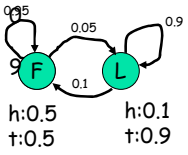
How to design an HMM for a new problem

- Architecture/topology design:
  - What are the states, observation symbols, and the topology of the state transition graph?
- Learning/Training:
  - Fully annotated or partially annotated training datasets
  - Parameter estimation by maximum likelihood or by EM
- Validation/Testing:
  - Fully annotated testing datasets
  - Performance evaluation (accuracy, specificity and sensitivity)

HMM model structure

- Duration modeling



What is the probability of staying with the fair coin for T time steps?

Inherent limitation of HMMs

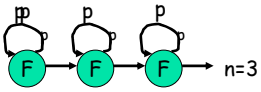
- The duration in state F follows an exponentially decaying distribution called a geometric distribution.

$$P(X = F^T) = (0.95)^{T-1}(0.05)$$

- The geometric distribution gives too much probability to short sequences of Fs and Ls and too little to medium and long sequences of Fs and Ls.

Duration modeling

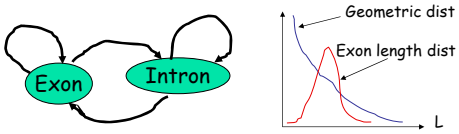
- To obtain non-geometric length distributions, we use an array of n F states, as follows:



$$P(|X| = L) = \binom{L-1}{n-1} p^{L-n}(1-p)^n$$

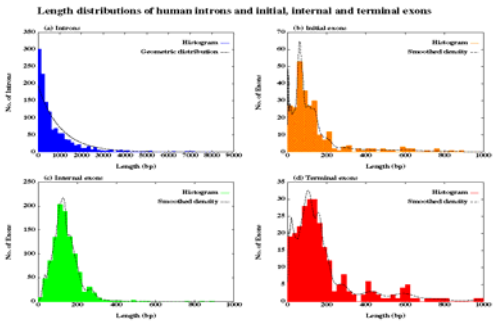
- Generated length distribution is a negative binomial.

Why does this matter?



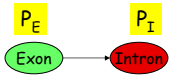
- Length of stay in "Exon" state determines length of predicted exons. Very short exons are rare.
- Similarly for introns. Introns shorter than 30 bp do not exist.

Length distributions of exons and introns



Generalized HMMs (semi-Markov HMMs)

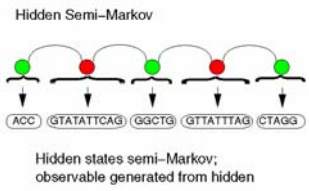
- Each state has a specified length distribution.



No self-transitions to generate extra symbols

- Pick a state to start at t=1.
- Repeat
  - Pick the length of stay (d) in current state from distribution P.
  - Emit d symbols in current state.
  - Pick a new state (according to a matrix) and transition to it at time t+d

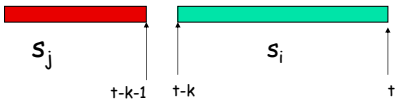
Example



Multiple symbols emitted in each state. One to one mapping between symbols and hidden states is lost in the generalized HMM.

Viterbi algorithm for gHMMs

- Just like Viterbi for HMMs, but we use the entire stay in state instead of a state at a given time.



$$\delta_t(i) = \max_{k=0..t-1} \max_{j \neq i} f_{t,i}(k, j)$$

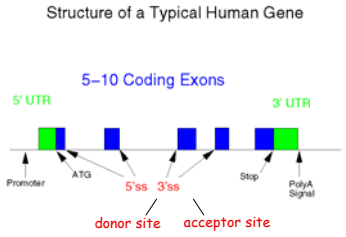
Probability of most likely path ending at t with stay of k+1 in state i following a stay in state j

$$f_{t,i}(k, j) = \left[\prod_{r=0}^{k} b_i(o_{t-r})\right] l_i(k) a_{ji} \delta_{t-k-1}(j)$$

