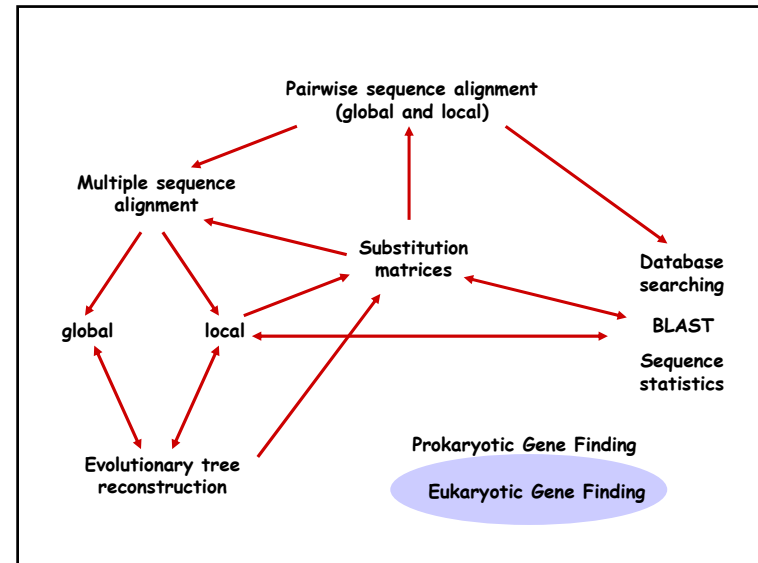


- Thurs, Nov 30: Online FCE's: Thru Dec 11  
Gene Finding 2
- Tues, Dec 5:  
PS5 due in my office at 5pm.  
Project presentation preparation – no class
- Thurs, Dec 7  
PS5 returned in class  
Final papers due  
Project presentations
- Monday Dec 18  
8:30am – 11:30am Final Exam, DH 1211



## Recent results in gene finding

Credits for slides:  
Serafim Batzoglou  
Marina Alexandersson  
Lior Pachter  
Sam Gross

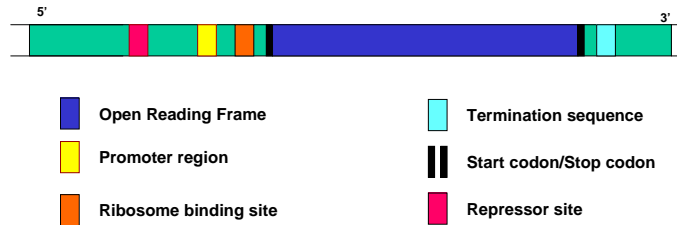
## Gene criteria

Snyder and Gerstein, Science 2003

- Open Reading Frames(ORFs) Computational
- Sequence features Computational
- Sequence conservation Computational
- Evidence for transcription Experimental
- Gene inactivation induces a phenotype Experimental

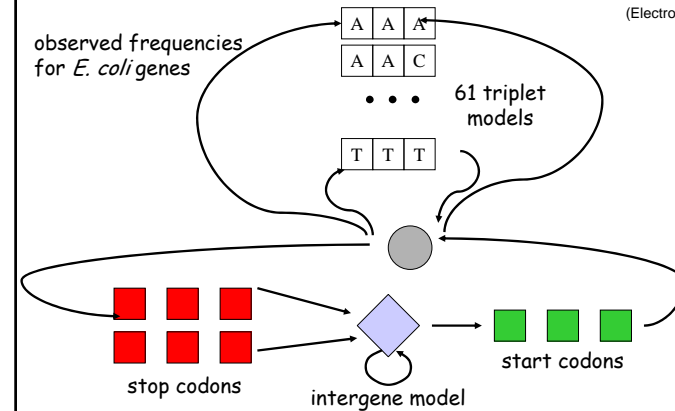
## Sequence features

- Coding statistics (e.g. codon bias)
- Gene structure



## An HMM that finds genes in *E. coli*

Krogh *et al*, 1995  
(Electronicreserves)



## 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
- Glimmer (Salzberg *et al*, 1998)
  - Finds 98% of all genes in a bacterial genome.

