

How to design an HMM for a new problem

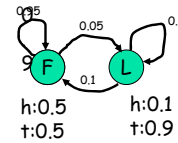
- Architecture/topology design:
 - What are the states, observation symbols, and the topology of the state transition graph?
- Learning/Training:
 - Fully annotated or partially annotated training datasets
 - Parameter estimation by maximum likelihood or by EM
- Validation/Testing:
 - Fully annotated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

(c) Devika Subramanian, 2006

69

HMM model structure

- Duration modeling



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

(c) Devika Subramanian, 2006

70

Duration modeling

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

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

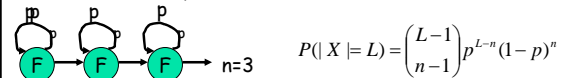
- This may be inappropriate for some applications.

(c) Devika Subramanian, 2006

71

Duration modeling

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

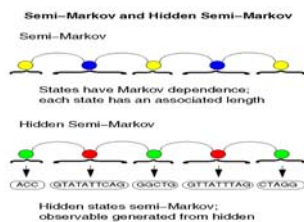


- Generated length distribution is a negative binomial.

(c) Devika Subramanian, 2006

72

Semi-Markov HMMs



(c) Devika Subramanian, 2006

73

Hidden Semi-Markov models

- Each state is associated with an explicit duration model of the form: $P(|X|=L)$, where $|X|$ is the length of the hidden state sequence in state X.

(c) Devika Subramanian, 2006

74

Genscan

- The Genscan HMM model
- Training Genscan
- Validating Genscan

(c) Devika Subramanian, 2006 75

Structure of a human gene

Structure of a Human Gene (PSA)

Exon-intron structure

(c) Devika Subramanian, 2006 76

Gene structure assumed by Genscan

Structure of a Typical Human Gene

(c) Devika Subramanian, 2006 77

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.

(c) Devika Subramanian, 2006 78

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

Genscan HMM

(c) Devika Subramanian, 2006 79

Genscan model components

- Vector of initial probabilities π
- State Transition probability Matrix T
- Set of length distributions $f_{Q(i)}$ conditional on state
- Sequence generating model $P(s|q,d)$ conditional on state and length.

(c) Devika Subramanian, 2006 80

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 intergenic length (bp)	83000	36000	5400	2600

(c) Devika Subramanian, 2006

81

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.

(c) Devika Subramanian, 2006

82

Transition probabilities

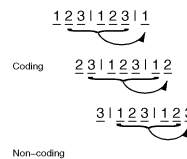
- Sure transitions are assigned probability 1.
- The others are set according to maximum likelihood values in training data.

(c) Devika Subramanian, 2006

83

Exon and intron models

Models of Coding and Non-Coding DNA



Phases of the exons

5th order inhomogeneous Markov model

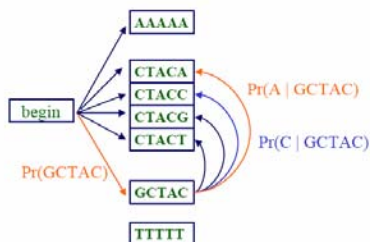
5th order homogeneous Markov model :

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

(c) Devika Subramanian, 2006

84

A Fifth Order Markov Chain



(c) Devika Subramanian, 2006

85

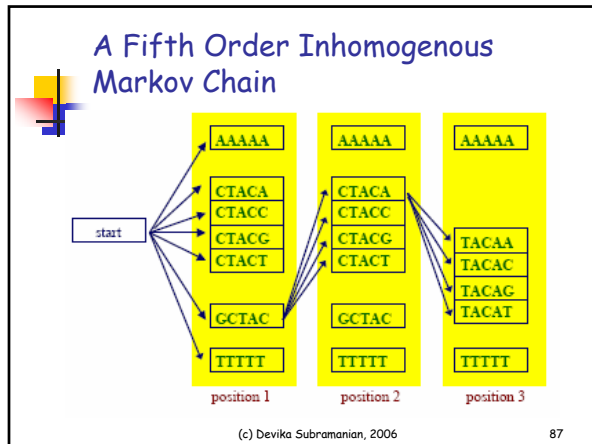
Inhomogeneous Markov Chains

- In the Markov chain models we have considered so far, the probabilities do not depend on where we are in a given sequence
- In an *inhomogeneous* Markov model, we have different distributions at different positions in the sequence.