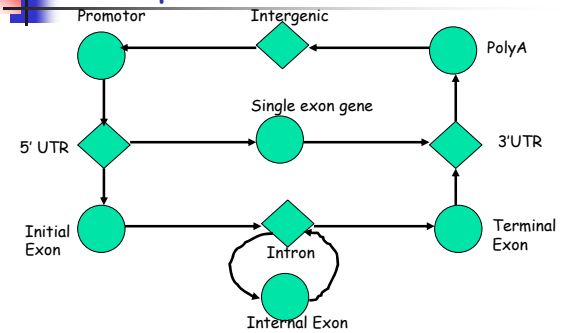
Genscan

- The Genscan HMM model
- Training Genscan
- Validating Genscan

Gene structure assumed by Genscan

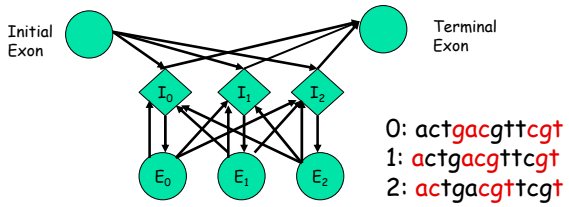


A simple model



Exon phases

Need to keep track of codon position in exon.

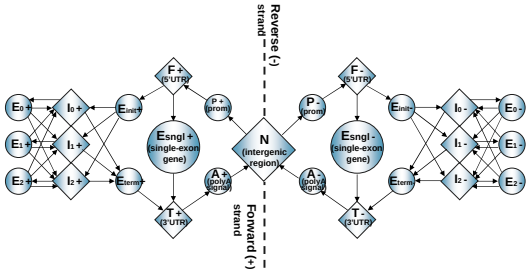


Genscan's architecture (1)

- HMM states for exons and introns in three different phases, single exon, 5' and 3' UTRs, promoter region, polyA site and intergenic region.
- Explicit length modeling of introns and exons.

("*Prediction of complete gene structures in human genomic DNA*"(1997) Burge and Karlin, *JMB* **268**, p. 86)

Genscan HMM



Genscan model components

- Vector of initial probabilities: π
- State Transition probability Matrix: a
- Set of length distributions: f_q conditional on state q.
- Emission probabilities: P(s|q,d) conditional on state and length.

Isochore groups

Group	I	II	III	IV
C + G% range	<43	43-51	51-57	>57
Number of genes	65	115	99	101
Est. proportion single-exon genes	0.16	0.19	0.23	0.16
Codelen: single-exon genes (bp)	1130	1251	1304	1137
Codelen: multi-exon genes (bp)	902	908	1118	1165
Introns per multi-exon gene	5.1	4.9	5.5	5.6
Mean intron length (bp)	2069	1086	801	518
Est. mean integenic length (bp)	83000	36000	5400	2600

Initial probabilities

	I	II	III	IV
Intergenic (N)	0.892	0.867	0.54	0.418
Intron (I0+,I1+,I2+,I0-,I1-,I2-)	0.095	0.103	0.338	0.388
5' Untranslated region (F+, F-)	0.008	0.018	0.077	0.122
3' Untranslated region (T+, T-)	0.005	0.011	0.045	0.072

All other probabilities set to zero.

Transition probabilities

- Probabilities of state transitions not present in model are zero.
- Deterministic transitions are assigned probability 1.
- The others transition probabilities are set according to maximum likelihood values in training data.

Length distribution for introns

- No introns < 65bp. After that geometric (exponential) distribution.
- Substantial difference between different C+G groups.
- So, intron length is modeled as geometric distribution with different parameters of different C+G groups.

Exon length distribution model

- Exons are very important to model.
- Substantial differences in length distribution between initial, internal and terminal exons.
- No substantial difference between different C+G compositional groups.
- Exon length means considered between 50 and 300 bps.
- Account for phase (3*codons + phase)

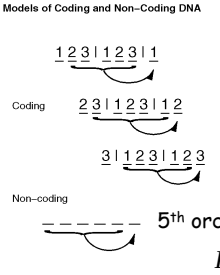
Other length distributions

- 5' UTR -> Geometric with mean 769bp
- 3' UTR -> Geometric with mean 457bp

Emission models

- **Exons** -- inhomogeneous 3-periodic 5th order Markov model.
- **Introns** and **intergenic regions** - homogeneous 5th order Markov model
- **5'** and **3' UTRs** - homogeneous 5th order Markov model

Emission models for exons and introns



5th order inhomogeneous Markov model

In an *inhomogeneous* Markov model, we have different distributions at different positions in the sequence.

