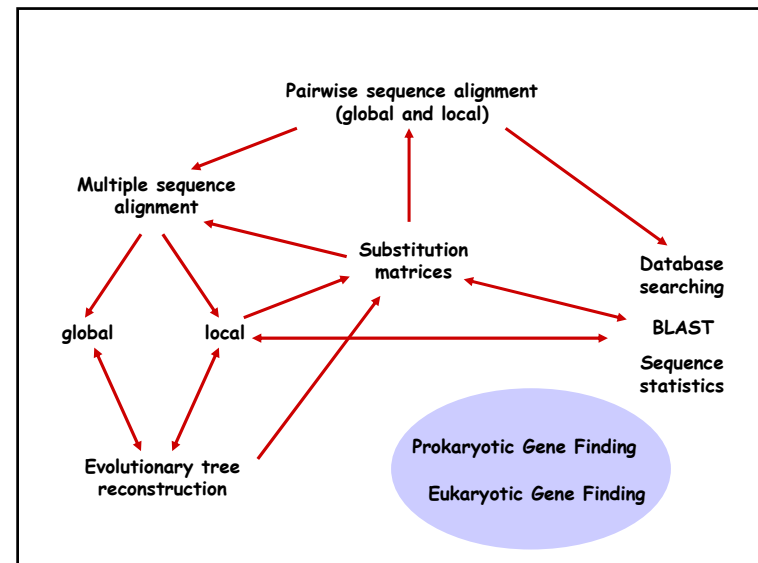


- Tues, Nov 28:
Gene Finding 1 Online FCE's: Thru Dec 11
- Thurs, Nov 30:
Gene Finding 2
- Tues, Dec 5:
PS5 due in my office at 5pm.
Project presentation preparation – no class
- Thurs, Dec 7
Final papers due
Project presentations
- Monday Dec 18
8:30am – 11:30am Final Exam, DH 1211



What is a Gene?

Snyder and Gerstein, Science 2003

- Something that encodes a heritable trait
- One gene, one enzyme
- One gene, one polypeptide
- One gene, one product (include RNA products)
- “A complete chromosomal segment responsible for making a functional product”
 - coding region
 - regulatory region
 - expressed product
 - functional product

Gene Finding Questions

- Identify protein coding region
- Identify Open Reading Frame
- Predict mRNA (including UTR's)
- Predict intron/exon structure
Eukaryotes only
- Regulatory signals
- Protein sequence

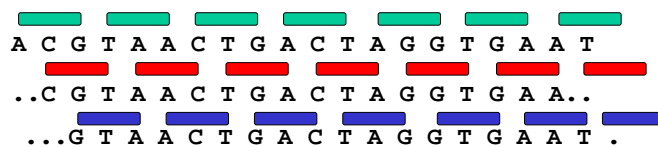
Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Prokaryotic Gene Finding

- **Identify Open Reading Frames (ORFs)**
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Reading Frames



- Each grouping of the nucleotides into consecutive triplets constitutes a reading frame.
- Three reading frames in the 5' → 3' direction
- Three in the reverse direction on the opposite strand.

Open Reading Frames

An ORF is a contiguous set of codons, each specifying an amino acid (starting with ATG).

GGAGCATGGTGCACCTGACTCCTGAGGTGACTTAGAC

M V H L T P E V T Stop

All coding sequences are ORF's, but not all ORF's encode proteins

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- **Coding Statistics**
- Identify individual gene architecture features
- Assemble an integrated gene description
- Homology

Coding Statistics

Fickett and Tung, 1992
Guigo and Fickett, 1995
(Electronic reserves)

- Codon usage
 - Determine codon (triplet) frequencies in known coding regions
 - Compare with codon frequencies in sliding window

ccgcctggcgtcgcggttgtttttcaatctctttcatctgca

CodingStatistics

Fickett and Tung, 1992
Guigo and Fickett, 1995
(Electronic reserves)

- Codon usage Species specific
- Codon pair preference Species specific

ccgcctggcgtcgcggttgtttttcaatctctttcatctgca

CodingStatistics

Fickett and Tung, 1992
Guigo and Fickett, 1995
(Electronic reserves)

- Codon usage Species specific
- Codon pair preference Species specific
- Amino acid usage Species specific

Gly Val Ala Val Cys Phe Ser
ccgcctggcgtcgcggttgtttttcaatctctttcatctgca

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

- Codon usage Species specific
- Codon pair preference Species specific
- Amino acid usage Species specific
- Amino acid pair preference Species specific

ccgcctggcggtcgcggtttgtttttcatctctcttcatctgca

Gly	Val	Ala	Val	Cys	Phe	Ser	Ser
-----	-----	-----	-----	-----	-----	-----	-----

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

- Codon usage Species specific
- Codon pair preference Species specific
- Amino acid usage Species specific
- Amino acid pair preference Species specific
- Third position Any organism
 - 3rd base tends to be the same much more often than chance

ccgcctggcggtcgcggtttgtttttcatctctcttcatctgca

Coding Statistics continued

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

CG content Species specific

In *E. coli*:

