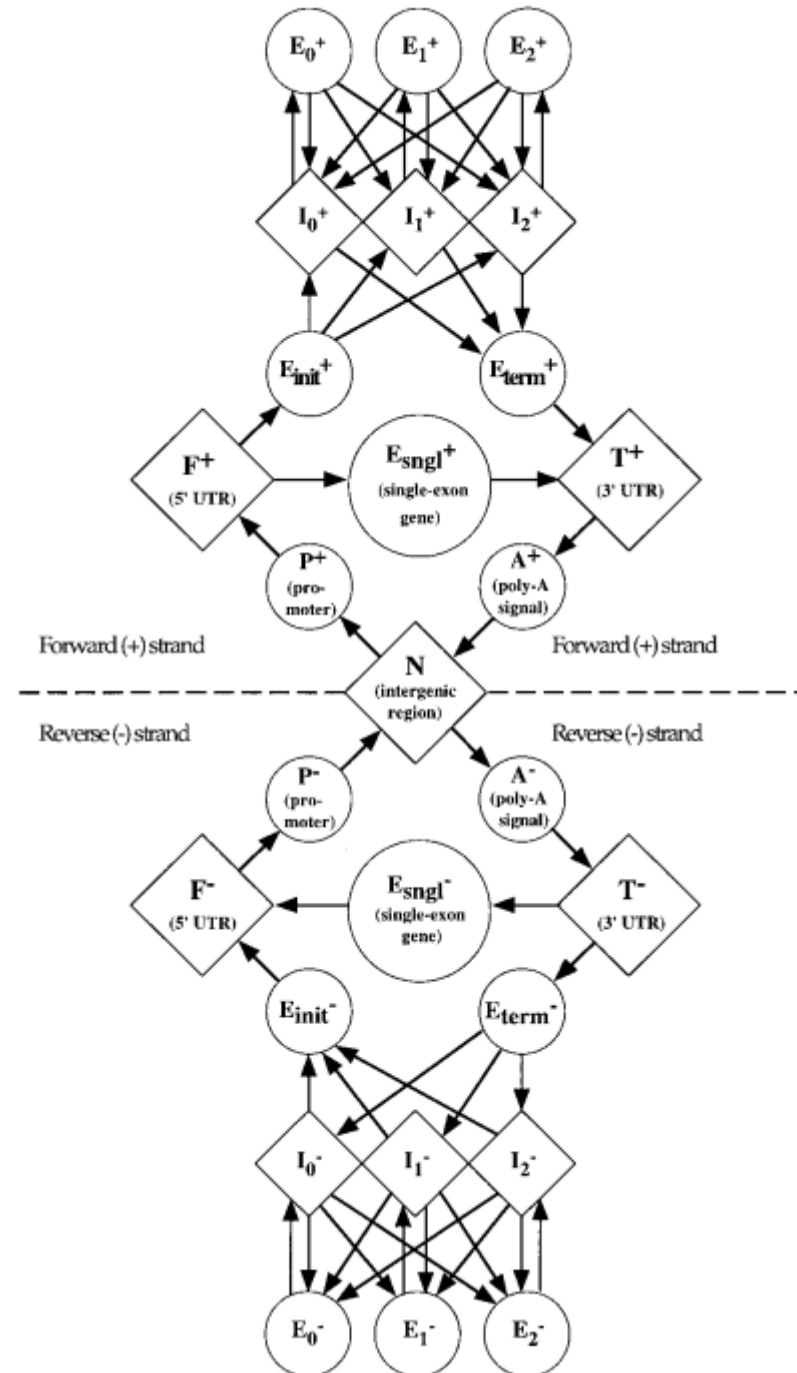


Finding Eukaryote Genes

Based on Burge & Karlin 1997

CCAATTAGTCCCATTGAAACTGTACCAGTAAAATTAAGCCAGGAATGGATGGCCC
AAAAGTTAAACAATGGCCATTGACAGAAGAGAAAATAAAAGCATTAAACAGCAATTTG
TAAAGAAATGGAAGAGGAAGGAAAAATTACAAGAATTGGGCCTGAAAATCCATATAA
CACTCCAGTATTTGCTATAAAAAAGAAGGACAGTACTAAGTGGAGGAAATTAGTAGA
TTTCAGGGAACTCAATAAAAGAACTCAAGACTTTTGGGAAGTTCAATTAGGAATACC
ACACCCAGCAGGGCTAAAAAAGAAAAAATCAGTGACAGTGCTTGATGTGGGGGAT
GCATATTTTTCAATCCCTTTAGATGAAAGCTTCAGGAAATATACGGCATTACCCATA
CCCAGCATAAACAATGAAGCACCAGGGATTAGATATCAATACAATGTGCTTCCACA
GGGATGGAAAATCTCACCAGCAATATTCCAGAGTAGCATGACAAGAATCTTAGAGC
CCTTTAGGAAACAAAACCCAGACATAGTTATCTATCAATATATGGATGACTTGTATG
TAGGATCTGACTTAGAAATAGGGCGACATAGAGCAAAAATAGAGGAGCTAAGAGAA
CACCTATTGAGATGGGGGCTTACCACACCAGACAAGAAACATCAGAAAAAACCCCC
ATTTCTTTGGATGGGGTATGAACTCCATCCTGACAAATGGACAGTACAGCCTATACT
GCTGCCAGAAAAGGATAGCTGGACTGTCAATGACATACAGAAGTTAGTGGGAAAAT
TAACTGGGCCAGTCAGATTTACCCAGGGATTAAAGTAAGGCAGCTGTGCAAACCTC
CTTAGGGGAGCCAAAGCATTAAACAGACATAGTACCACTGACTGAAGAAGCGCCAAT
TAGTCCCATTGAAACTGTACCAGTAAAATTAAGCCAGGGATGGATGGCCCAAAG
TTAAACAATGGCCATTGACAGAAGAGAAAATAAAAGCATTAAACAGCAATTTGTAAAG
AAATGGAAAAGGAAGGAAAAATTACAAAAATTGGGCCTGAAAATCCATATAACACTC
CAGTATTTGCTATAAAAAAGAAGGACAGTACTAAGTGGAGAAAATTAGTAGATTTCA
GGGAACCTCAATAAGAGAACTCAAGACTTTTGGGAAGTTCAATTAGGAATACCACAC
CCAGCAGGGTTAAAAAAGAAAAGATCAGTGACAGTGCTTGATGTGGGGGATGCATA
TTTTTCAATCCCTTTAGATGAAAACCTTCAGGAAATATACTGCATTACCCATACCTAGC
ATAAACAATGAAGCACTAGGGATTAGATATCAATATAATGTGCTTCCACAGGGATGG
CCAAAGCACTAACAGACGTAGTGCCACTAACTGAAGAACCAATTAGTCCCAAAGCA