5th order homogeneous Markov model :

$$P(o_t \mid o_{t-1}o_{t-2}o_{t-3}o_{t-4}o_{t-5})$$

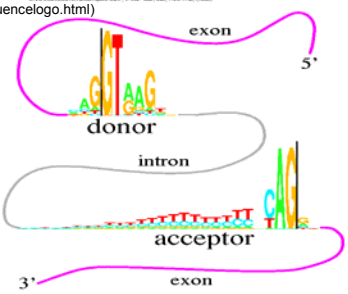
Genscan architecture (2)

- Weighted matrix (WMM) and weighted arrays (WAM) for acceptor splice site, polyA site and promoter region.
  - WMM: $p_j(i)$ is probability of nucleotide j at position i.
  - WAM: $p_{j,k}(i)$ is probability of nucleotide k at position i conditional on nucleotide j at position i-1.
- Decision tree (maximal dependence decomposition) for donor sites.
- Different model parameters for regions with different GC content.

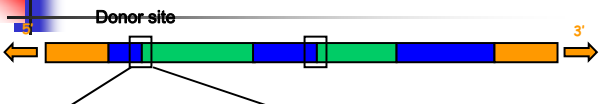
Splice Site Detection

(http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

Donor: 7.9 bits
Acceptor: 9.4 bits
(Stephens & Schneider, 1996)

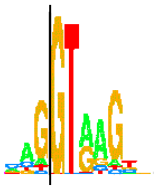


Splice site detection



Position

%	-8	…	-2	-1	0	1	2	…	17
A	26	…	60	9	0	1	54	…	21
C	26	…	15	5	0	1	2	…	27
G	25	…	12	78	99	0	41	…	27
T	23	…	13	8	1	98	3	…	25



Weighted matrix

- Computed by measuring the frequency of every element of every position of the site (weight)

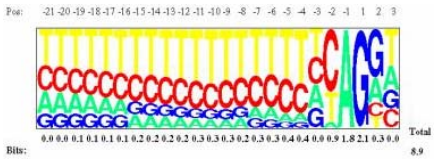
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

	1	2	3	4	5	6
A	0	6	0	3	4	0
C	0	0	1	0	1	0
G	1	0	0	3	0	0
T	5	0	5	0	1	6

- Score for any putative site is the sum of the matrix values (converted in probabilities) for that sequence (log-likelihood score)

Acceptor splice site

Consensus region from -20 to +3



A weighted matrix model for scoring potential splice sites.

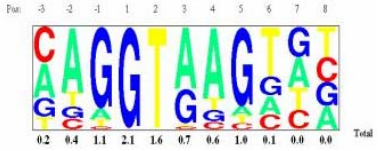
Promotor model

- Promoters
  - 30% of them lack apparent TATA signal
  - So, split model:
    TATA containing promoter
    - Generated with probability 0.7
    - 15 bp TATA-box WMM and 8 bp cap site WMM
  - TATA-less
    - Generated with probability 0.3
    - Modeled as intergenic-null regions of 40bp

Transcriptional and Translational Signals

- PolyA signal
  - 6 base pairs WMM (AATAAA)
- Translation Initiation signal
  - 12 base pairs WMM (6 base pairs prior to start codon)
- Translation termination signal
  - 1 of 3 stop codons according to observed frequency
  - Next 3 nucleotides using WMM

Donor splice site



Donor splice site model

- Consensus region -3 to +6 (3 on exon, 6 on intron)
- WMM or WAM not sufficient to model because of dependencies on non-adjacent nucleotides.

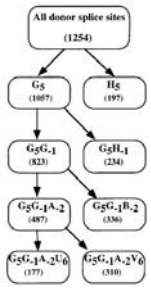
MDD algorithm

Absence of nucleotide G at position +5 implies a great consensus matching at position -1.

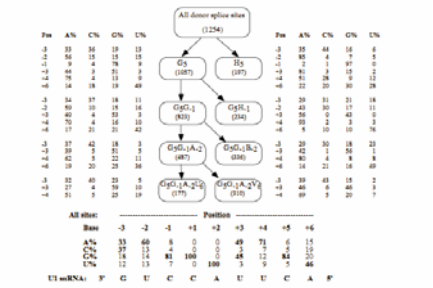
H = A/C/U

B=C/G/U

V=A/C/G



MDD algorithm



Using Genscan for gene finding

- Model's goal is to generate "Optimal Parse"
- Parse (X) consists of
  - Ordered set of states = $\{s_1, s_2, \ldots, s_n\}$ where $s_i \in \{S_j / j=1 \text{ to } 27\}$
  - Associated lengths (durations) $(d) = \{d_1, d_2, \ldots, d_n\}$
  - It generates DNA sequence O of length $L = \Sigma_{i=1 \text{ to } n} d_i$.

