}(N))(\mathrm{TP}(N) + \mathrm{FN}(N))}}$$

4. $-1 \leq \mathrm{CC} \leq 1$
5. $\mathrm{CC} = 1$: perfect prediction
6. $\mathrm{CC} = -1$: each coding nucleotide predicted as non-coding & vice versa

Accuracy at Exon Level

- true positive: if 5' & 3' boundaries predicted correctly
- false positive: if no overlap with real exon
- sensitivity: proportion of true positives among true exons
- specificity: proportion of true positives among predicted exons
- summary statistic: $\text{sensitivity} + \text{specificity}/2$

Accuracy at Gene Level

- true positive
  1. all coding exons identified
  2. every intron/exon boundary correct
  3. all exons included
- false negative (missed): no exons overlap with any predicted gene
- false positive (wrong): none of its exons overlapped by any real gene

Summary Accuracy Measures

	nucleotide	exon	gene
Sensitivity S_n	proportion of coding nucleotides correctly predicted as coding	proportion of correctly predicted exons among actual exons	proportion of completely correctly predicted genes among actual genes
Specificity S_p	proportion of nucleotides predicted as coding and which are actually coding	proportion of correctly predicted exons among all predicted exons	proportion of completely correctly predicted genes among all predicted genes

Examples Gene Prediction Accuracy

	Nucleotide		Exon				Gene					
	S_n	S_p	S_n	S_p	FN(E)	FP(E)	S_n	S_p	FN(G)	FP(G)	SG	JG
(1)	0.89	0.77	0.65	0.49	0.10	0.32	0.30	0.27	0.09	0.24	1.10	1.06
(2)	0.86	0.83	0.58	0.34	0.21	0.47	0.26	0.10	0.14	0.30	1.06	1.11
(3)	0.96	0.92	0.70	0.57	0.08	0.17	0.40	0.29	0.05	0.11	1.17	1.08
(4)	0.97	0.91	0.77	0.55	0.05	0.20	0.44	0.28	0.05	0.13	1.15	1.09
((5)	0.81	0.86	0.42	0.41	0.24	0.29	0.14	0.12	0.16	0.24	1.23	1.08
(6)	0.97	0.91	0.68	0.53	0.05	0.20	0.35	0.30	0.07	0.15	1.04	1.12
(7)	0.96	0.63	0.63	0.41	0.12	0.50	0.33	0.21	0.05	0.55	1.22	1.06

(1) FGENES, (2) GeneID, (3) GENIE, (4) GENIE EST-version, (5) GRAIL, (6) HMMGENE, and (7) MAGPIE

Additional Accuracy Measures

1. *ab initio* prediction: poor prediction of initial & terminal exons $\rightsquigarrow$
   1. chimeric predictions
   2. split predictions
2. Additional measures:
   1. joined genes, JG: predicted genes overlapping real genes / number of joined predictions
   2. split genes, SG: predicted genes overlapping real genes / number of split precdictions

Principle

- Mimic the cell's transcription, splicing, & translation machinery
- Look for signals:
  1. Transcription signals
  2. Splicing signals
  3. Translation signals
- Look at sequence composition
  1. GC-composition
  2. Codon usage

Transcription Signals

1. Transcription start site (TSS): CAP signal, single purine (A/G)
2. TATA-box: AT-rich region of 6 bp about 30 bp upstream from TSS
3. Transciption termination:
   1. polyadenylation signal: consensus AATAAA hexamer
   2. degenerate signal 20-30 bp downstream from polyadenylation signal

NB:

- only $\approx$ 70% of human promoters contain TSS & TATA box
- polyadenylation signal absent from 50% of genes

Therefore: hard to find beginning & end of gene.

Splicing Signals

1. donor site (5' end of intron): GT
2. acceptor site (3' end of intron): AG plus upstream polypyrimidine tract
3. branch point:
   - close to 3' end of intron, upstream of poly-pyrimidine tract
   - degenerate motif always containing A; mammalian consensus: YNYTRAY