## Prokaryotic vs. Eukaryotic Genes

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

### Genomic data:

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

# 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 5<sup>th</sup> 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<sub>sngl</sub>**: single exon (intronless) (translation start -> stop codon)
- **E<sub>init</sub>**: initial exon (translation start -> donor splice site)
- **E<sub>k</sub>**: phase k internal exon (acceptor splice site -> donor splice site)
- **E<sub>term</sub>**: terminal exon (acceptor splice site -> stop codon)
- **I<sub>k</sub>**: phase k intron:
- **T**: 3' untranslated region
- **A**: poly-A signal

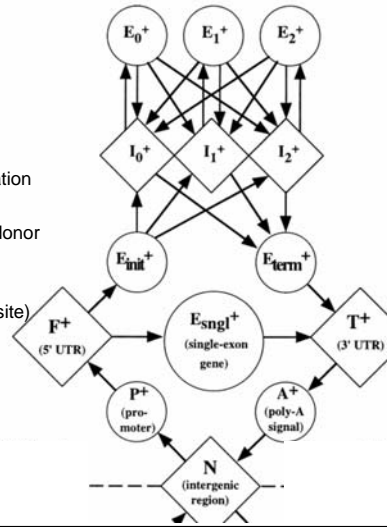
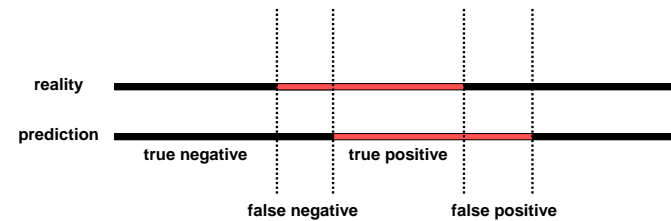


Fig. 3, Burge and Karlin 1997

# Performance measures

Burset & Guigo, 1996

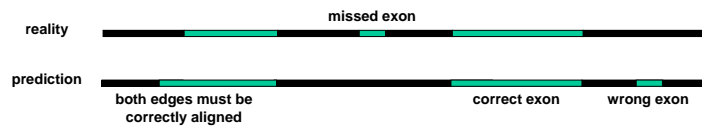


**Nucleotide Level**

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

## Performance measures

Burset & Guigo, 1996



**Exon Level**

$$S_n = \frac{\text{correct\_exons}}{\text{actual\_exons}} \quad S_p = \frac{\text{correct\_exons}}{\text{predicted\_exons}}$$

## Gene prediction performance

Burset & Guigo, 1996; Burge and Karlin, 1997

	Nucl S <sub>n</sub>	Exon S <sub>n</sub>	Exon S <sub>p</sub>	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

Genes with all exons correctly predicted by Genscan: 43%

Data set: Each sequence contains exactly one gene.

## Gene prediction performance with more challenging benchmarks

Guigo *et al*, 2000.

h178:

Single gene data

- 178 sequences
- 1 gene/sequence
- Nucleotides in gene regions: 53%
- Coding nucleotides: 21%

Gen178:

Semi-artificial genomes\*

- 42 sequences
- 4.1 genes/sequence
- Nucleotides in genic regions: 8.6%
- Coding nucleotides: 2.3%

\* Multiple genes interspersed with random sequence

## Genscan performance

Guigo *et al*, 2000.

	Nucl S <sub>n</sub>	Exon S <sub>n</sub>	Exon S <sub>p</sub>	Missing Exons	Wrong Exons
h178	0.93	0.78	0.75	0.08	0.10
Gen178	0.89	0.64	0.44	0.14	0.41

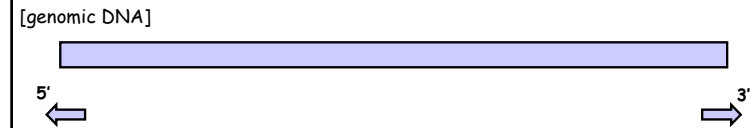
- Correct genes: 10% - 15%
- Gen178 does not contain repeats, pseudogenes, huge introns with huge introns, ... Results are probably still overly optimistic
- A lot of room for improvement...