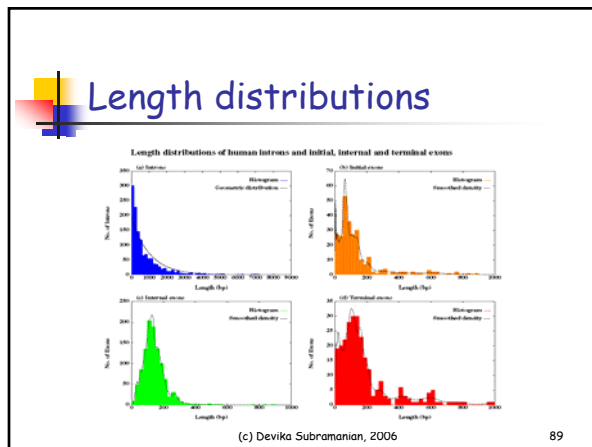
$$a_{x_1 x_2}^1 \quad a_{x_2 x_3}^2 \quad a_{x_3 x_4}^3 \quad a_{x_4 x_5}^1 \quad a_{x_5 x_6}^2$$

(c) Devika Subramanian, 2006

86



- ### Exon/intron/UTR model
- **Exons** -- inhomogeneous 3-periodic fifth order Markov model.
 - **Introns and intergenic regions** - homogeneous 5th order Markov model
 - **5' and 3' UTRs** - homogeneous 5th order Markov model
- (c) Devika Subramanian, 2006 88



- ### 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.
- (c) Devika Subramanian, 2006 90

- ### 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 ($3 \times \text{codons} + \text{phase}$)
- (c) Devika Subramanian, 2006 91

- ### Other length distributions
- 5' UTR -> Geometric with mean 769bp
 - 3' UTR -> Geometric with mean 457bp
- (c) Devika Subramanian, 2006 92

Genscan architecture (2)

- Weighted matrix and weighed arrays for acceptor splice site, polyA site and promoter region.
- Decision tree (maximal dependence decomposition) for donor sites.
- Different model parameters for regions with different GC content.

(c) Devika Subramanian, 2006

93

Signal models

- WMM (Weight Matrix Method)
 - $p_j(i)$ is probability of nucleotide j at position i .
 - Multiplicative.
- WAM (Weight Array Model)
 - Markov chains. $p_{j,k}(i-1,i)$ is probability of nucleotide k at position i conditional on nucleotide j at position $i-1$.
- MDD (Maximal Dependence Decomposition)

(c) Devika Subramanian, 2006

94

Weighted matrix

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

TACGAT		1	2	3	4	5	6
TATAAT	A	0	6	0	3	4	0
TATAAT	C	0	0	1	0	1	0
GATACT	G	1	0	0	3	0	0
TATGAT	T	5	0	5	0	1	6
TATGTT							

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

(c) Devika Subramanian, 2006

95

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

(c) Devika Subramanian, 2006

96

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

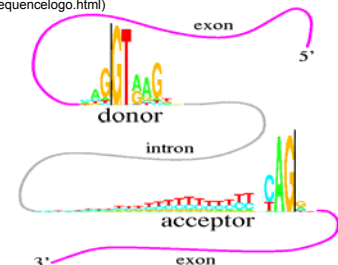
(c) Devika Subramanian, 2006

97

2. Splice Site Detection

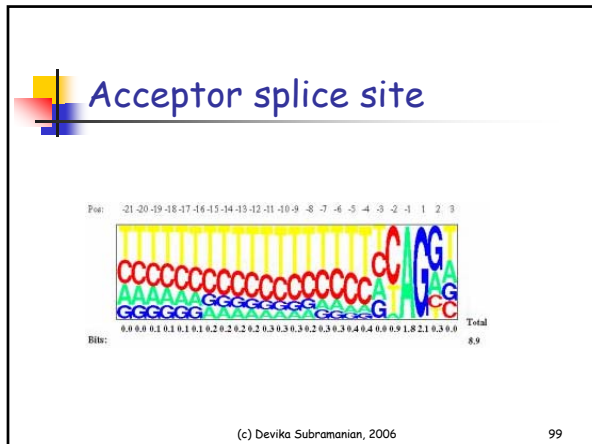
(<http://www-imb.ncifcrf.gov/~toms/sequencelogo.html>)