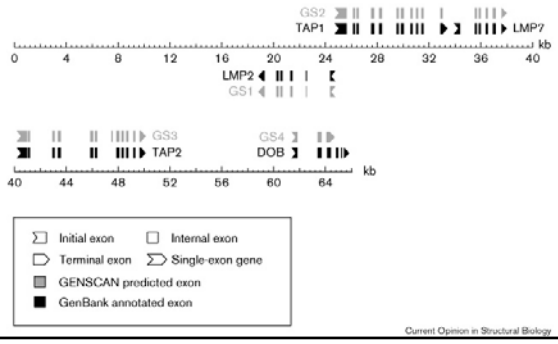
Running the model

- An initial state s_1 is chosen according to an initial distribution π on the states, i.e. $\pi_i = P(s_1=S_i)$
- A length distribution d_1 is generated conditional on $s_{1,i.e.}$ $f_{s1}(d_1)$
- A sequence segment s_1 of length d_1 is generated conditional of s_1 and d_1 i.e. $P(s_i|s_1,d_1)$
- Subsequent state s_2 is generated, conditional on s_1. First order Markov. $a_{ij} = P(s_{k+1}= S_j \,|s_k=S_i\}$

Using model

- Optimal parse can be computed by Viterbi algorithm for generalized HMMs (see Rabiner's extension in section 4D, pages 269-270).

Genscan output



Genscan

- The Genscan HMM model
- Training Genscan
- Validating Genscan

Evaluating gene finders

- Calculating accuracy of programs' predictions
- Several evaluation studies:
  - Burset and Guigó, 1996 (vertebrate sequences)
  - Pavy *et al.*, 1999 (*Arabidopsis thaliana*)
  - Rogic *et al.*, 2001 (mammalian sequences)

Accuracy Metrics

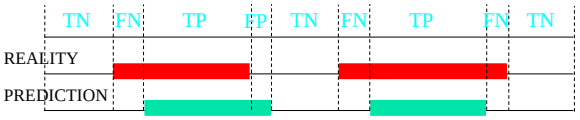
		actual class: positive	actual class: negative
predicted	positive	true positives (TP)	false positives (FP)
predicted	negative	false negatives (FN)	true negatives (TN)

$$\text{sensitivity} = \frac{TP}{\text{all pos}} = \frac{TP}{TP+FN}$$

$$\text{specificity} = \frac{TP}{\text{predicted pos}} = \frac{TP}{TP+FP}$$

Measures of Prediction Accuracy

Nucleotide level accuracy



Sensitivity $Sn = \frac{TP}{TP + FN}$

Specificity $Sp = \frac{TP}{TP + FP}$

Measures of Prediction Accuracy

Exon level accuracy

$$ESn = \frac{TE}{AE} \qquad ESp = \frac{TE}{PE}$$

$$AC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TP}{TP+FP} + \frac{TN}{TN+FP} + \frac{TN}{TN+FN}\right) - 1$$

$$CC = \frac{(TP * TN) - (FN * FP)}{((TP+FN) * (TN+FP) * (TP+FP) * (TN+FN))^{\frac{1}{2}}}$$

Evaluation Results

Programs	# of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	*Sp*	*AC*	*CC*	*ESn*	*ESp*	*(ESn+Esp)/2*	*ME*	*WE*	*PCn*	*PCp*	*OL*
FGENES	195 (5)	0.86	0.88	0.84 ± 0.19	0.83	0.67	0.67	0.67 ± 0.32	0.12	0.09	0.20	0.17	0.02
GeneMark.hmm	195 (0)	0.87	0.89	0.84 ± 0.18	0.83	0.53	0.54	0.54 ± 0.36	0.13	0.11	0.29	0.27	0.09
Genie	195 (15)	0.91	0.90	0.89 ± 0.16	0.88	0.71	0.70	0.71 ± 0.30	0.19	0.11	0.15	0.15	0.02
Genscan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.32	0.08	0.09	0.21	0.19	0.02
HMMgene	195 (5)	0.93	0.93	0.91 ± 0.13	0.91	0.76	0.77	0.76 ± 0.30	0.12	0.07	0.14	0.14	0.02
Morgan	127 (0)	0.75	0.74	0.70 ± 0.21	0.69	0.46	0.41	0.43 ± 0.26	0.20	0.28	0.28	0.25	0.07
MZEF	119 (8)	0.70	0.73	0.68 ± 0.21	0.66	0.58	0.59	0.59 ± 0.28	0.32	0.23	0.08	0.16	0.01