## Innovations in gene prediction since 2000

- Spliced alignment with proteins or ESTs
  - Genewise, Procrustes
- Dual-genome predictors
  - SLAM, TWINSKAN, SGP2
- Multi-genome predictors
  - PhyloHMMs (Exoniphy), NSCAN
- Also, better models of gene features (e.g., splice sites, UTRs) and better identification of pseudogenes.

## Spliced alignments

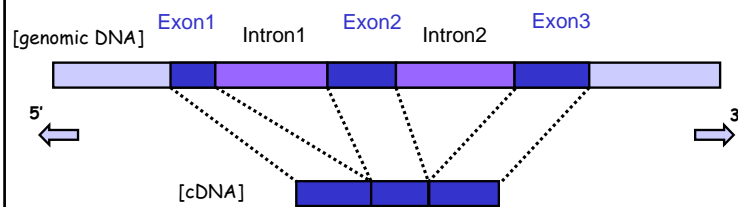
Procrustes (Gelfand et al,96), Genewise (Birney & Durbin, 97)



Align genomic DNA with proteins or cDNAs

## Spliced alignments

Procrustes (Gelfand et al,96), Genewise (Birney & Durbin, 97)



Align genomic DNA with proteins or cDNAs

## Spliced alignments

Procrustes (Gelfand et al,96), Genewise (Birney & Durbin, 97)

```

TTCATGAGGTGAGgtgaatagt.....cgtaattagGTCTTCTGGGGCC
|||||<--15907-->|||||
TTCATGAGGTGAG_____GTCTTCTGGGGCC
    
```

Methods include gene feature models, e.g., splice sites, frameshifts, penalise stop codons

Spliced alignment methods are

- More accurate for known genes
- Less accurate for unknown genes

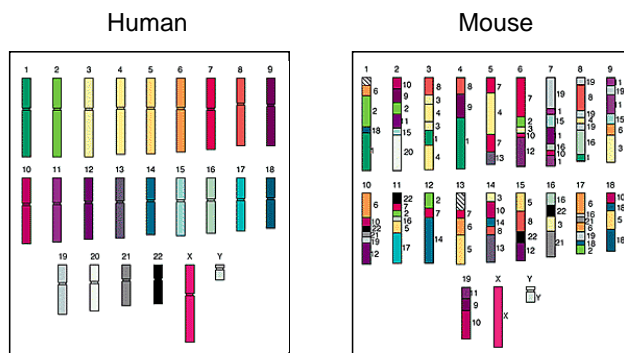
## Innovations in gene prediction since 2000

- Spliced alignment with proteins or ESTs
  - Genewise, Procrustes
- Dual-genome predictors
  - SLAM, TWINSKAN, SGP2
- Multi-genome predictors
  - PhyloHMMs (Exoniphy), NSCAN

## Dual genome predictors

- TWINSKAN (Brent), SGP2 (Guigo)
  - Predict genes in pairwise alignments
- SLAM (Pachter)
  - Simultaneous alignment and gene prediction