Non-standard donor dinucleotide: GC; $< 1\%$ in mammals

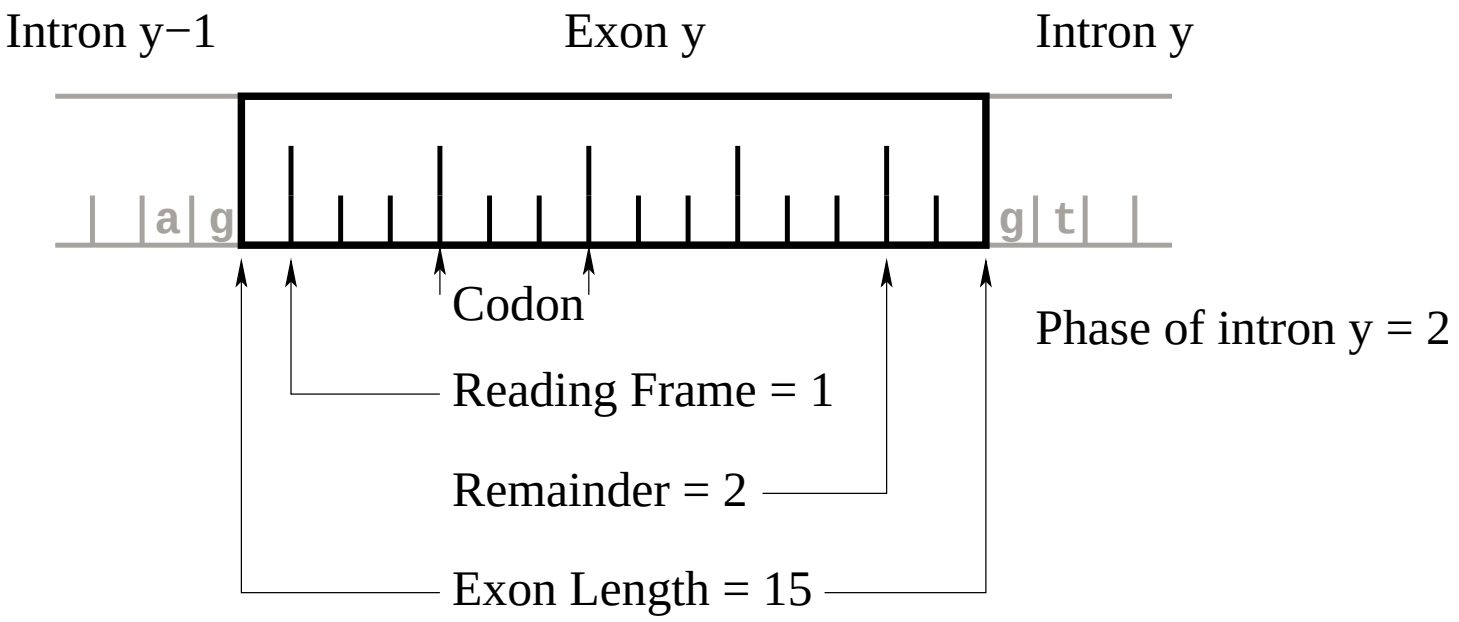
Translation Signals

1. Kozak signal: `gccaccATGG`
2. stop codon: `TA[AG]`, `TGA`
3. rare contexts: `TGA` encodes selenocysteine

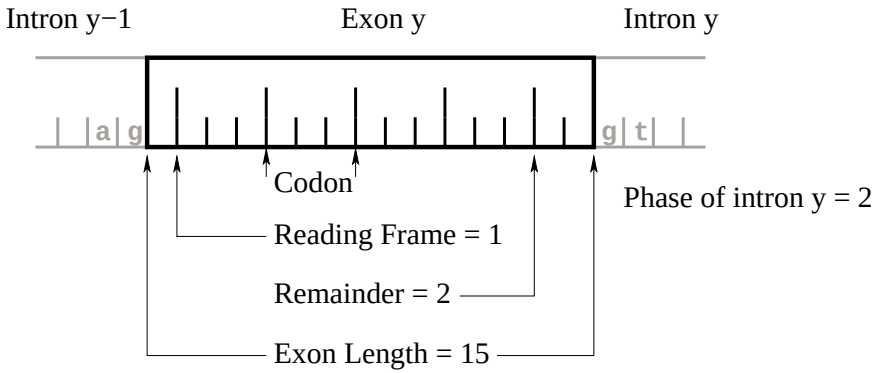
Programs that use signals for gene prediction: *search by signal*.

Reading Frame

- first translated exon: start codon ATG sets reading frame
- internal exon: offset at beginning of exon when exon is translated



Reading Frame of Internal Exon



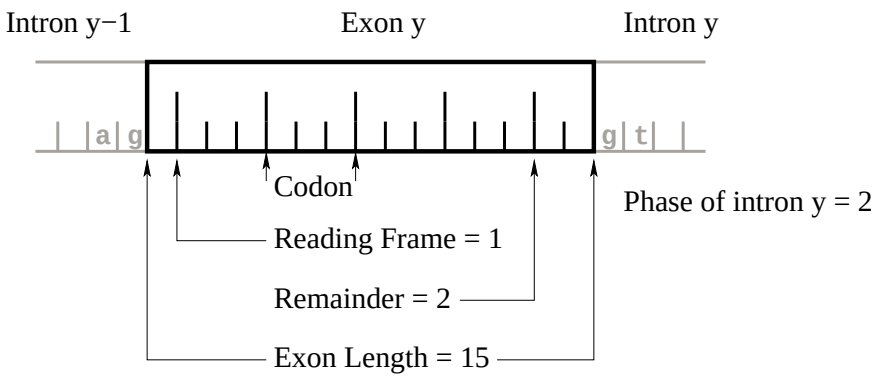
$$\begin{aligned} \text{frame}(1) &= 0 \\ \text{frame}(y) &= (3 + \text{frame}(y-1) - \text{length}(y-1)\%3)\%3\,, \end{aligned}$$

