How does eukaryotic gene prediction work?

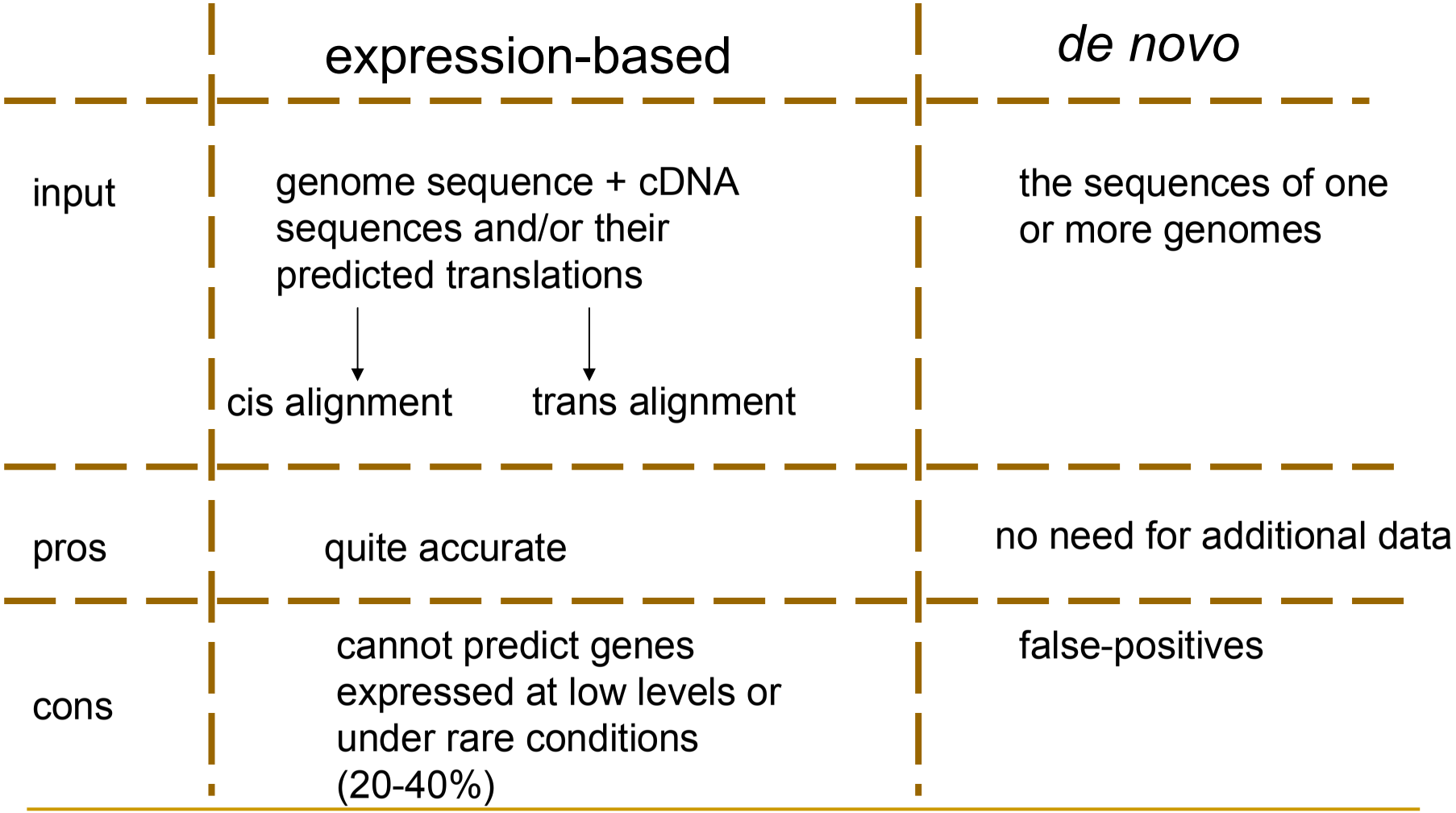
Age Tats

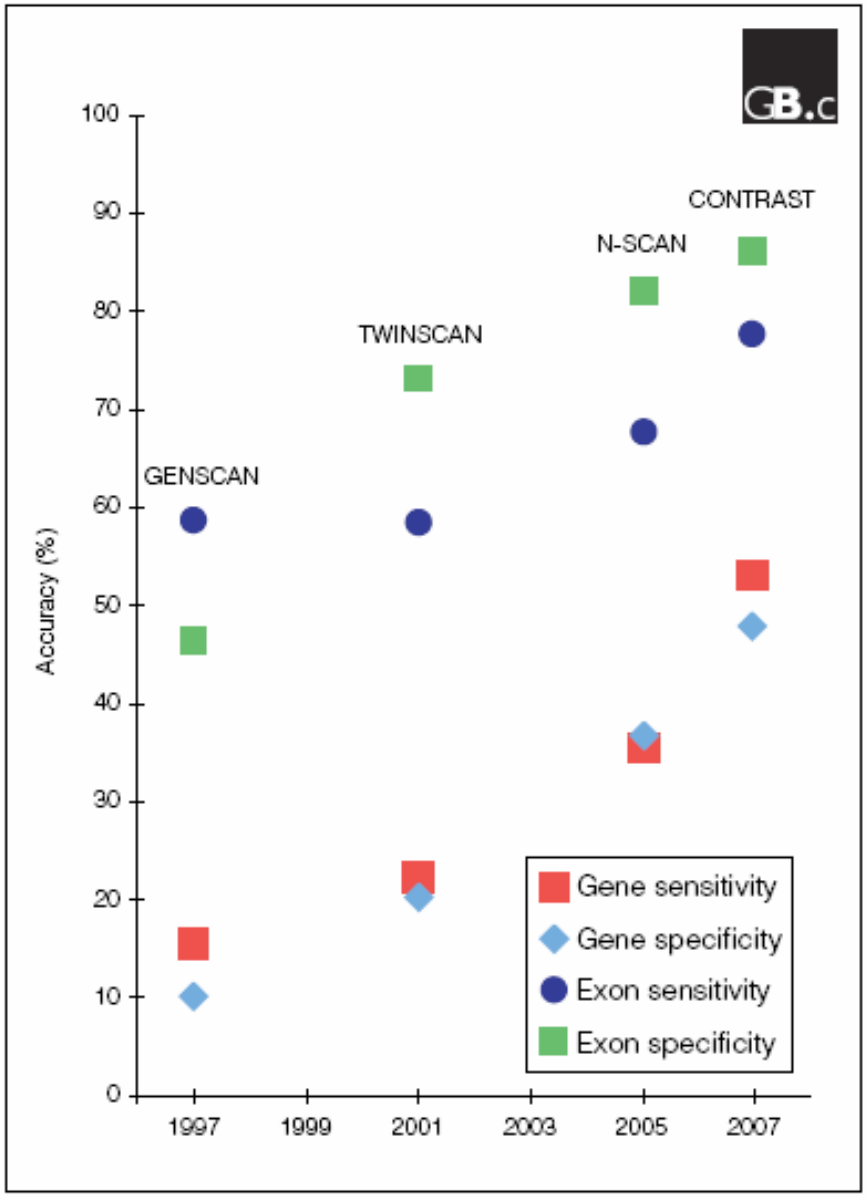
JClub Feb 26th, 2008

Gene prediction

- complete exon-intron structures of the protein-encoding portions of transcripts (open reading frames)
- 5’ untranslated regions
- only the boundaries of isolated exons

Major approaches





Flicek (2007) *Genome Biology* **8**:233

Generative (GHMM) versus discriminative models (CRF)

Generative models Pr(parse|seq) = C Pr(parse)Pr(seq|parse)