Coding regions are embedded in segments of 53% G+C content, about 1000 ases long

Non-coding regions are embedded in segments of 46% G+C content, about 500 bases long

aa, at, ta, tt occur more frequently than expected in coding regions

tgccgcctggcggtcgcggtttctttttcatctctcttcatctg
acggcggaaccgcagcgccaagaaaaagtagagagaagtagac

CodingStatistics

Fickett and Tung,1992
Guigo and Fickett,1995
(Electronic reserves)

- Codon usage Species specific
- Codon pair preference Species specific
- Amino acid usage Species specific
- Amino acid pair preference Species specific
- Third position Any organism
- CG content Species specific

Look for variations in these measures in coding and non-coding regions (both *intergenic* and *intragenic*).

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- **Identify individual gene architecture features**
- Assemble an integrated gene description
- Homology



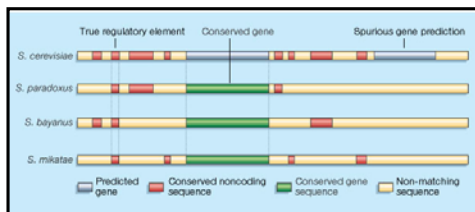
Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- **Assemble an integrated gene description**
- Homology

Prokaryotic Gene Finding

- Identify Open Reading Frames (ORFs)
- Coding Statistics
- Identify individual gene architecture features
- Assemble an integrated gene description
- **Homology**

Homology



Salzberg, Nature 2003

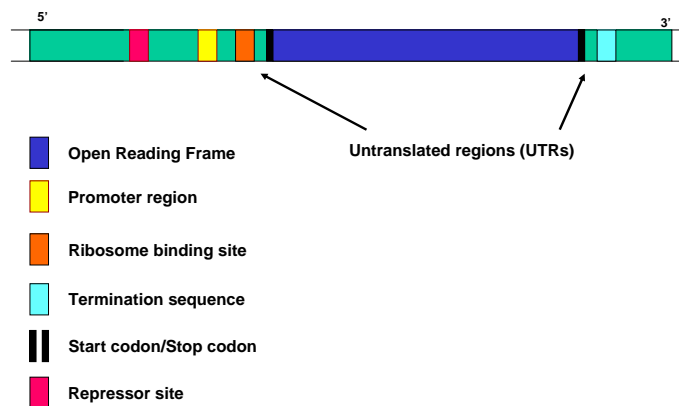
Prokaryotic Gene Finding

- Genome length: 0.5Mbp – 10Mbp
- Coding density: ~90%
- Long ORFs are usually real genes

Early approaches

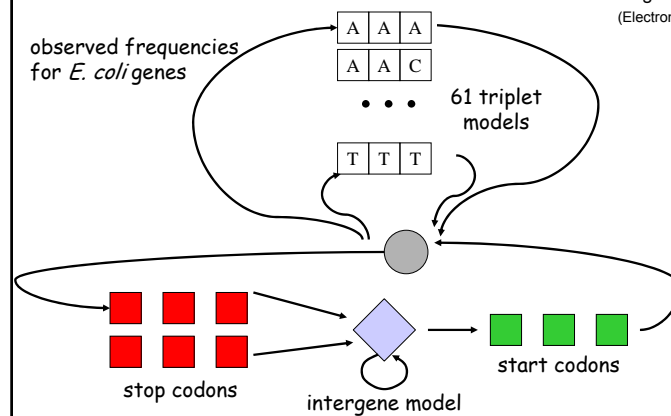
- Identify ORFs
- Score windows with coding statistics
- Identify gene structure elements
- Parse into a coherent gene model surrounded by intergenic DNA.