where $y > 1$ is y-th exon.

Reading Frame Scheme

$\text{length}(y-1)\%3 \rightarrow$ $\text{frame}(y-1) \quad \downarrow$	0	1	2
0	0	2	1
1	1	0	2
2	2	1	0

Intron Phase

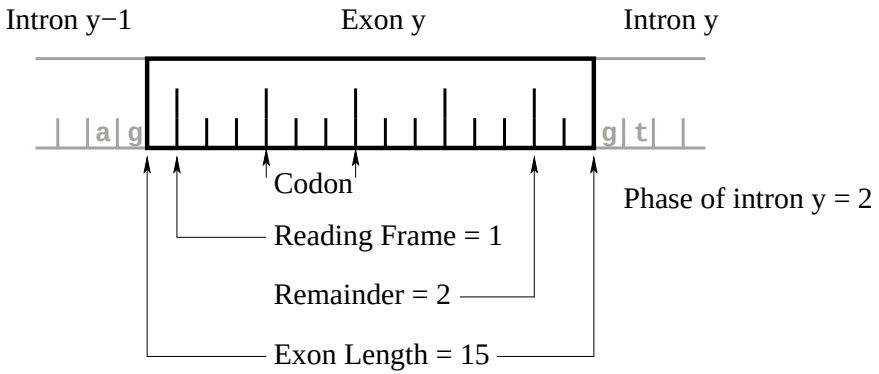


Possible intron positions

- Between codons: phase 0
- Between first & second nucleotide of codon: phase 1
- Between second & third nucleotide of codon: phase 2

Phase of intron = 3 - frame of 3' exon

Remainder of Exon



- Number of nucleotides at 3' end of exon that do not form complete codon. $0 \leq$ remainder ≤ 2.
- Remainder of exon = phase of 3' intron

Sequence Characteristics of Exons in CDS

Type	5' signal		3' signal		Frame	Remainder
First	initiation codon	ATG	donor site	gt	0	0,1,2
Internal	acceptor site	ag	donor site	gt	0,1,2	0,1,2
Terminal	acceptor site	ag	stop codon	TAR, TGA	0,1,2	0
Single	initiation codon	ATG	stop codon	TAR, TGA	0	0