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

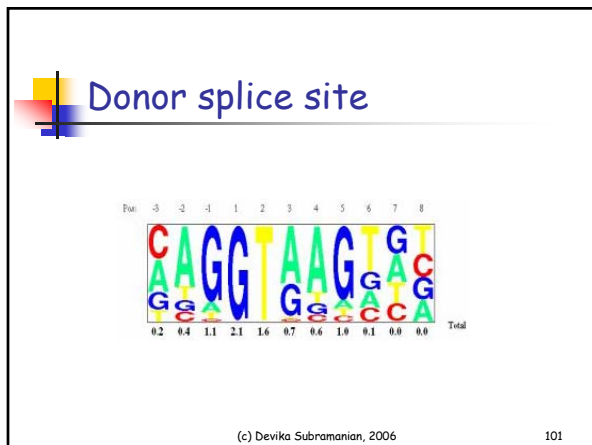


(c) Devika Subramanian, 2006

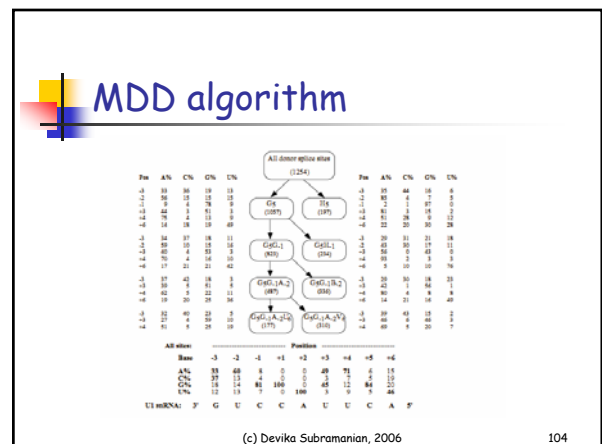
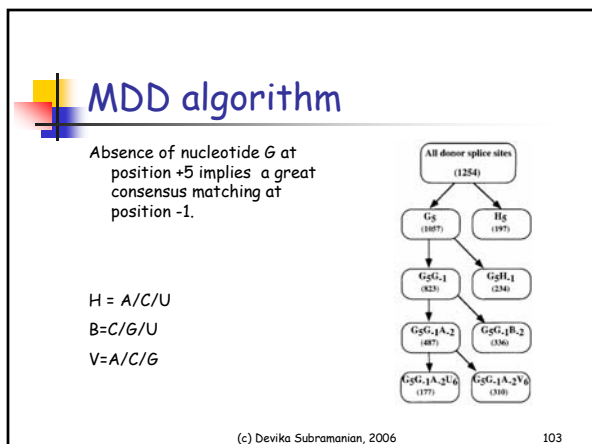
98



- ## Acceptor splice site model
- Consensus region from -20 to +3
 - Windowed second-order WAM model (WWAM)
- (c) Devika Subramanian, 2006 100



- ## 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.
- (c) Devika Subramanian, 2006 102



