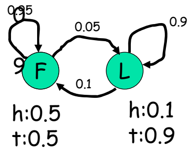
How to design an HMM for a new problem

- 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?

Duration modeling

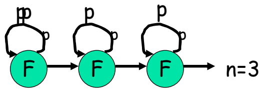
- The duration in state F follows an exponentially decaying distribution called a geometric distribution.

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

- This may be inappropriate for some applications.

Duration modeling

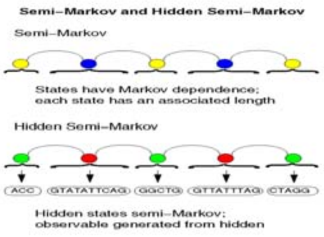
- To obtain non-geometric length distributions, we use an array of n F states, as follows:



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

- Generated length distribution is a negative binomial.

Semi-Markov HMMs


Semi–Markov and Hidden Semi–Markov
Semi–Markov
States have Markov dependence;
each state has an associated length
Hidden Semi–Markov
ACC
GTATATTCAG
GGCTG
GTTATTTAG
CTAGG
Hidden states semi–Markov;
observable generated from hidden

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.

Genscan

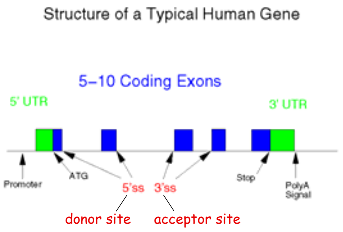
- The Genscan HMM model
- Training Genscan
- Validating Genscan

Structure of a human gene

Structure of a Human Gene (PSA)

Exon-intron structure

Gene structure assumed by Genscan

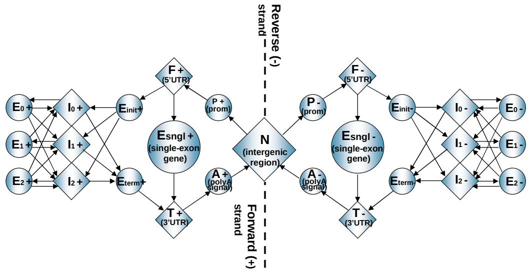


Genscan's architecture (1)

- 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 T
- Set of length distributions $f_{Q(i)}$ conditional on state
- Sequence generating model 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

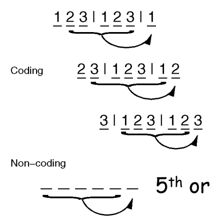
All other probabilities set to zero.

Transition probabilities

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

Exon and intron models

Models of Coding and Non–Coding DNA



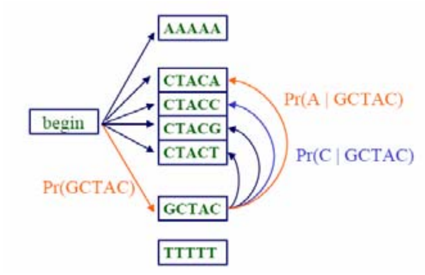
Phases of the exons

5th order inhomogeneous Markov model

5th order homogeneous Markov model :

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

A Fifth Order Markov Chain

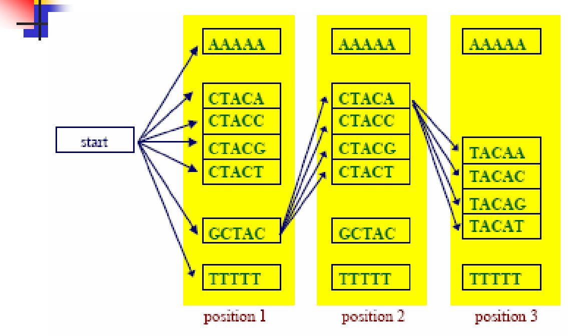


Inhomogenous Markov Chains

- 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.

$$a^1_{x_1x_2}\ a^2_{x_2x_3}\ a^3_{x_3x_4}\ a^1_{x_4x_5}\ a^2_{x_5x_6}$$

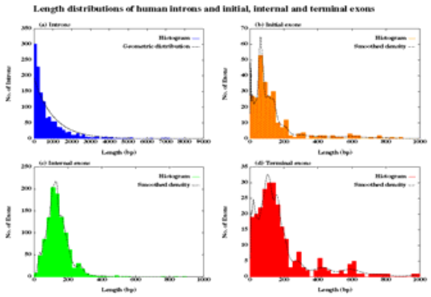
A Fifth Order Inhomogenous Markov Chain



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

Length distributions


Length distributions of human introns and initial, internal and terminal exons

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

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.