Upper case: exonic; lower case: intronic

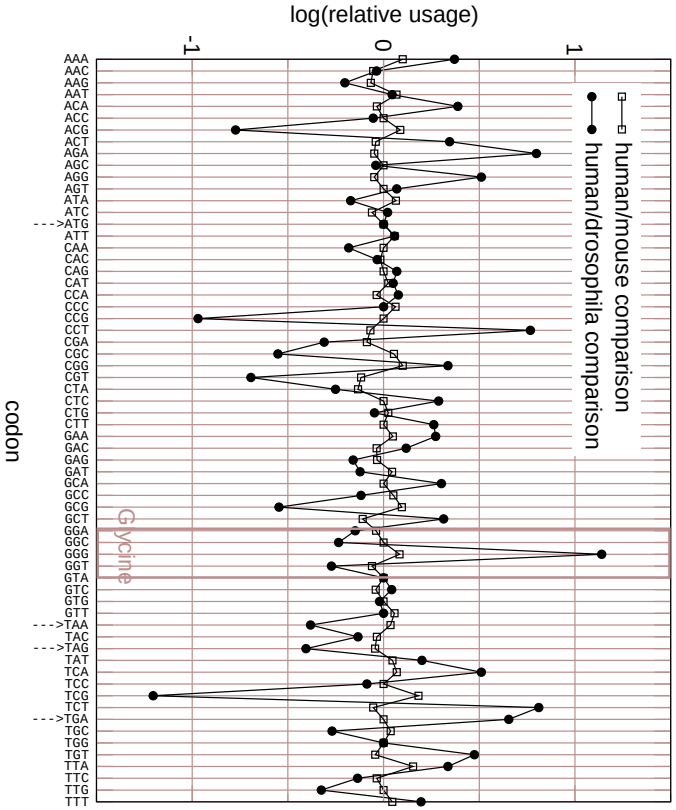
Genetic Code

5' end	second position				3' end
	T	**C**	**A**	**G**	
T	Phe/F	Ser/S	Tyr/Y	Cys/C	**T**
					C
	Leu/L		Ter/*	Ter/*	**A**
				Trp/W	**G**
C		Pro/P	His/H	Arg/R	**T**
					C
			Gln/Q		**A**
					G
A	Ile/I	Thr/T	Asn/N	Ser/S	**T**
					C
			Lys/K	Arg/R	**A**
	Met/M				**G**
G	Val/V	Ala/A	Asp/D	Gly/G	**T**
					C
			Glu/E		**A**
					G

Codon Usage

1. Preference within set of degenerate codons; codon bias
2. Variation across genes: high codon bias in highly expressed genes
3. Variation across species
4. Search by content: optimization for each species necessary

Codon Usage in Mouse, Human, & *Drosophila*



Construct Primitive Donor Finder

1. Align sequences containing donor site (GT)
2. Construct profile
3. Search sequence with profile

Multiple Sequence Alignment

fragment	accession number	5'-end of intron	position	sequence
f_1	U04239	donor 1	868-876	CTG**gt**gagt
f_2	U04239	donor 2	992-1000	GGG**gt**aagt
f_3	U04239	donor 3	1481-1489	CGG**gt**aagt
f_4	U04239	donor 4	2356-2364	AGG**gt**aagt
f_5	U04239	donor 5	2669-2677	AAA**gt**aagt
f_6	U04239	donor 6	3291-3299	TAG**gt**aact
f_7	M61127	donor 1	226-234	TGG**gt**ttgt
f_8	M61127	donor 2	539-547	TAG**gt**gagt

Exonic: upper case; intronic: lower case.

Frequency Matrix

	-3	-2	-1	0	1	2	3	4	5
A	2/8	3/8	1/8	0/8	0/8	5/8	7/8	0/8	0/8
C	2/8	0/8	0/8	0/8	0/8	0/8	0/8	1/8	0/8
G	1/8	4/8	7/8	8/8	0/8	2/8	0/8	7/8	0/8
T	3/8	1/8	0/8	0/8	8/8	1/8	1/8	0/8	8/8

Position Weight Matrix (PWM)

$\log_2(O/E)$, assume equal distribution of $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

	-3	-2	-1	0	1	2	3	4	5
A	0	0.58	-1	$-\infty$	$-\infty$	1.32	1.81	$-\infty$	$-\infty$
C	0	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	-1	$-\infty$
G	-1	1	1.81	2	$-\infty$	0	$-\infty$	1.81	$-\infty$
T	0.58	-1	$-\infty$	$-\infty$	2	-1	-1	$-\infty$	2

Example Computation

Entry 0.58 at position 2 for nucleotide A:

$$\log_2 \left(\frac{3/8}{1/4} \right) \approx 0.58 \,.$$

Information Content of MSA

- Uncertainty at site x:

$$H(x) = -\sum_{i=A,C,G,T} f(i,x)\log_2 f(i,x)\,,$$

  where $f(i,x)$ is the relative frequency of nucleotide i at position x.
- $0 \leq H \leq 2$.
- Information content $\equiv$ decrease in uncertainty:

$$R(x) = 2 - H(x)\,.$$

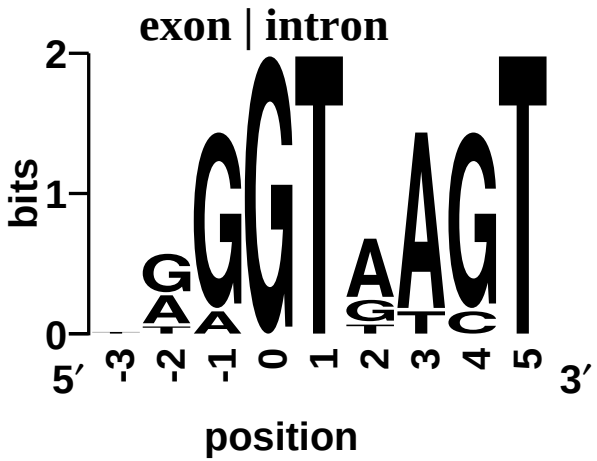
Information Content Around Donor Site

position x	information content $R(x)$
-3	0.09
-2	0.59
-1	1.46
0	2.00
1	2.00
2	0.70
3	1.46
4	1.46
5	2.00