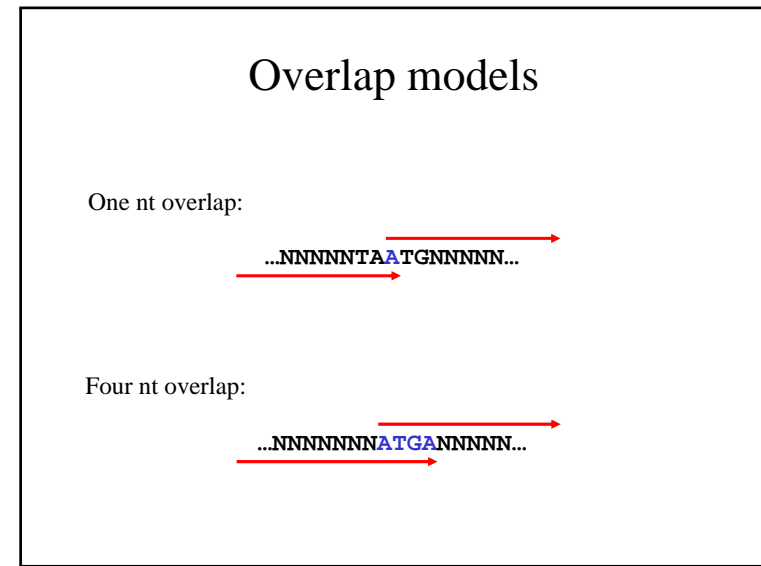
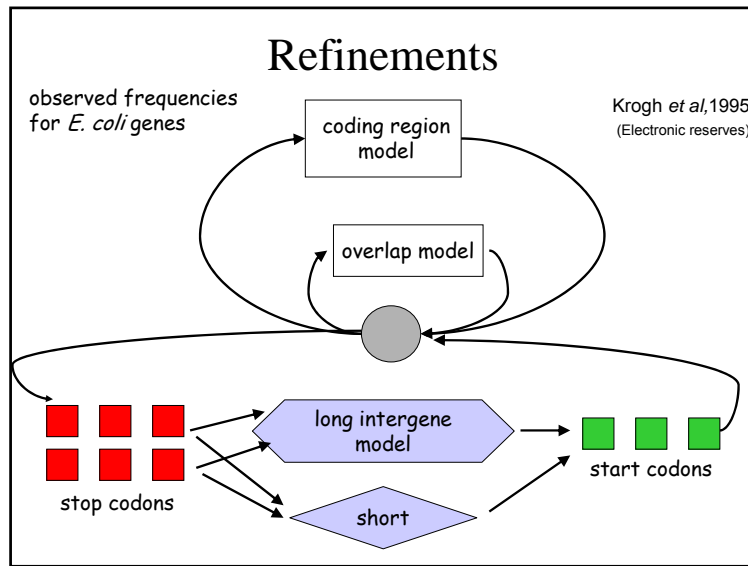
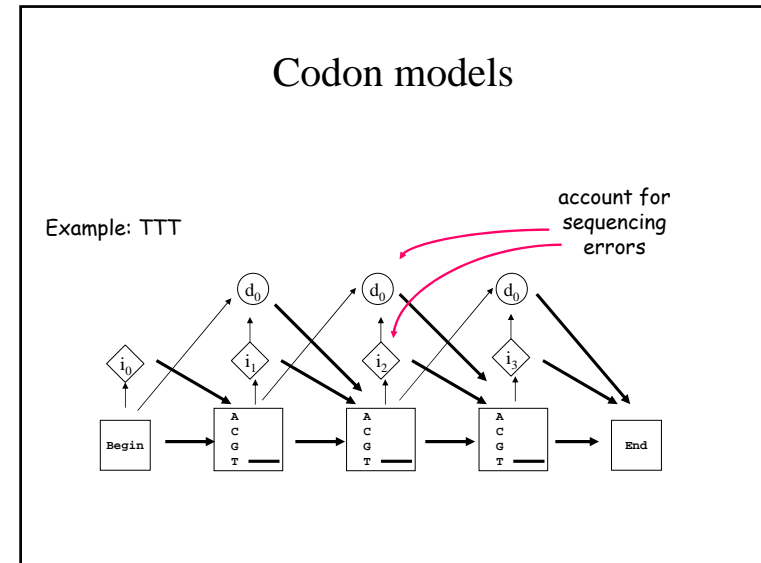
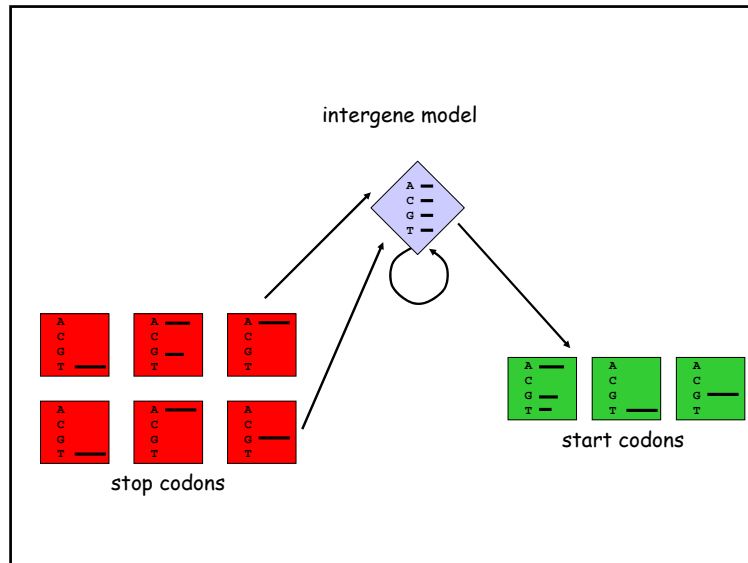
Prokaryotic gene model