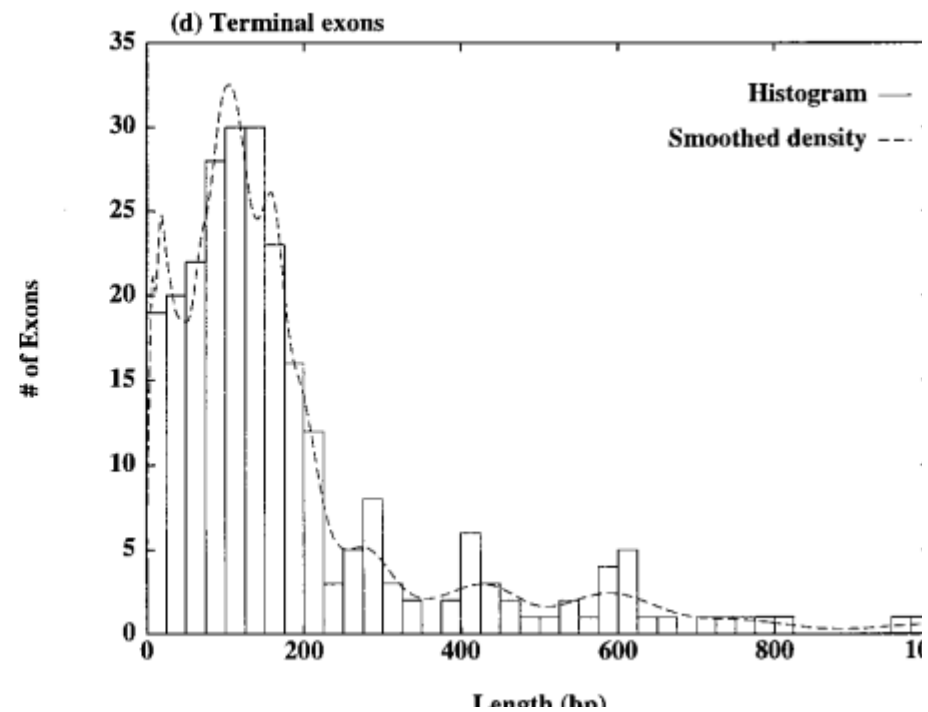
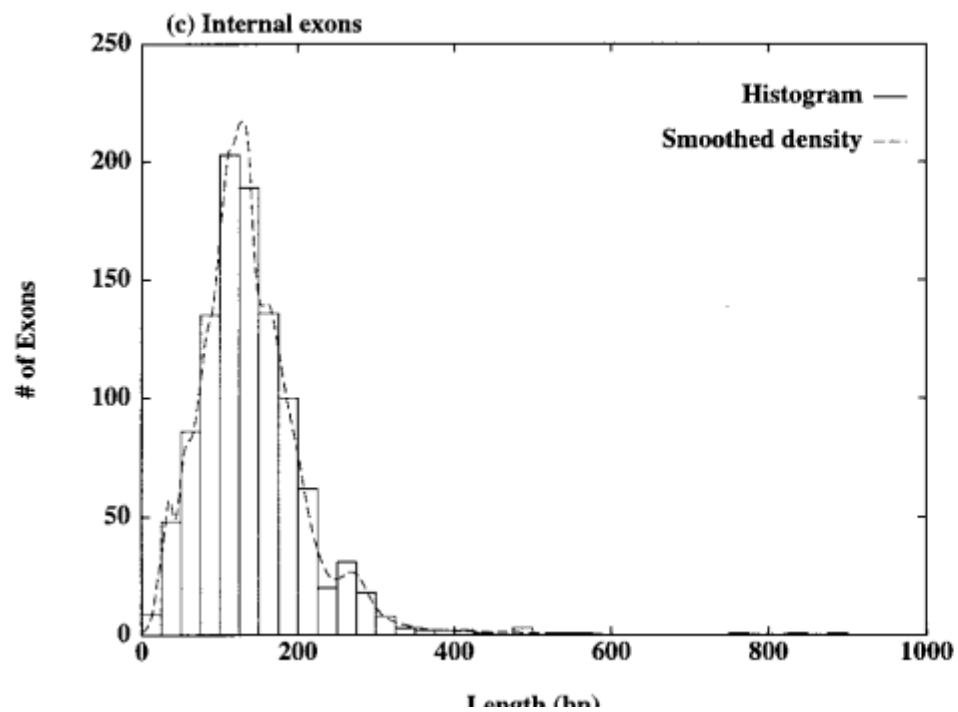