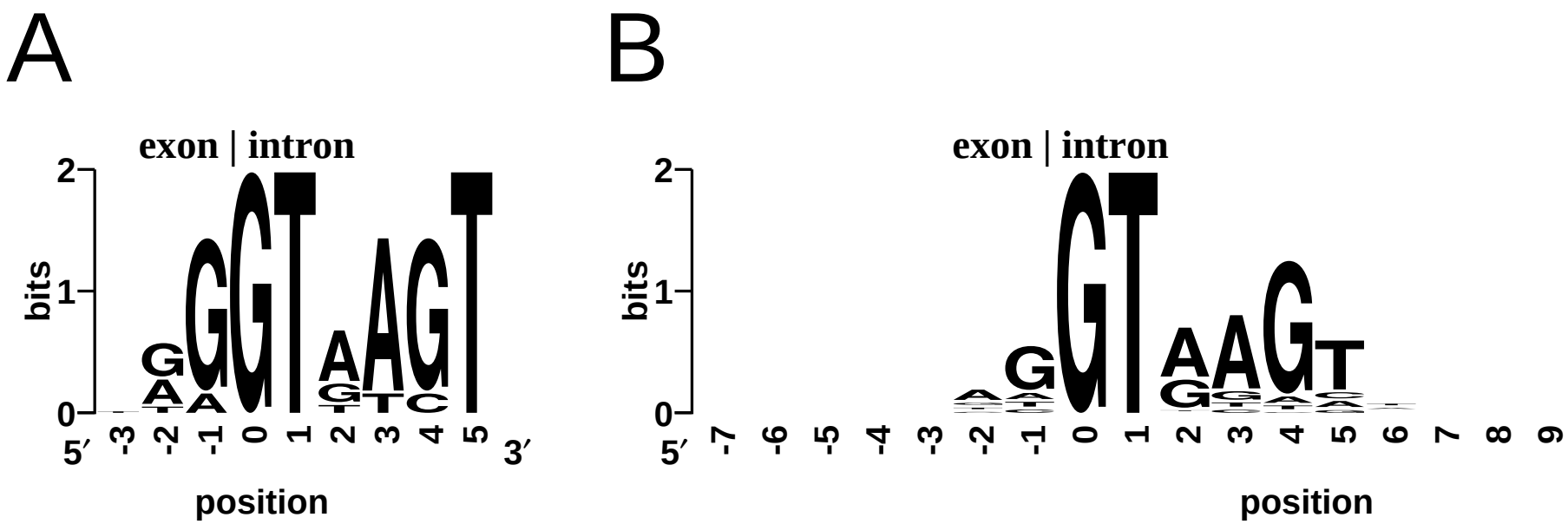
Example computation, position 3:

$$2+\frac{2\times 2/8\times\log(2/8)+1/8\times\log(1/8)+3/8\times\log(3/8)}{\log(2)}\approx 0.09$$