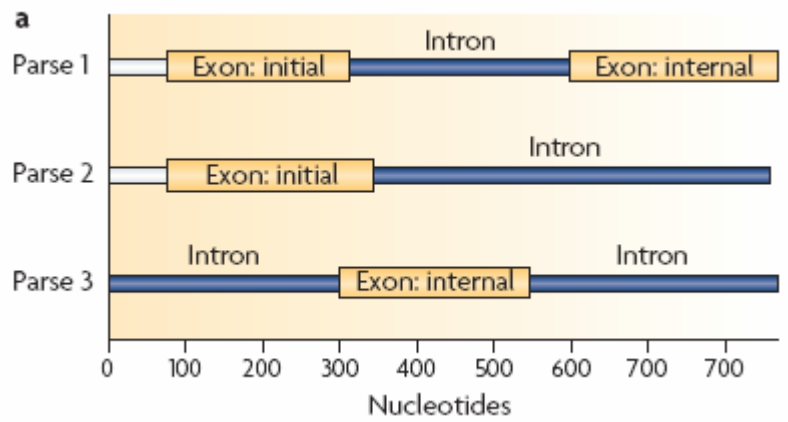
Pr(seq|parse 1) =

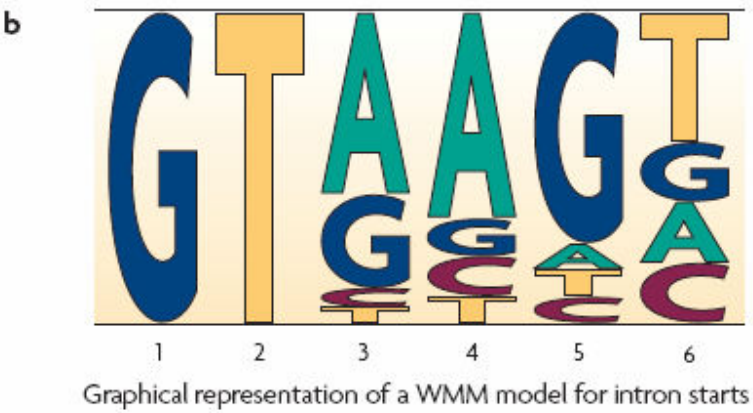
Pr(seq(1,80)|Intergenic(1,80))x

Pr(seq(80,320)|Exon:initial(80,320))x

Pr(seq(320,600)|Intron(320,600))x

Pr(seq(600,800)|Exon:internal(600,800))





Graphical representation of a WMM model for intron starts

Brent (2008) *Nature Reviews, Genetics*, **9**, 62-73

Generative (GHMM) versus discriminative models (CRF)

Generative models $\Pr(parse|seq) = C \Pr(parse)\Pr(seq|parse)$