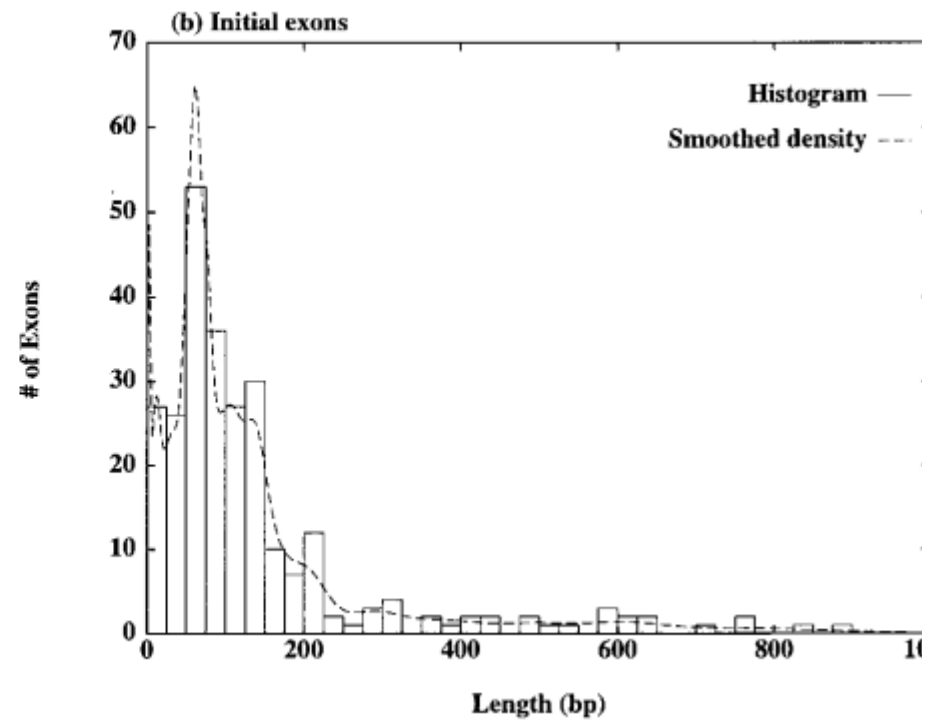
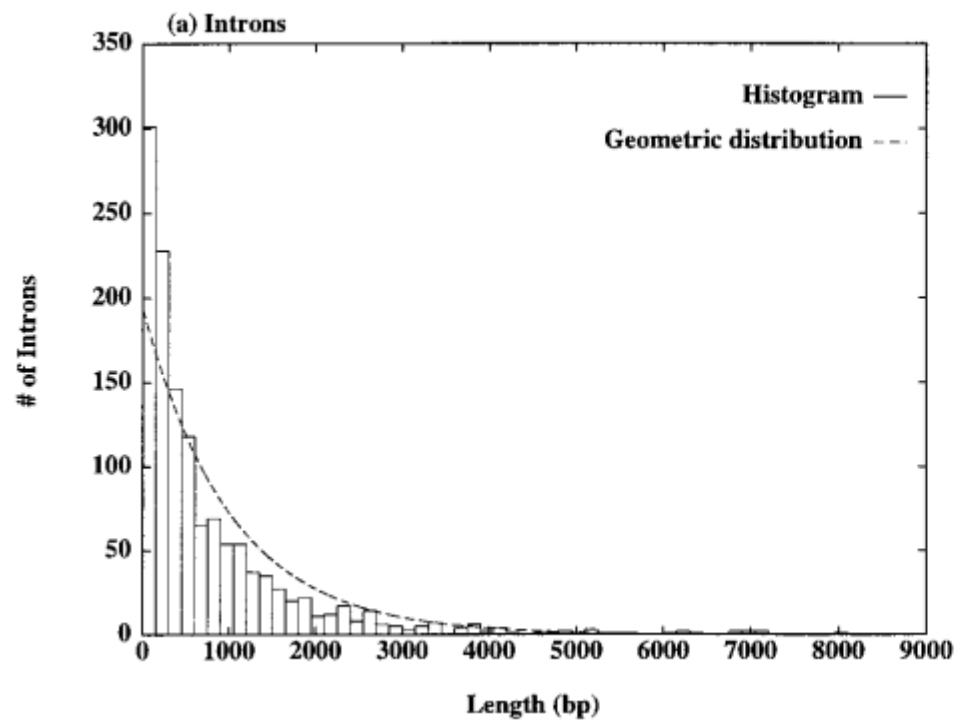
Model