Sequence Logo, Example Data

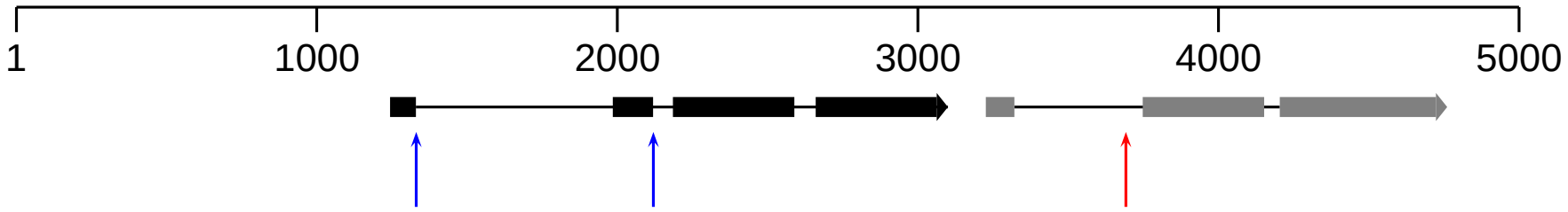


Sequence Logo, 757 Sequences



(A) 8 donor sequences of example data (B) 757 donor sequences from Drosophila

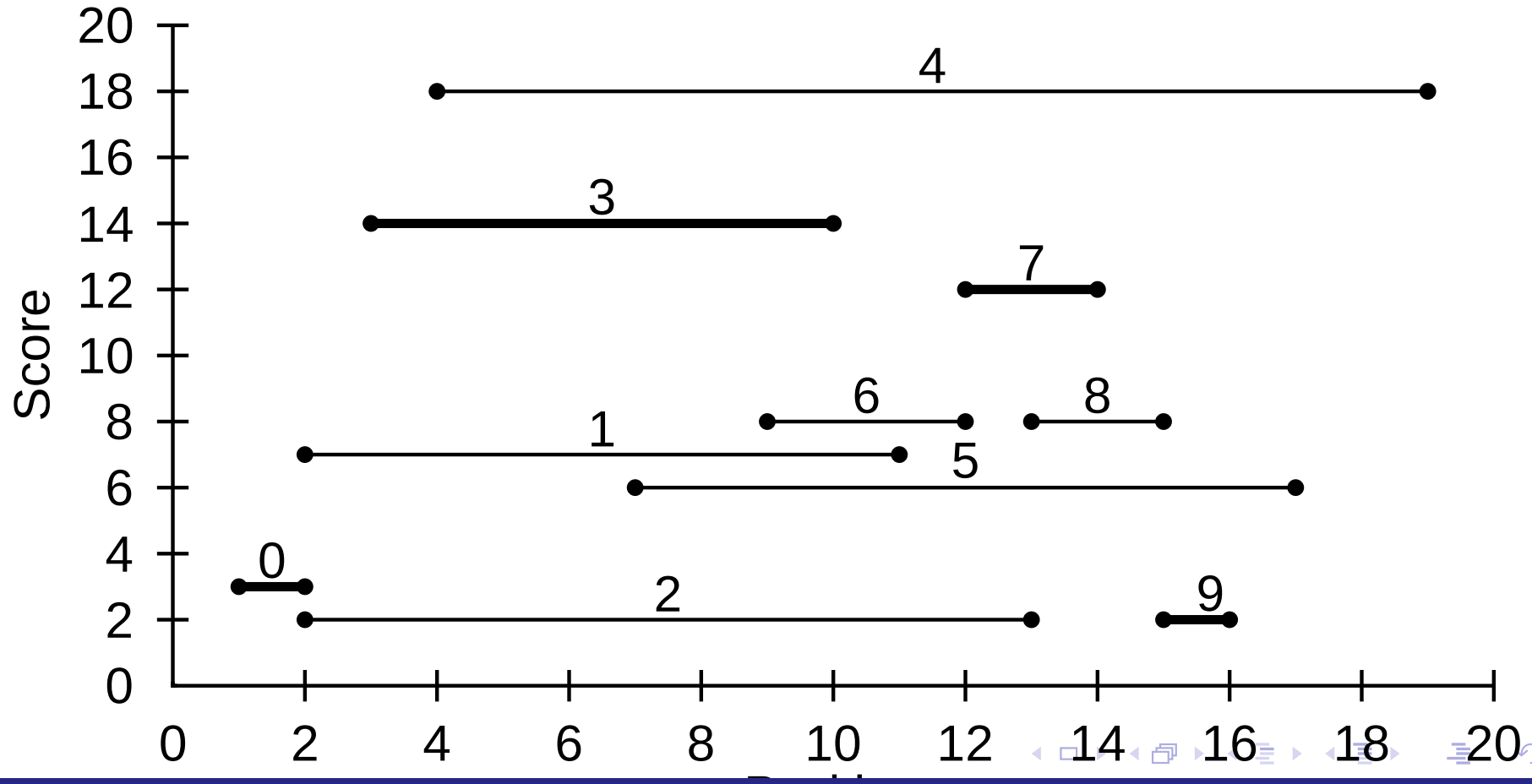
Application



position*	score	sequence
1331	+10.52	CAA**GT**AAGT
2120	+10.52	AAG**GT**AACT
3692	+4.39	GAA**GT**TACT

- sensitivity: $2/5$
- specificity: $2/3$

Problem



Algorithm

Require: e //array of n intervals
Require: b //sorted array of the $2n$ borders of the n intervals
  //Forward Algorithm
  $s \leftarrow 0$
  for all $0 \leq i < n$ **do**
    $p_i \leftarrow -1$ //initialize backpointers to "NULL"
    $v_i \leftarrow$ score of e_i
  $l \leftarrow -1$ //initialize index to last element added to optimal chain
  for $i \leftarrow 0$ to $2n - 1$ **do**
    if b_i is left border of interval e_j **then**
      $v_j \leftarrow v_j + s$
      $p_j \leftarrow l$
    else
      if $v_j > s$ **then**
        $s \leftarrow v_j$
        $l \leftarrow j$
  output s
  //Traceback
  while $l \neq -1$ **do**
    output e_i
    $l \leftarrow p_l$

Example—1

#	Position	Interval	Side
0	1	0	l
1	2	2	l
2	2	1	l
3	2	0	r
4	3	3	l
5	4	4	l
6	7	5	l
7	9	6	l
8	10	3	r
9	11	1	r

#	Position	Interval	Side
10	12	7	l
11	12	6	r
12	13	8	l
13	13	2	r
14	14	7	r
15	15	9	l
16	15	8	r
17	16	9	r
18	17	5	r
19	19	4	r