An HMM that finds genes in *E. coli*

Krogh et al, 1995
(Electronic reserves)



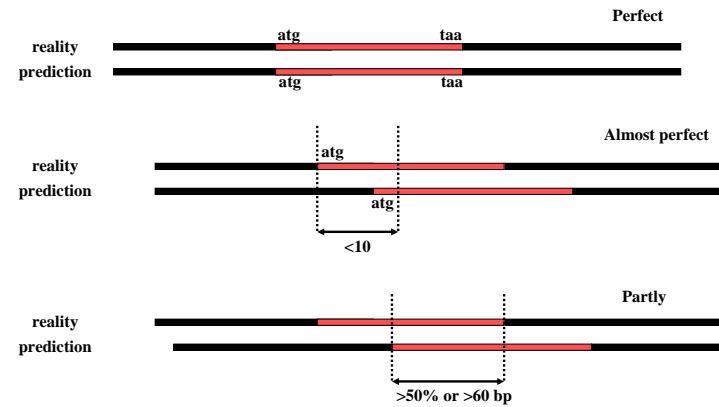


Parameter estimation

- Data: 429 *E. coli* contigs
- Trained intergenic models with non-coding DNA
- Transitions into codon models set to observed codon frequencies in coding regions

	Training	Test
Contigs	300	129
Base pairs	1,271,528	324,684
Genes	1007	251
Av length	1008	1015

Performance measures



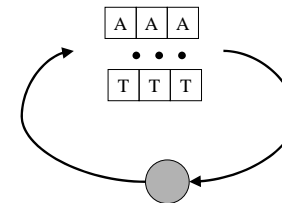
Prediction results on test set

Perfect	Almost perfect	Partly	Missed	Predicted
86%	5.1%	4.7%	4.2%	286

- About half of the false negatives were genes with unusual codon usage.
- Overlapping genes are sometimes mistaken as frameshifts.
- Predicted genes: 286
 - 95: recently discovered *E. coli* genes.
 - 63: similar to known proteins
 - 128: false positives?

Outstanding Problems

- Model cannot account for drift in CG content
- Does not take position dependencies into account



Outstanding Problems

- Model cannot account for drift in CG content
- Does not take position dependencies into account
- Solution:
 - k th order Markov chain
 - looks back k positions

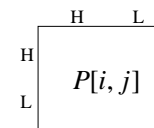
First-order Markov chain

Example: transmembrane region model



H: hydrophobic
L: hydrophilic

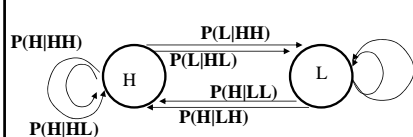
Transition matrix:



$$P(x_t = i | x_{t-1} = j)$$

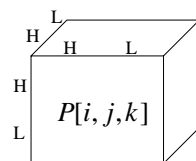
Second-order Markov chain

Example: transmembrane region model



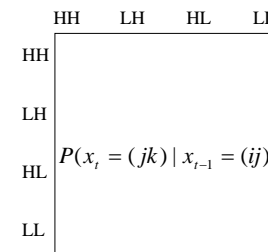
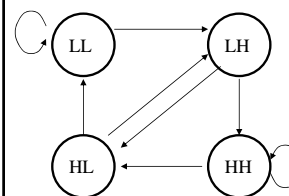
H: hydrophobic
L: hydrophilic

Transition matrix:



$$P(x_t = i | x_{t-1} = j, x_{t-2} = k)$$

A second-order Markov chain can be expressed as a first order Markov chain with more states and transitions



Glimmer

Salzberg *et al*,1998

- Prokaryotic gene finder
- Finds 98% of all genes in a bacterial genome
- Genome independent
 - Uses all large, non-overlapping ORFs as training data
- *k*th order Markov chain
 - (looks back *k* positions)
- Higher order Markov models require *more* training data