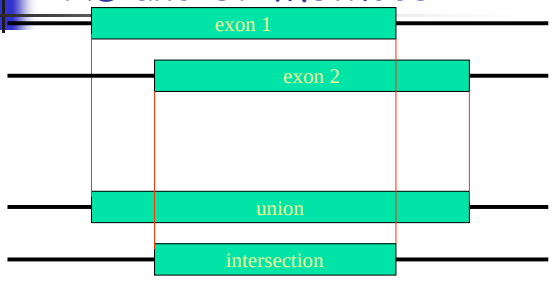
Genscan and Chromosome 22

- I. Dunham, Nature 402:489-95, 1999
- Chromosome 22
  - Annotated genes: 94% predicted partially
  - Annotated exons: 84% predicted partially
  - Predicted exons: 30% more than annotated exons. How many of them are real exons?

Integrated approaches for gene finding

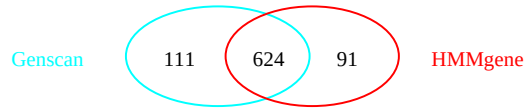
- Programs that integrate results of similarity searches with *ab initio* techniques (GenomeScan, FGENESH+, Procrustes)
- Programs that use synteny between organisms (ROSETTA, SLAM)
- Integration of programs predicting different elements of a gene (EuGène)
- Combining predictions from several gene finding programs (combination of experts)

AND and OR Methods



Combining Genscan and HMMgene

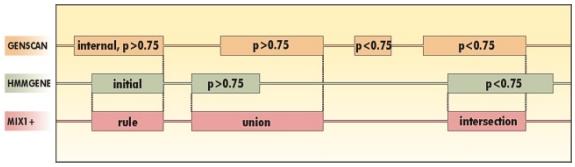
- High prediction accuracy as well as reliability of their exon probability make them good candidates.



- Genscan predicted 77% of exons correctly, HMMgene 75%, both 87%

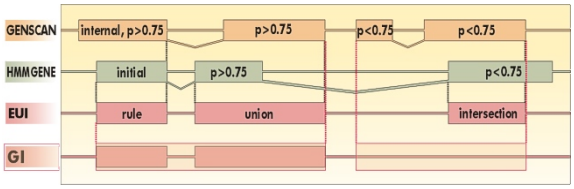
EUI Method (exon union – intersection)

1. Union of exons with $p \geq 0.75$
2. Intersection of exons with $p < 0.75$
3. Rule for initial exon



Gene intersection (GI) method

1. Intersection of genes
2. Apply EUI method to exons completely belonging to GI genes



EUI with reading frame consistency

1. Assign probabilities to GI genes. Determine position of acceptor and donor site in a reading frame.
2. GI gene with higher probability imposes the reading frame. Choose only EUI exons contained in GI genes that are in a chosen reading frame.

Results – Burset/Guigó dataset

METHODS	#no prediction	Nucleotide accuracy			Exon accuracy				
		Sn	Sp	AC	ESn	ESp	(ESn+ESp)/2	ME	WE
Genscan	8	0.94	0.93	0.92	0.78	0.81	0.80	0.09 (203)	0.05 (188)
HMMgene	38	0.93	0.94	0.92	0.81	0.83	0.82	0.14 (308)	0.04 (139)
EUI	20	0.94	**0.96**	**0.93**	**0.83**	**0.88**	**0.85**	0.12 (250)	**0.03** (98)
GI	43	0.91	**0.97**	**0.93**	**0.82**	**0.90**	**0.86**	0.18 (386)	**0.02** (67)
EUI_frame	27	0.93	**0.96**	**0.93**	**0.83**	**0.88**	**0.85**	0.13 (286)	**0.03** (87)

Summary: Eukaryotic gene finding

- Overall accuracy usually below 50%
  - Human gene finding is hardest
  - Very long introns, and lots of them
- Leading methods: HMMs and variants
- New ideas needed
- New opportunity: use sequence of related species