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)

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)

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

Promotor model

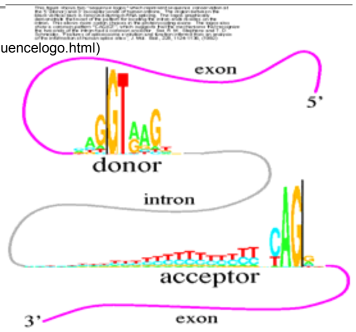
- Promoters
  - 30% of them lack apparent TATA signal
  - So, split model:

    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

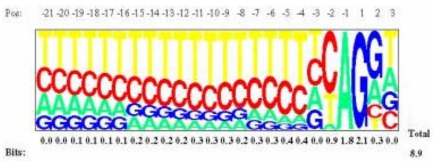
2. Splice Site Detection

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

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



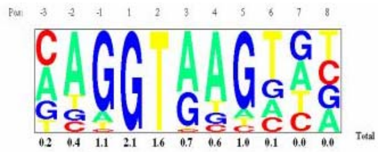
Acceptor splice site



Acceptor splice site model

- Consensus region from -20 to +3
- Windowed second-order WAM model (WWAM)

Donor splice site

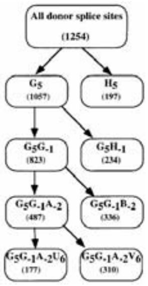


Donor splice site model

- 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.

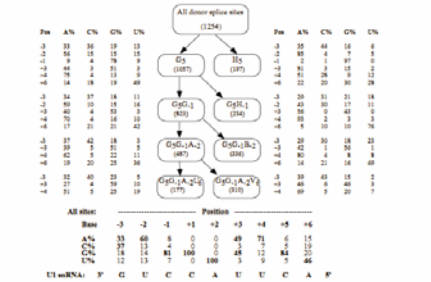


H = A/C/U

B=C/G/U

V=A/C/G

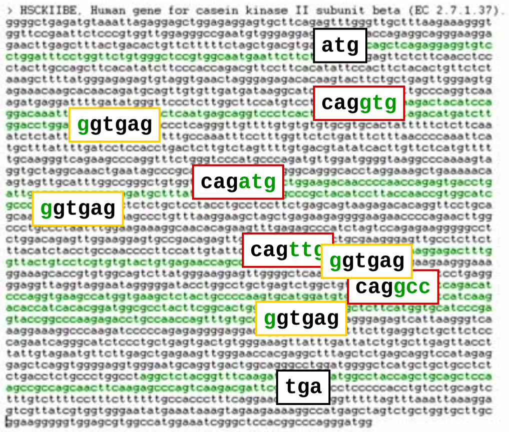
MDD algorithm



Pos	A%	C%	G%	U%
-3	33	36	19	13
-2	56	15	15	15
-1	9	4	78	9
+3	44	3	51	3
+4	75	4	13	9
+6	14	18	19	49
-3	34	37	18	11
-2	59	10	15	16
+3	40	4	53	3
+4	70	4	16	10
+6	17	21	21	42
-3	37	42	18	3
+3	39	5	51	5
+4	62	5	22	11
+6	19	20	25	36
-3	32	40	23	5
+3	27	4	59	10
+4	51	5	25	19

Pos	A%	C%	G%	U%
-3	35	44	16	6
-2	85	4	7	5
-1	2	1	97	0
+3	81	3	15	2
+4	51	28	9	12
+6	22	20	30	28
-3	29	31	21	19
-2	43	30	17	11
+3	56	0	43	0
+4	93	2	3	3
+6	5	10	10	76
-3	29	30	18	23
+3	42	1	56	1
+4	80	4	8	8
+6	14	21	16	49
-3	39	43	15	2
+3	46	6	46	3
+4	69	5	20	7