Discriminative models $\Pr(parse|seq) = Ce^{WF(seq,parse)}$

WF(seq,parse) is a weighted sum of feature functions:

$$\sum_j W_j F_j(seq, parse)$$

F_j – predefined function, can depend on any part of the input sequence

W_j – the weight of each function, learned from examples of correct parses during training

CONTRAST:
CONditionally **TRA**ined **S**earch for **T**ranscripts

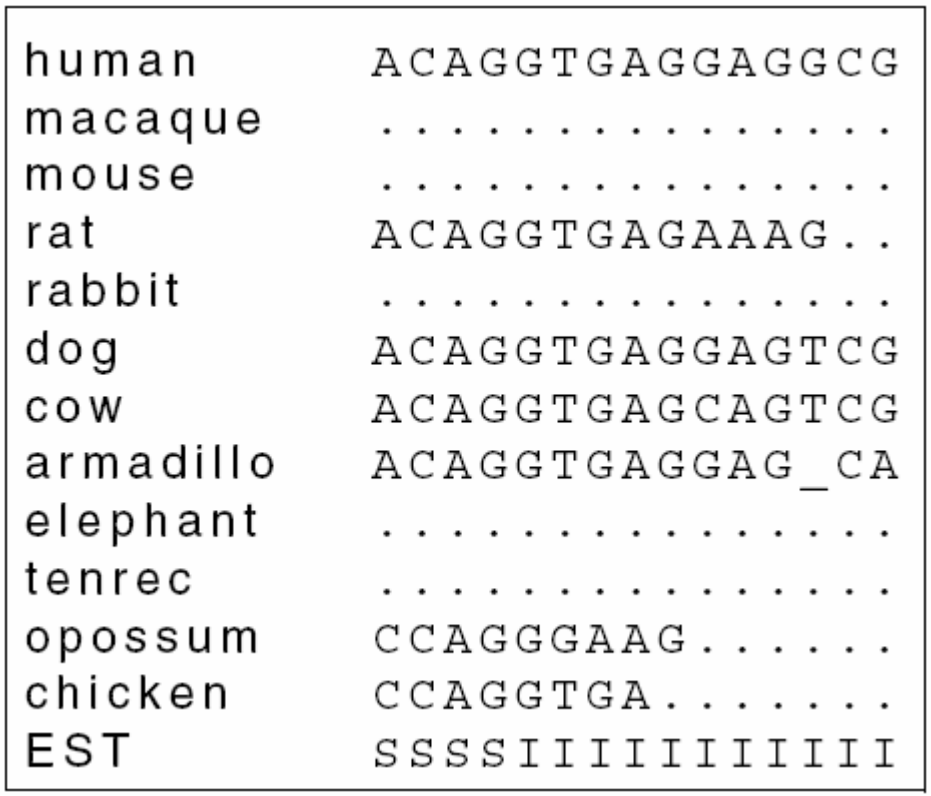
Gross et al (2007) Genome Biology 8:R269