Example—2

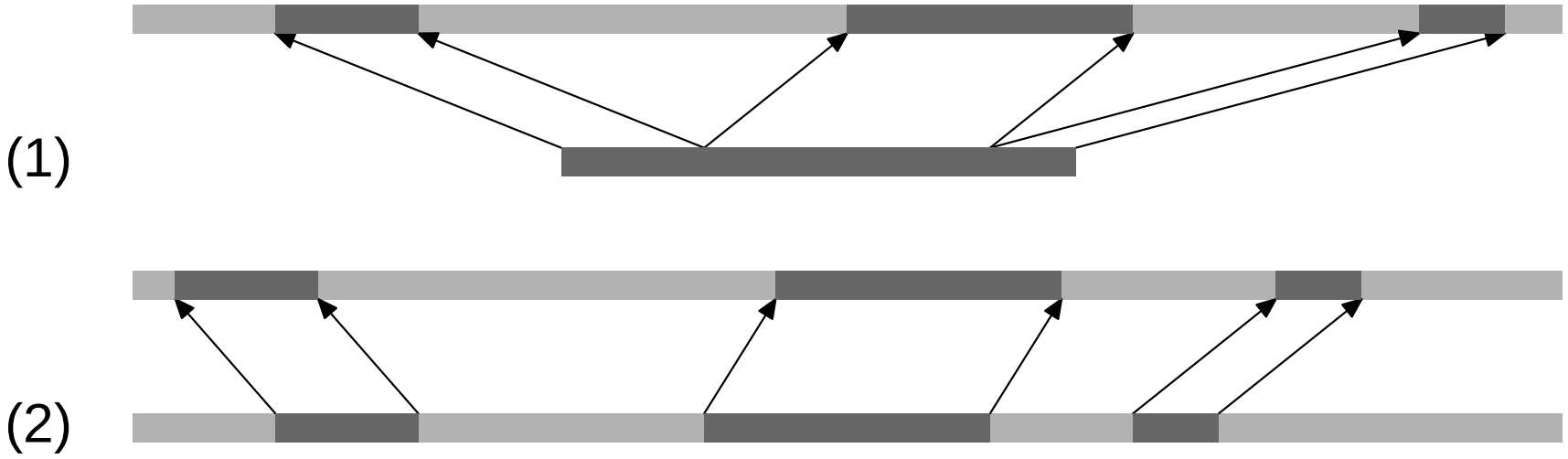
A

i	v_i
0	3
1	7
2	2
3	14
4	18
5	6
6	8
7	12
8	8
9	2

B

i	v_i
0	3
1	7
2	2
3	17
4	21
5	9
6	11
7	29
8	25
9	31

Spliced Alignment

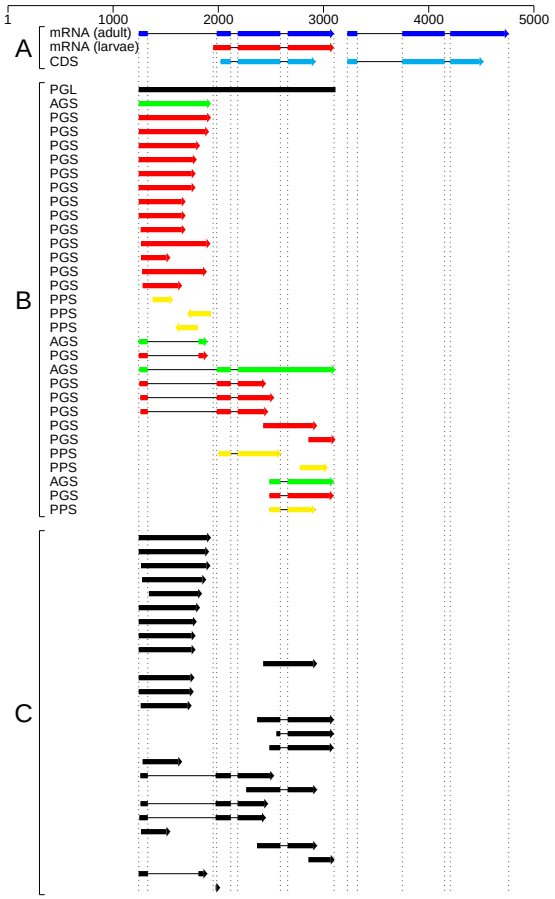


(1) genomic query and protein, cDNA or EST subject sequence,
(2) genomic query and homologous genomic subject sequence