Some Problems

Snyder and Gerstein, *Science* 2003

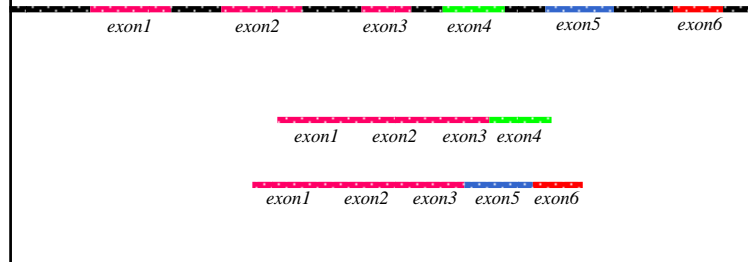
- Overlapping genes

aggcctatgacgcctctcccagcatgggacctgaggctcctgtccccactagtagggcctgc

Some Problems

Snyder and Gerstein, *Science* 2003

- Overlapping genes
- Alternate splicing



Some Problems

Snyder and Gerstein, *Science* 2003

- Overlapping genes
- Alternate splicing
- Pseudogenes

cctatgacgcctctcccagcatgggacctgaggctcctgtccccactagtagggcctgctcc

cctatgacgcctctcccagcatgagcctgaggctcctgtccccactagtagggcctgctcc

cctatgacgcctctcccagcatgagcctgaggctcctgtccccactagtagggcctgctcc

Gene Finding Challenges

- Small protein-coding genes (<100 aa's)
- Non-protein-coding RNA genes
- Regulatory regions
- Genes with sparse conserved positions and little sequence similarity; e.g., beta-defensins

Salzberg, Nature, 2003

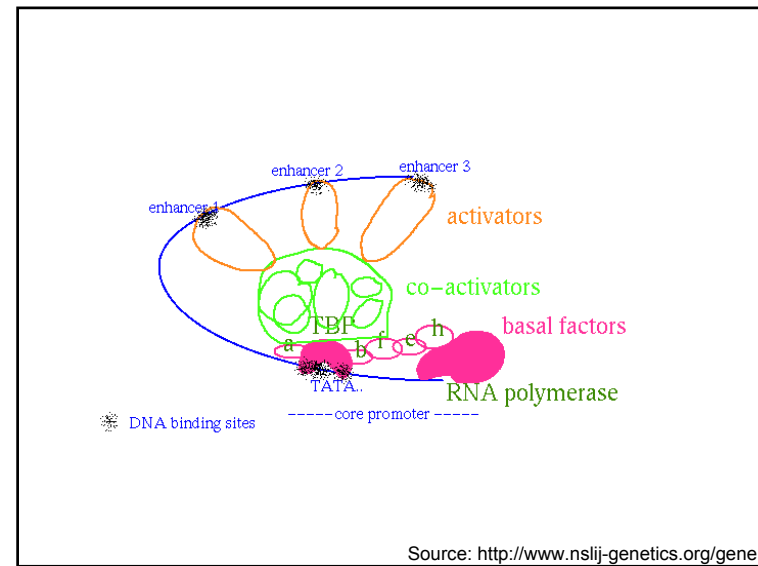
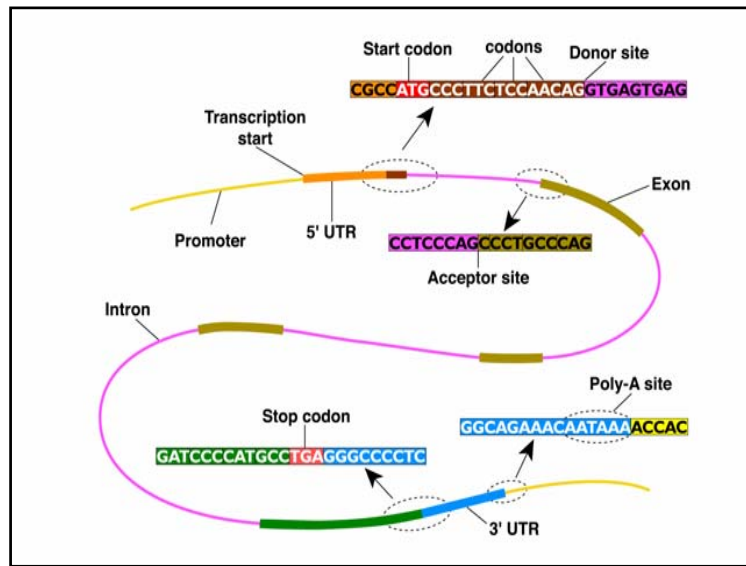
Schutte *et al.*, PNAS, 2001

Prokaryotic vs. Eukaryotic Genes

- | | |
|--|---|
| <ul style="list-style-type: none"> • Prokaryotes <ul style="list-style-type: none"> – small genomes (0.5Mb to 10Mb) – high gene density (90%) – no introns (or splicing) – no RNA processing – simple regulatory regions – most long ORF's are genes | <ul style="list-style-type: none"> • Eukaryotes <ul style="list-style-type: none"> – large genomes – low gene density (3% - 50%) – intron/exon structure – splicing – complex regulatory regions |
|--|---|

Genomic data:

Must handle multiple genes and/or gene fragments in input sequence.