Two types of hidden states

States with geometric duration (diamonds on previous figure)

States that can generate arbitrary sequences of observed states (circles)



A version of the forward algorithm for this model

Calculate $f_i(j)$ for any sequence position j and all D states i

Initialization: $f_i(1) = \pi_i p_i P_i(S_1)$

Induction:

$$\begin{aligned} f_i(j+1) &= f_i(j) p_i P_i(S_{j+1}) \\ &+ \sum_{\epsilon_k \in L_j} f_{x_k}(a_k - 1) (1 - p_{x_k}) T_{x_k, y_k} \alpha_{y_k}(\lambda_k) P_{y_k}(S_{a_k, j}) T_{y_k, i} p_i P_i(S_{j+1}) \end{aligned}$$

A Comparison of GENSCAN with other gene prediction programs

Program	Sequences	Sn	Accuracy per nucleotide			Sn	Accuracy per exon			
			Sp	AC	CC		Sp	Avg.	ME	WE
GENSCAN	570 (8)	0.93	0.93	0.91	0.92	0.78	0.81	0.80	0.09	0.05
FGENEH	569 (22)	0.77	0.88	0.78	0.80	0.61	0.64	0.64	0.15	0.12
GeneID	570 (2)	0.63	0.81	0.67	0.65	0.44	0.46	0.45	0.28	0.24
Genie	570 (0)	0.76	0.77	0.72	n/a	0.55	0.48	0.51	0.17	0.33
GenLang	570 (30)	0.72	0.79	0.69	0.71	0.51	0.52	0.52	0.21	0.22
GeneParser2	562 (0)	0.66	0.79	0.67	0.65	0.35	0.40	0.37	0.34	0.17
GRAIL2	570 (23)	0.72	0.87	0.75	0.76	0.36	0.43	0.40	0.25	0.11
SORFIND	561 (0)	0.71	0.85	0.73	0.72	0.42	0.47	0.45	0.24	0.14
Xpound	570 (28)	0.61	0.87	0.68	0.69	0.15	0.18	0.17	0.33	0.13
GeneID+	478 (1)	0.91	0.91	0.88	0.88	0.73	0.70	0.71	0.07	0.13
GeneParser3	478 (1)	0.86	0.91	0.86	0.85	0.56	0.58	0.57	0.14	0.09