> HSKC11BE. Human gene for casein kinase II subunit beta (EC 2.7.1.37).

```

ggggctgagatgtaaatagaggagctggagagagagcttccaggtttgggtttgctttaagaaagggt
ggttcogattctcccggttgaggagcogaaatgaggaggaaccagggagggagggagga
gaacttgagctttactgacctgttctttttctagctgacctggatgagctcagagagaggtgtc
ctggattctcggttctgtggtctcgtggcctcgtggcaatgaaatctctcaggttccttcaacctcc
ctacttgcgcgttccacttcttcccccagaggttccctccacatctccacttctcaactgttctct
aaagctttatggagagagagtgagtggaactaggagagagaccccaagctacttctgctgagttggagtg
agaacaagcacaacagatgcagttgtgtgatataaggcatctgcccaggttcaaa
agaaggagattttgataagggttccctttgggttccatgtccctcagatgacttcaaca
ggacaatctggtgagctcctcaatggagcaggtccctccactgagcatgactctt
ggacctggaggtgaggttctcaggttctgtctagtttgggtggcactatttctctctcaaa
atctctatctttgccaatctcccttgggttctctgattcttcaaccccaatctca
tgccttatcttgatctccaccctgactcttctctagtttggtagctatatacactgttctctatgtttt
tgcagggttcagaagcaggttctcgggtccatgcccagatgttggatggggaaggcccaaaagta
gttctgagccaactgaatggccggctgggaggggagccctaggaaagctgaaaaca
agtgtgcatcttggccggctgtggcagatgcttggtagagacaccccacagagtgacctg
attggatgctttaccctgacctctggccgttaagagacacaggttccctgca
ggctggagagcctgttaaggaagctagctgagagagggggaagaccagaacttgg
cccggcccaarrtgggaaggaagcacaagaaatttggagagccactagtcagagaggggggacct
ctggacagagttggaagagtgccgacagaggttttgcacaaagaggttggcttctct
ttacatctactgccaaccccttccatgtatttggagagactttg
gttactgtctctgtgtactgtgaggaacacaggttgggtctcaagagaggttgg
ggaagcaccgtgtggcagctcttatgggaagaggttgggtctcaactctgagg
ggaggttagttaggaataggggtacctggcctgtgagctgtggctggctcagacat
cccaggtggaagcactgttgaagctactgcccacaatggatggatctcactcaag
acacatcacacaggaaggcctactctggcactctctctcaggttctcaggttctcagga
gtaccggcccaagagacctgccaaccagtttggctggagagagctatgaaggtca
aagaaagcccaagctcccccagagagggaggaattcttgaggtctgctctcc
cagaatcagggcaactccctgctgagtgacttgggaagatttatgattctgtgctgaggttacct
tatgtagaagttcttggctgagaagttgggaaccagaggtctttagcttgagcaggtccatagag
gagctcaggtgggaggtgggaatgcaggtgactggcagaccctgaatggggctcatgctgctcctct
ctgaccttgcctggcctaggctctcaggttcaaggttggcttaccagagctgaggttcaa
agcccgagcaacttcaagagcaggtcaagagagcttctgaggtttttagtttaaatgaagga
gtctttatctggtggagattgaataaagtagagaaagggcaagagctagctgctgctgctgctg
ggaggggggtggagctggcctggaaatcgggtccaccggccagggatgg

```

105

Exon emission models

- Inhomogeneous 3-periodic fifth order Markov model.
- Different model for C+G group I.
- Maintain phase.

(c) Devika Subramanian, 2006 106

Non-coding emission models

- For UTR, intergenic and intron regions,
 - Homogeneous fifth-order Markov model

(c) Devika Subramanian, 2006 107

Using Genscan for gene finding

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

(c) Devika Subramanian, 2006 108

Running the model

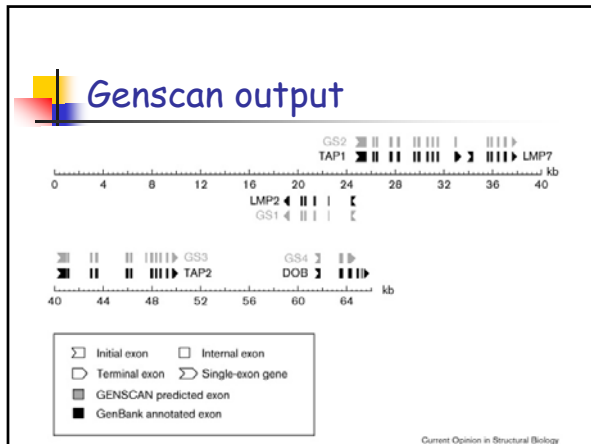
- An initial state s_1 is chosen according to an initial distribution π on the states, i.e. $\pi_i = P(s_1=S_i)$
- A length distribution d_1 is generated conditional on s_1 , i.e. $f_{s_1}(d_1)$
- A sequence segment s_1 of length d_1 is generated conditional of s_1 and d_1 i.e. $P(s_1|s_1, d_1)$
- Subsequent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