Source: <http://www.nslj-genetics.org/gene>

Typical human gene sizes	
Average gene length	30kb
Coding region	1-2kb
Exon length	150 - 200 bp
Exon count	5-6
Single exon genes	8%

Genome statistics			
	Size	Gene number	Density (1 gene per)
Human	3300Mb	30K	100,000
Fly	180Mb	13.6K	9000
<i>C. elegans</i>	97Mb	19.1K	5000
Yeast	12Mb	6.3K	2000
<i>E. coli</i>	4.8Mb	3.2K	1400
<i>H. influenzae</i>	1.8Mb	1.7K	1000

http://www.ornl.gov/TechResources/Human_Genomefaq/compngen.html

GenScan

Burge and Karlin, 1997

Architecture:

- Individual modules: intergenic region, promoter, 5'UTR, exon/intron, post-translation region
- Semi Hidden Markov Model – various length distributions
- Different statistical models for each module:
 - weight matrices + extensions, 3-periodic 5th order Markov chains

Incorporates:

- Descriptions of transcriptional, translational and splicing signals
- Compositional features of exons, introns, intergenic, C+G regions

GenScan

Burge and Karlin, 1997

Larger predictive scope than previous models

- Partial genes
- Multiple genes separated by intergenic DNA
- Genes on either/both DNA strands

Proposed pipeline

- Screen for repetitive elements
- Predict protein sequences with GENSCAN
- BLAST predictions to find homologs
- Refine using spliced alignment of prediction with homolog (e.g., Gelfand, Mironov, Pevzner, 96)
- Verify experimentally

GenScan States

- **N**: intergenic region
- **P**: promoter
- **F**: 5' untranslated region
- **E_{sngl}**: single exon (intronless) (translation start -> stop codon)
- **E_{init}**: initial exon (translation start -> donor splice site)
- **E_k**: phase k internal exon (acceptor splice site -> donor splice site)
- **E_{term}**: terminal exon (acceptor splice site -> stop codon)
- **I_k**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

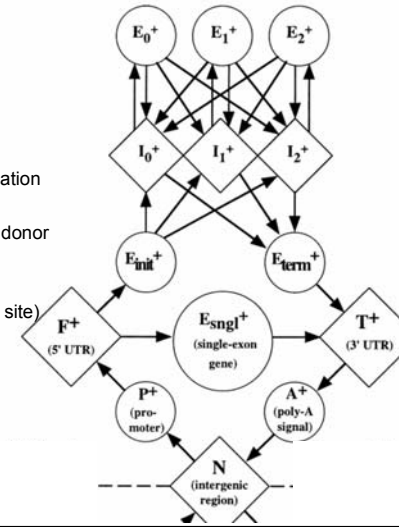


Fig. 3, Burge and Karlin 1997

GenScan Training Set

2.5M base pairs

142 Single Exon Genes (SEGs)

238 multi-exon gene

1492 Exons

1254 Introns

An additional 1619 coding sequences (no introns)

Promoter model based on published sources.

Initial and transition probabilities

Trained separately for four categories of G+C content

< 43% (G+C)

43% - 51% (G+C)

51% - 57% (G+C)

> 57% (G+C)

- Gene density varies with G+C content
- Genes in A+T rich regions had fewer introns

GenScan States

- **N**: intergenic region
- **P**: promoter
- **F**: 5' untranslated region
- **E_{sngl}**: single exon (intronless) (translation start -> stop codon)
- **E_{init}**: initial exon (translation start -> donor splice site)
- **E_k**: phase k internal exon (acceptor splice site -> donor splice site)
- **E_{term}**: terminal exon (acceptor splice site -> stop codon)
- **I_k**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

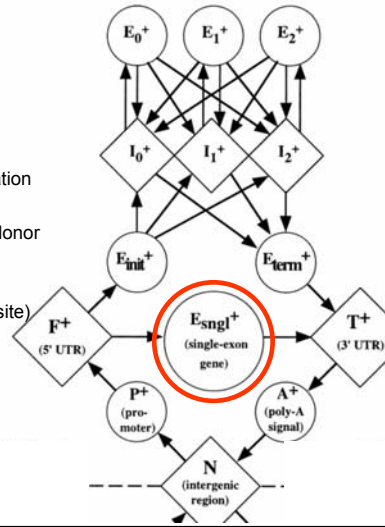
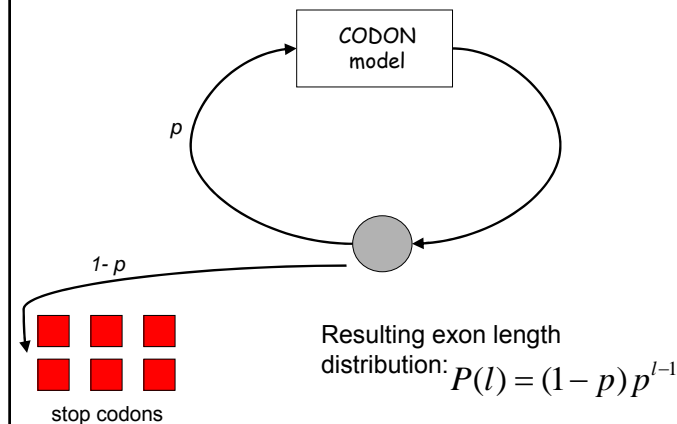


Fig. 3, Burge and Karlin 1997

How to model sequences with lengths that are not geometrically distributed?