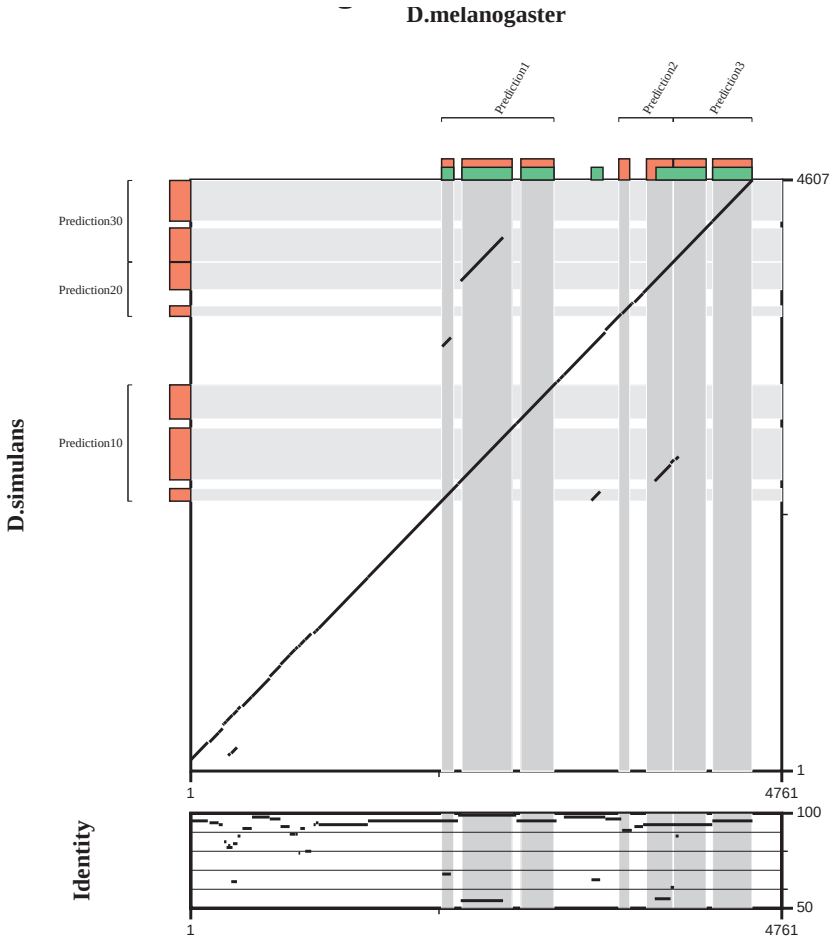
Implementations

1. `GeneSeqer`: computes spliced alignments of query & all matching ESTs
2. `SLAM`: pairwise alignment between two genomic sequences

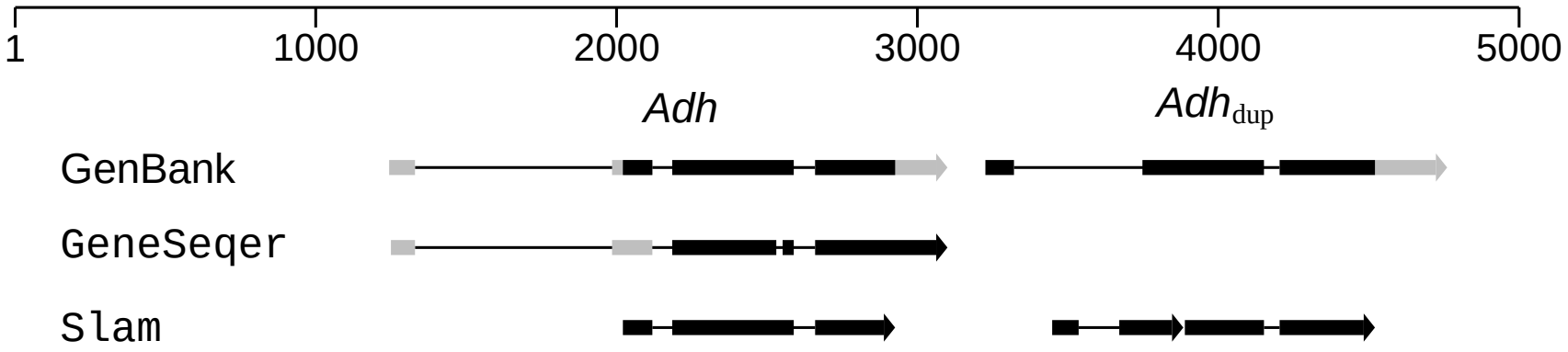
Spliced Alignments



Genomic Comparison



Results


1
1000
2000
3000
4000
5000
Adh
Adh dup
GenBank
GeneSeqer
Slam