Directly incorporates information from multiple alignments rather than employing phylogenetic models.

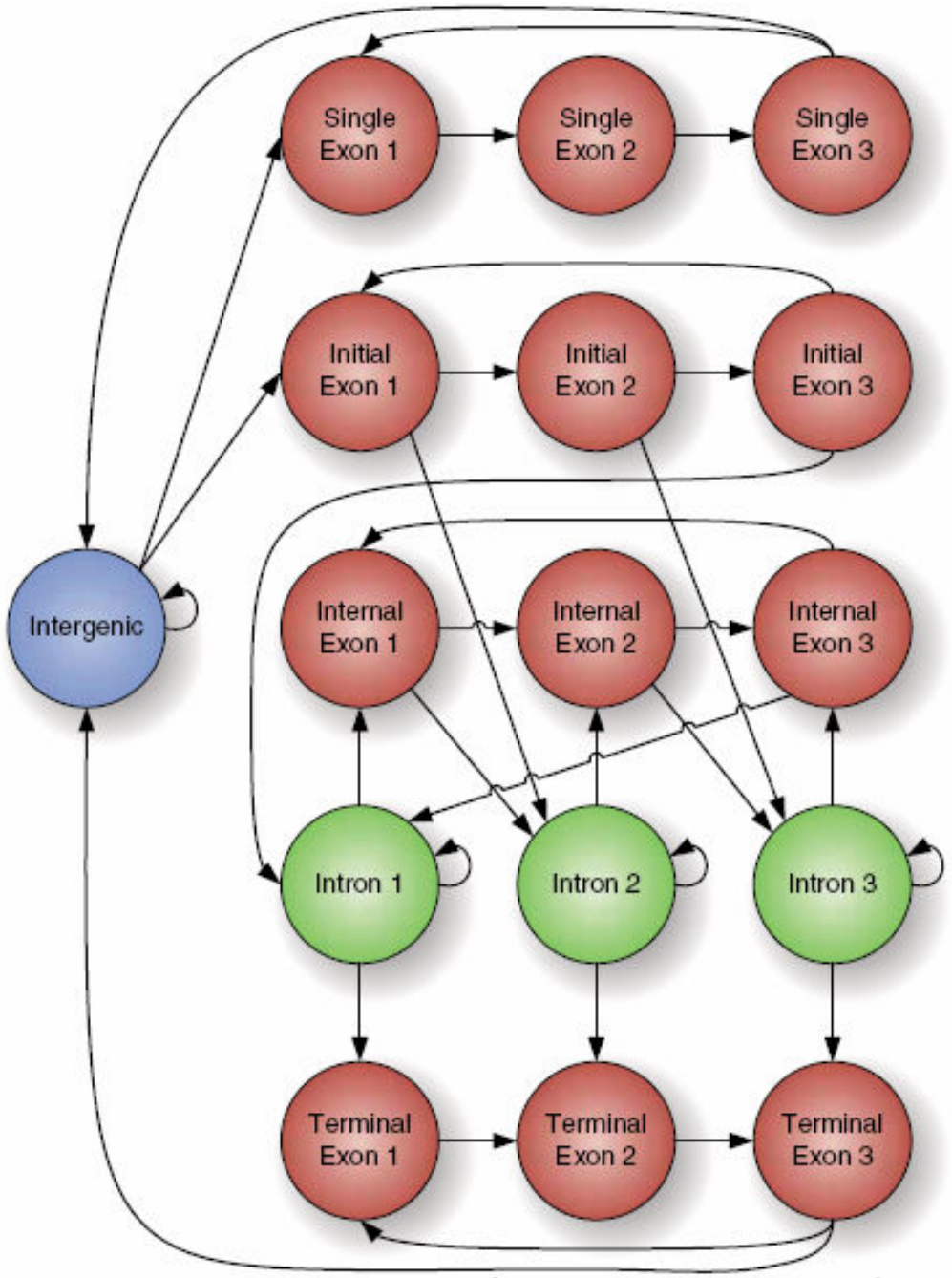
Uses

-SVMs for its coding boundary classifiers

-CRF (conditional random field) for its global model of gene structure



Gross et al (2007) *Genome Biology* 8:R269



Gross et al (2007) *Genome Biology* 8:R269

Boundaries of coding regions

Table 6

Coding region boundary classifiers. Coding region boundary classifiers. Window sizes and positions are shown for the five coding region boundary classifiers used by CONTRAST. Coordinates are defined such that the boundary occurs between the adjacent positions -1 and 1 (that is, either position -1 is coding and position 1 is coding or the reverse is true), with coordinates increasing in the 5' to 3' direction. Each classifier's require consensus sequence is shown in the second column.

	Consensus	5' end	3' end	Length
Start codon	$A_1T_2G_3$	-8	6	14
Stop codon	$T_1A_2A_3$, $T_1A_2G_3$, $T_1G_2A_3$	1	6	6
Donor splice GT	G_1T_2	-3	8	11
Donor splice GC	G_1C_2	-3	8	11
Acceptor splice	$A_{-2}G_{-1}$	-27	3	30

Gross et al (2007) Genome Biology 8:R269

Global model of gene structure

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{e^{\mathbf{w}^{\mathrm{T}}\mathbf{F}(\mathbf{x},\mathbf{y})}}{\sum_{\mathbf{y}'} e^{\mathbf{w}^{\mathrm{T}}\mathbf{F}(\mathbf{x},\mathbf{y}')}}$$