Semi-hidden Markov model

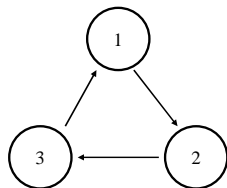
- Set of states: Q_1, Q_2, \dots
- Transition matrix $P(Q(t)|Q(t-1))$
- Initial distribution $P(Q(0))$
- Each state has
 - a length distribution
 - a sequence generating model
- Emission:
 - Each state emits a **sequence**, according to a particular distribution, of length, d , according to a particular length frequency distribution

Semi-hidden Markov model cont'd

- A parse ϕ of length L is
 - A state sequence: Q_1, Q_2, \dots
 - A sequence of lengths: d_1, d_2, d_3, \dots
- An observed sequence, s , is scored using a modified Viterbi algorithm

$$\phi_{opt} = \arg \max P(\phi | s)$$

Single Gene Exons



- 3-periodic
- 5th-order Markov chain
- Length distribution:
 - taken empirically from single exon lengths
- Sequence model
 - Trained with coding sequences

GenScan States

- **N**: intergenic region
- **P**: promoter
- **F**: 5' untranslated region
- **E_{sngl}**: single exon (intronless) (translation start -> stop codon)
- **E_{init}**: initial exon (translation start -> donor splice site)
- **E_k**: phase k internal exon (acceptor splice site -> donor splice site)
- **E_{term}**: terminal exon (acceptor splice site -> stop codon)
- **I_k**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

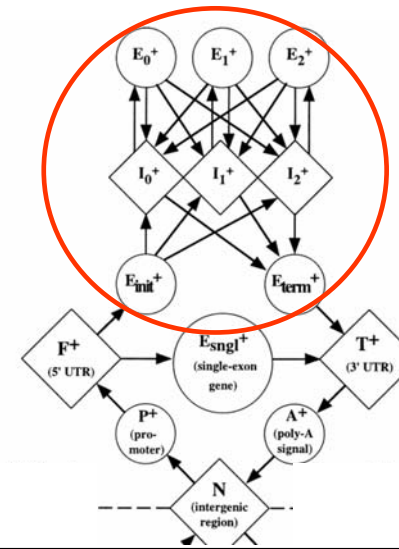


Fig. 3, Burge and Karlin 1997

Intron/Exon model

phase 0 intron:

GAGCATGACTXXXXXXXXXXXXXXXXGAGGTGCACCTGGTGA CTTAGAC . .

phase 1 intron:

GAGCATGACTGXXXXXXXXXXXXXXXXAGGTGCACCTGGTGA CTTAGAC . .

phase 2 intron:

GAGCATGACTGAXXXXXXXXXXXXXXXXXGGTGCACCTGGTGA CTTAGAC . .

- Exon phase = phase of previous intron
- Donor and acceptor models incorporated in intron models
- Length and sequence distribution determined empirically

GenScan States

- **N**: intergenic region
- **P**: promoter
- **F**: 5' untranslated region
- **Esngl**: single exon (intronless) (translation start -> stop codon)
- **Einit**: initial exon (translation start -> donor splice site)
- **Ek**: phase k internal exon (acceptor splice site -> donor splice site)
- **Eterm**: terminal exon (acceptor splice site -> stop codon)
- **I_k**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

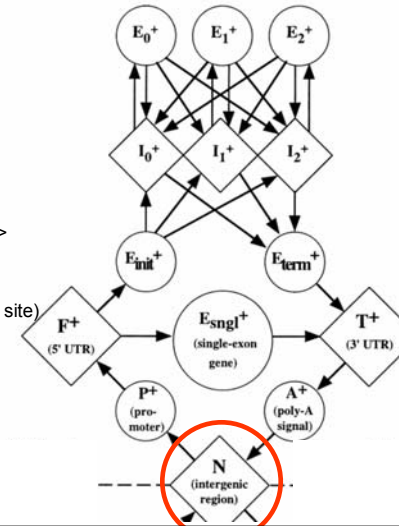


Fig. 3, Burge and Karlin 1997

Intergenic models

- Lengths are geometrically distributed with mean: $\frac{3 \times 10^9}{60,000}^*$
- Sequence model
5th order Markov model (*highest order trainable with data available*).
- Similar models used for untranslated regions