Evaluating performance

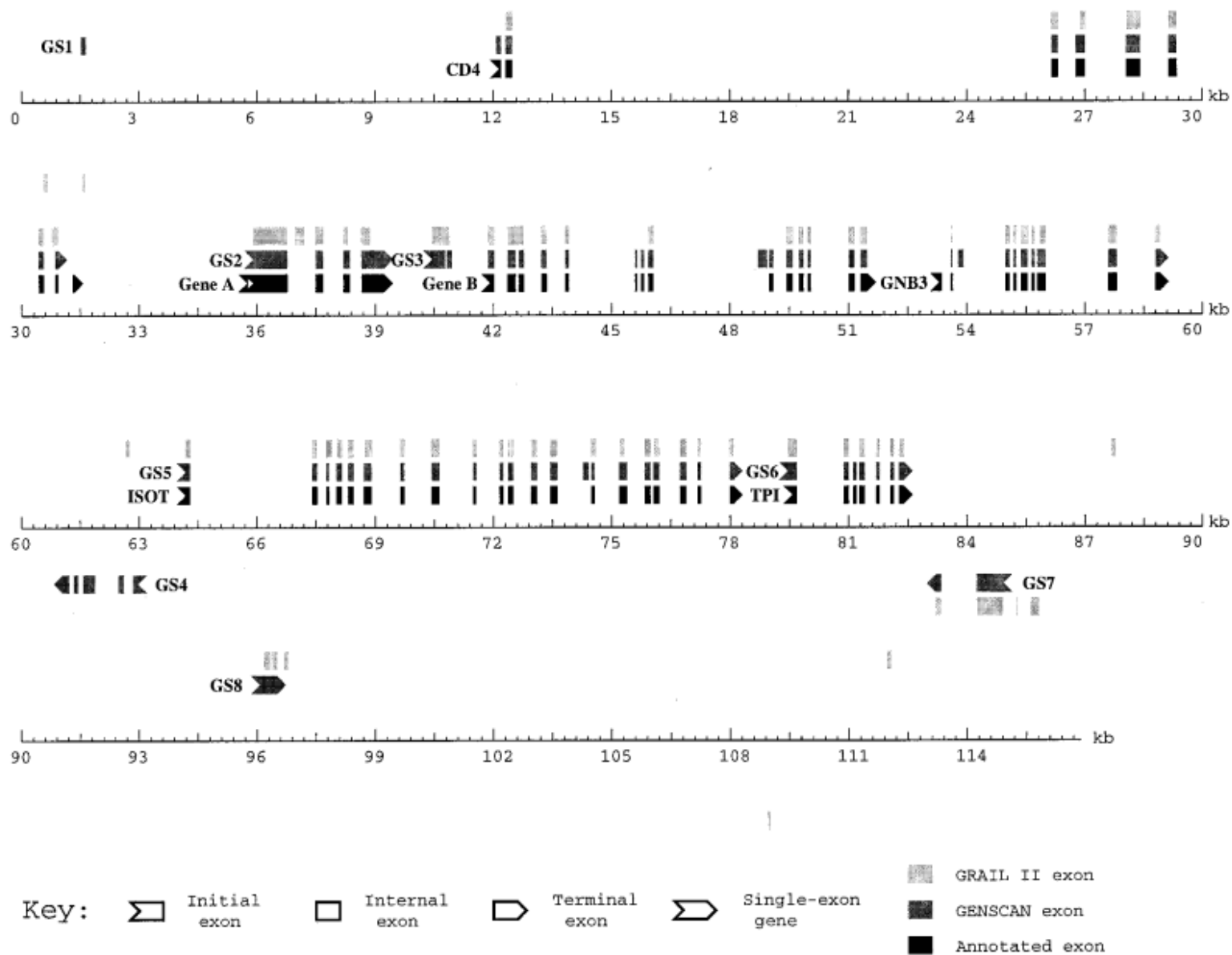
Sensitivity (Sn): TP/AP

Specificity (Sp): TP/PP

Correlation coefficient (CC): $CC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}};$

Approximate correlation (AC): $AC = \frac{1}{2} \left[\frac{TP}{AP} + \frac{TP}{PP} + \frac{TN}{AN} + \frac{TN}{PN} \right] - 1.$

Performance on arbitrary genomic DNA sequence



Maximum Dependence Decomposition (heuristic)

For each position i calculate

$$S_i = \sum_{j \neq i} \chi^2(C_i, X_j)$$

where C_i is the consensus nucleotide at position i

Choose $i_1 = \operatorname{argmax}_i(S_i)$

Divide sequences into those that have the consensus nucleotide at position i and the remaining sequences. Repeat procedure on subdivisions until no significant dependencies found