(c) Devika Subramanian, 2006 109

Using model

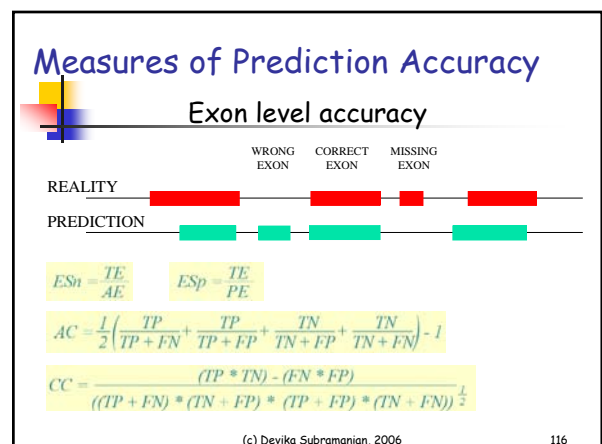
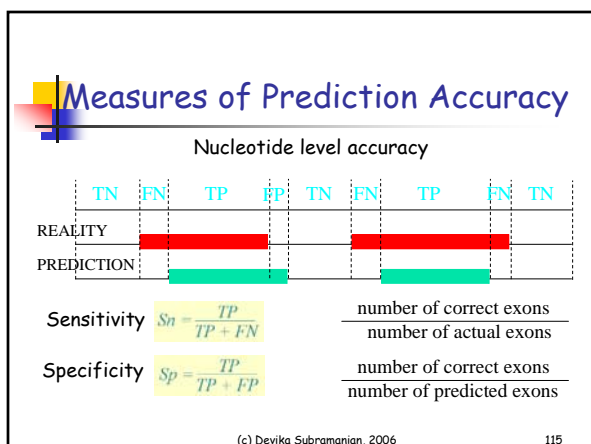
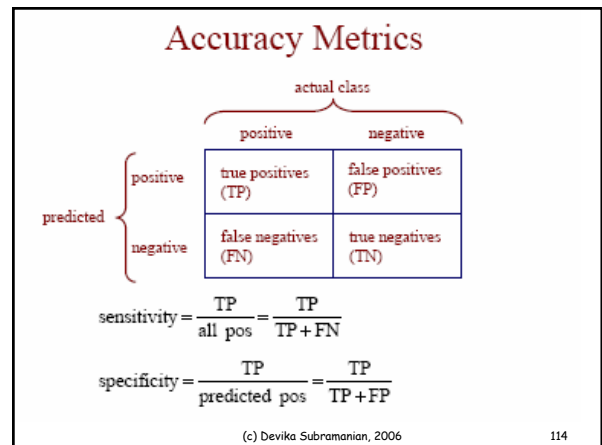
- Optimal parse can be computed by Viterbi algorithm (see Rabiner's extension in section 4D, pages 269-270).

(c) Devika Subramanian, 2006 110



- ## Genscan
- The Genscan HMM model
 - Training Genscan
 - Validating Genscan
- (c) Devika Subramanian, 2006 112

- ## 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)
- (c) Devika Subramanian, 2006 113



Evaluation Results

Programs	# of sequences	Nucleotide accuracy				Exon accuracy							
		Se	Sp	AC	CC	ESe	ESp	(ESe+ESp)/2	ME	WE	PCa	PCp	OE
PODSIES	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-ES	195 (6)	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
Gene	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
GeneScan	195 (3)	0.95	0.90	0.91 ± 0.12	0.91	0.70	0.70	0.70 ± 0.32	0.08	0.08	0.21	0.19	0.02
ISMGene	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
Mapon	127 (6)	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
MEXP	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

(c) Devika Subramanian, 2006

117

GeneScan 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?

(c) Devika Subramanian, 2006

118