* Estimated human gene number in 1997

GenScan States

- **N**: intergenic region
- **P**: promoter
- **F**: 5' untranslated region
- **Esngl**: single exon (intronless) (translation start -> stop codon)
- **Einit**: initial exon (translation start -> donor splice site)
- **Ek**: phase k internal exon (acceptor splice site -> donor splice site)
- **Eterm**: terminal exon (acceptor splice site -> stop codon)
- **I_k**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

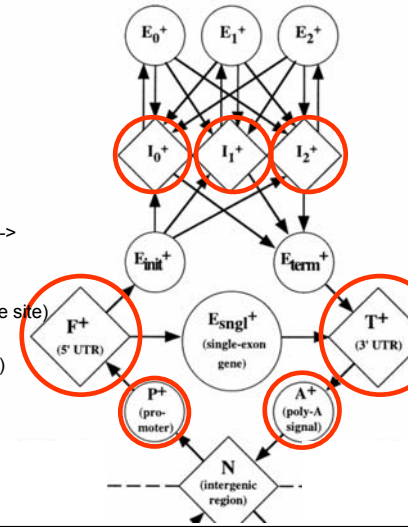
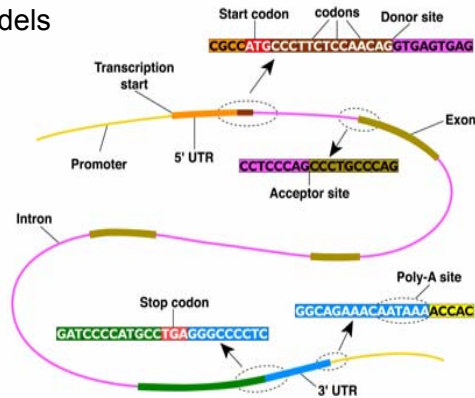


Fig. 3, Burge and Karlin 1997

Signal Models

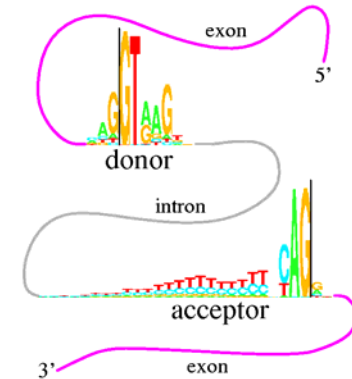
TATA box, polyA signal, 5'UTR

- Fixed length models
- PSSMs
- No positional dependence



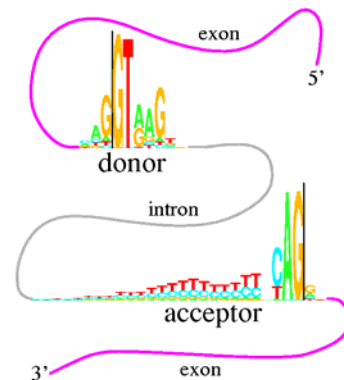
Acceptor Model

- Fixed length model
- “Weight array model”
 - Models dependence on the previous position



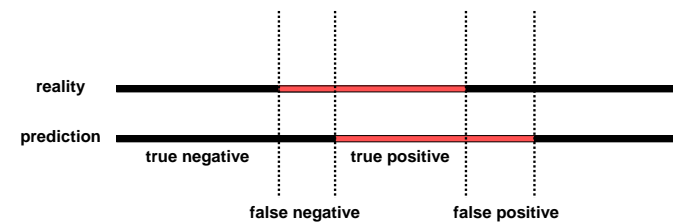
Donor Model

- Fixed length model
- “Maximal Dependence Decomposition”
 - Models dependencies between nonadjacent elements



Performance measures

Burset & Guigo, 1996

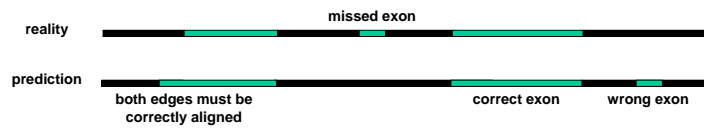


Nucleotide Level

$$S_n = \frac{TP}{TP + FN}$$

Performance measures

Burset & Guigo, 1996



Exon Level $S_n = \frac{\text{correct_exons}}{\text{actual_exons}}$ $S_p = \frac{\text{correct_exons}}{\text{predicted_exons}}$

Genscan Performance

	Nucl S_n	Exon S_n	Exon S_p	Missing Exons	Wrong Exons
Genscan	0.93	0.78	0.81	0.09	0.05
FGENEH	0.77	0.61	0.64	0.15	0.12
GENEID+	0.91	0.73	0.70	0.07	0.13
GENEPARSER3	0.86	0.56	0.58	0.14	0.09

Proportion of genes with all exons correctly predicted:

$$\frac{243}{570} = 0.43$$