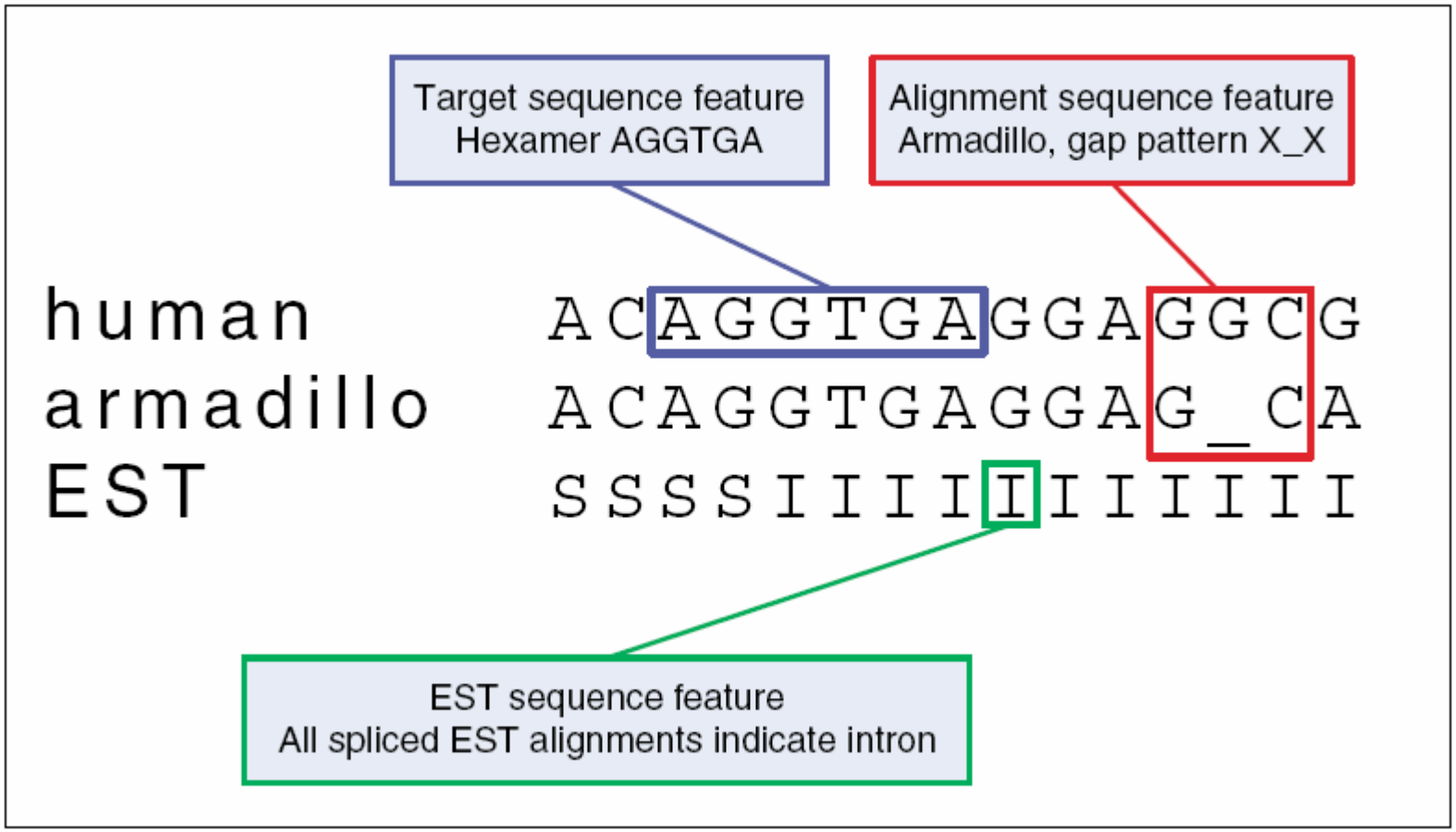
$$\mathbf{F}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{L} \mathbf{f}(y_{i-1}, y_i, i, \mathbf{x}).$$

f – feature mapping, a vector-valued function which determines the information used to calculate the score of a position.

$$\mathbf{f}(y_{i-1}, y_i, i, \mathbf{x}) = \begin{bmatrix} 1\{y_{i-1} = \text{Intron and } y_i = \text{Exon}\} \\ 1\{y_i = \text{Exon and } x_i = \text{'G'}\} \\ \cdots \end{bmatrix}$$