All sites:				Position						
	Base	-3	-2	-1	+1	+2	+3	+4	+5	+6
	A%	33	60	8	0	0	49	71	6	15
	C%	37	13	4	0	0	3	7	5	19
	G%	18	14	81	100	0	45	12	84	20
	U%	12	13	7	0	100	3	9	5	46
U1 snRNA:	3'	G	U	C	C	A	U	U	C	A 5'



```
> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggag          accagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtga          cagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttct          gagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatc              tgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcct              agactacatcca
ggacaaatt           ctcaatgagcaggtccctcact              agacatgatctt
ggacctgga           ccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctatt           tttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgc           ggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggt           ctggaagacaaccccaaccagagtgacctg
attg           gatgctttat           gcccgctacatccttaccaaccgtggcatc
gccc           tctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaa           agccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaagggggcct
ctggacagagttggaaggagtgccgacagagttc           tgcgaagggagttgcctcttct
ttacatctacctgccaaccccttccattgtattc                      aggagactttg
gttactgtcctcgtgtgtactgtgagaaccagcc                      aagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaa                cctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctg             cagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtc            catcaag
acaccatcacacggatggcgcctacttcggcactgc           gctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcc           agggagagtcattaagggtca
aaggaaaggcccaagatcccccagagaggggaggac           ttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtggggaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagat       atggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcc       cctcccccacctgtcctgcagtc
tttgtcttttcctttctttttgccacccttcaggaac         ggtttttagtttaaattaaagga
gtcgttatcgtggtgggaatatgaaatataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
ggaagggggtggagcgtggccatggaaatcgggctccacggcccagggatgg
```

Exon emission models

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

Non-coding emission models

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

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 ε $\{S_j$ / j=1 to 27$\}$
  - Associated lengths (durations) (d) = $\{d_1,d_2,\ldots,d_n\}$
  - It generates DNA sequence O of length $L = \Sigma_{i=1\text{ to }n} d_i$.

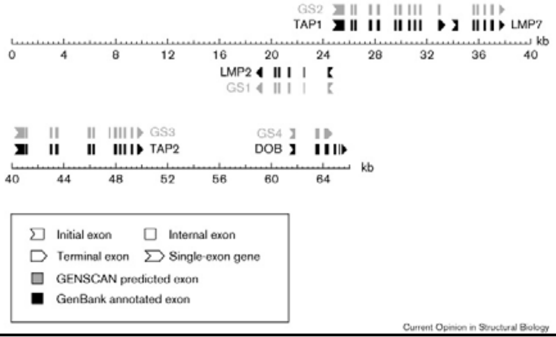
Running the model

- 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 (see Rabiner's extension in section 4D, pages 269-270).

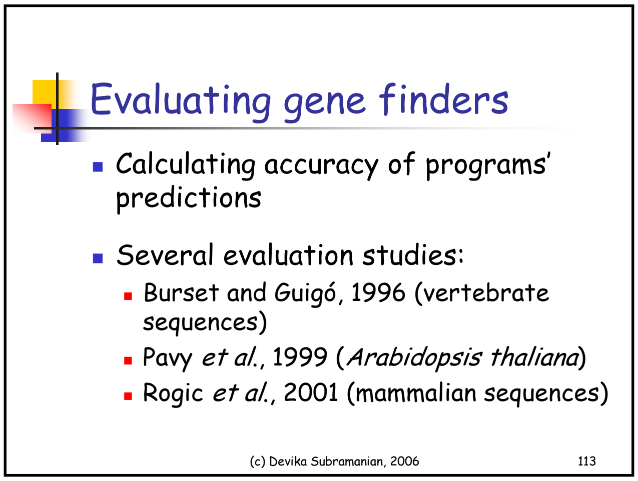
Genscan output



Current Opinion in Structural Biology

Genscan

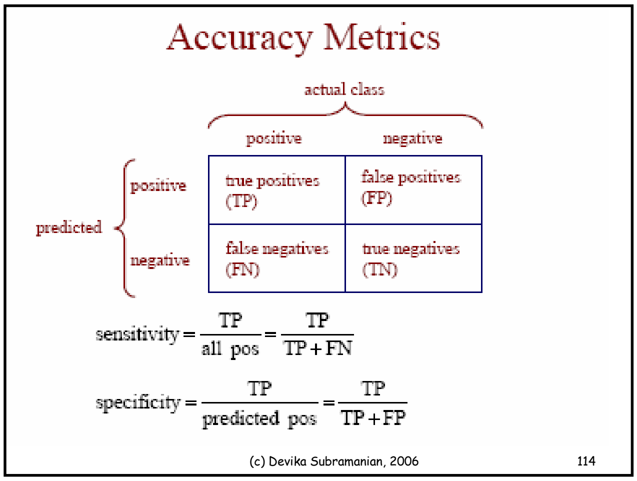
- The Genscan HMM model
- Training Genscan
- Validating Genscan