Gross et al (2007) *Genome Biology* 8:R269

Feature mapping in CONTRAST



Gross et al (2007) *Genome Biology* 8:R269

Human gene prediction

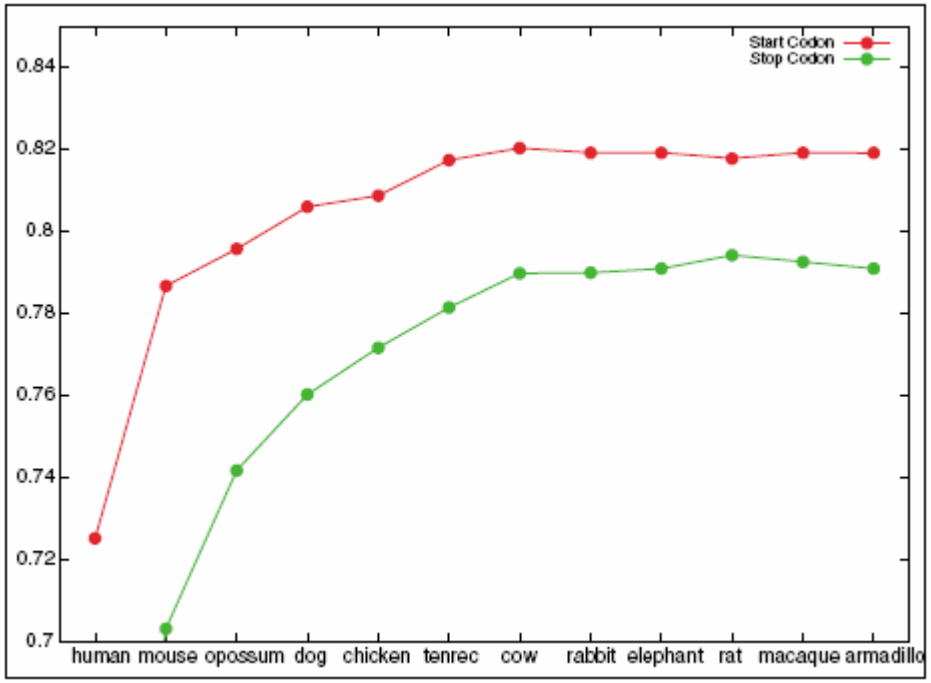
Table 1

De novo gene prediction performance for human. Sensitivity (Sn) and specificity (Sp) were evaluated at the gene, exon and nucleotide levels and reported as percentages. Also shown are the average number of genes and exons predicted for each cross-validation fold. The column headings indicate the predictor and informants used.

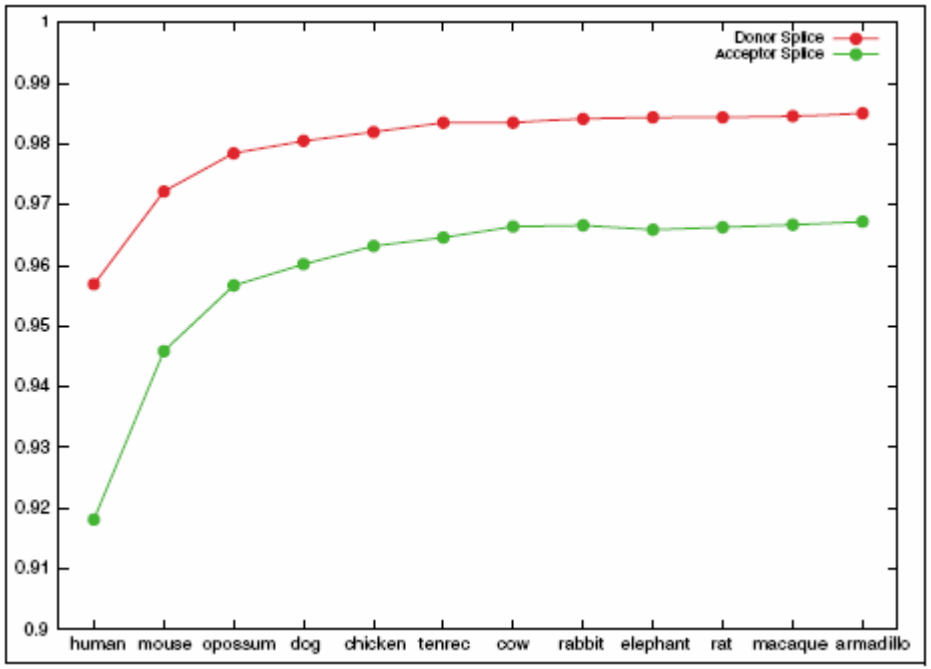
	N-SCAN (mouse)	CONTRAST (mouse)	CONTRAST (11 informants)
Gene Sn	35.6	50.8	58.6
Gene Sp	25.1	29.3	35.5
Exon Sn	84.2	90.8	92.8
Exon Sp	64.6	70.5	72.5
Nucleotide Sn	90.8	96.0	96.9
Nucleotide Sp	67.9	70.0	72.0
Genes predicted	22,596	27,614	26,260
Exons predicted	196,643	211,431	210,180

Gross et al (2007) *Genome Biology* 8:R269

Start and stop codon classifier accuracy



Splice site classifier accuracy



Gross et al (2007) *Genome Biology* 8:R269