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)

(c) Devika Subramanian, 2006 113



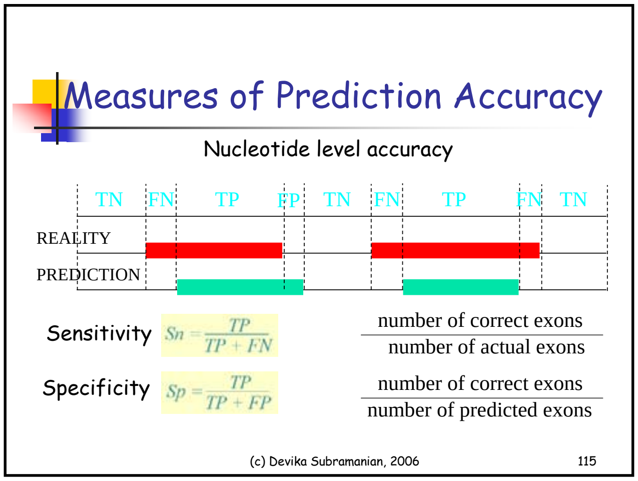
Accuracy Metrics

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

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

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

(c) Devika Subramanian, 2006 114



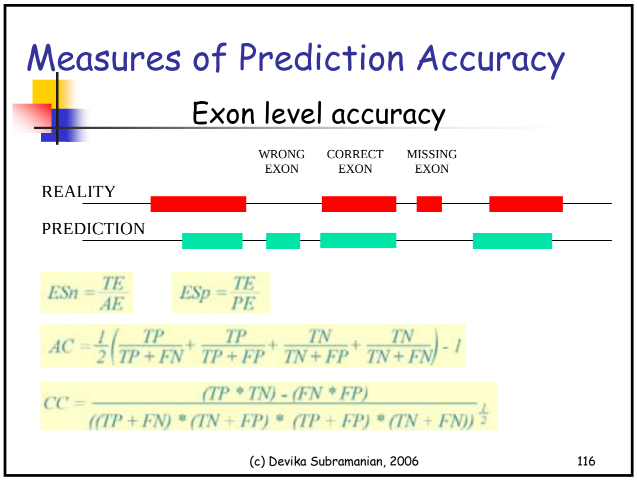
Measures of Prediction Accuracy

Nucleotide level accuracy

Sensitivity $Sn = \frac{TP}{TP + FN}$ $\frac{\text{number of correct exons}}{\text{number of actual exons}}$

Specificity $Sp = \frac{TP}{TP + FP}$ $\frac{\text{number of correct exons}}{\text{number of predicted exons}}$

(c) Devika Subramanian, 2006 115



Measures of Prediction Accuracy

Exon level accuracy

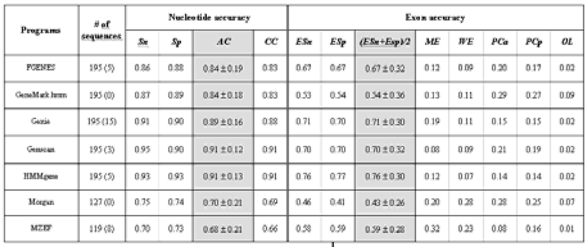
$ESn = \frac{TE}{AE}$ $ESp = \frac{TE}{PE}$

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

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

(c) Devika Subramanian, 2006 116

Evaluation Results



Programs	# of sequences	Nucleotide accuracy				Exon accuracy							
		Sn	*Sp*	*AC*	*CC*	*ESn*	*ESp*	*(ESn+Esp)/2*	*ME*	*WE*	*PCa*	*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?