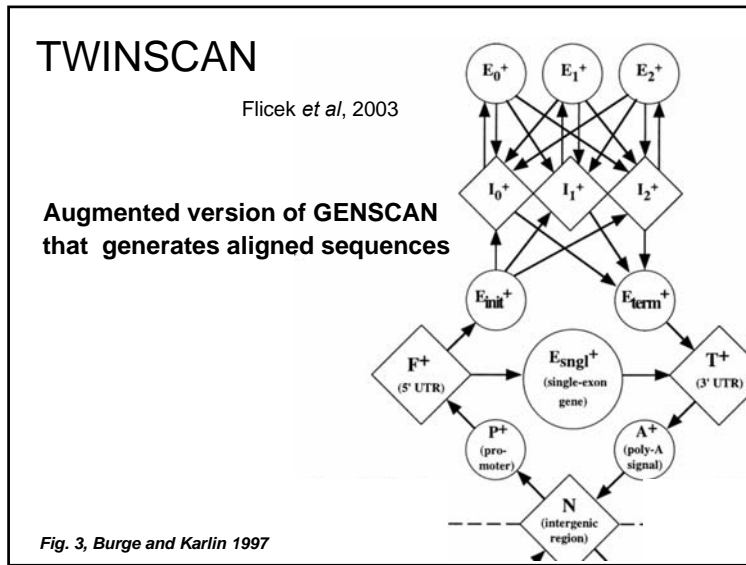
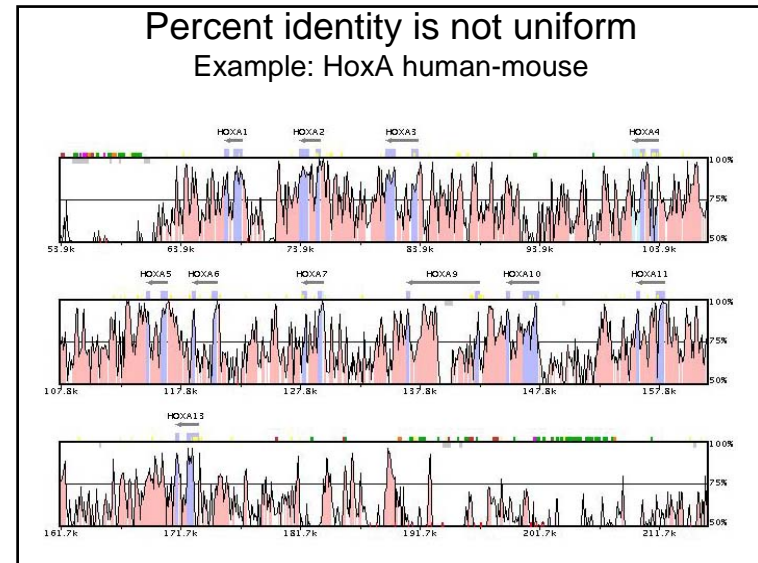
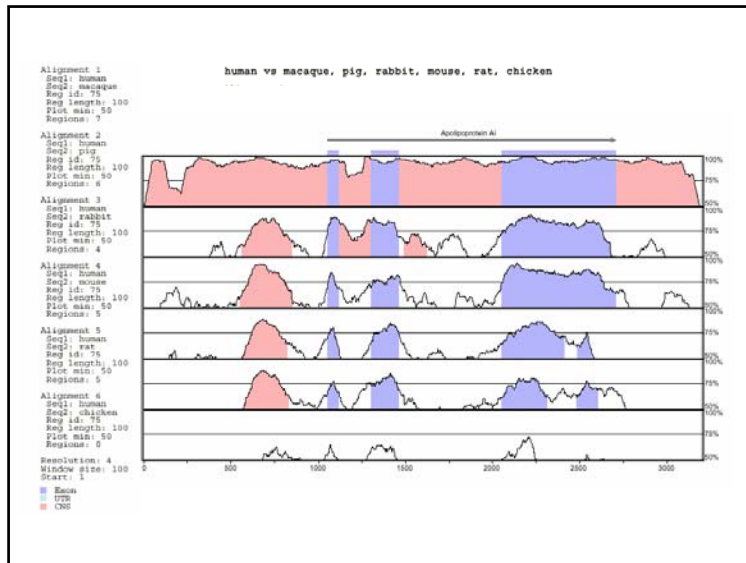
## Human-mouse homology



## Comparison of 1196 orthologous genes

Makalowski et al., 1996

- Sequence identity between genes in human/mouse
  - exons: 84.6%
  - introns: 35%
  - 5' UTRs: 67%
  - 3' UTRs: 69%



- ### TWINSKAN Algorithm
- Align the two sequences (eg. human and mouse)
 

Human: **ACGGCGACGUGCACGU**  
           | | : | : | | | | - | | | | : |  
 Mouse: **ACUGUGACG-GCACUU**
  - New "alphabet": 4 x 3 = 12 letters  
 $\Sigma = \{ A-, A:, A|, C-, C:, C|, G-, G:, G|, U-, U:, U| \}$
  - Mark each human base as gap (-), mismatch (:), match (|)  
**A| C| G: G| C: G| A| C| G| U| G| C| A| C| G: U|**

## TWINSKAN Algorithm

- Run Viterbi using emissions  $e_k(b)$   
where  $b \in \{ A-, A:, A|, \dots, T| \}$

**Note:**

Emission distributions  $e_k(b)$  estimated from real genes from human/mouse

$e_i(x|) < e_E(x|)$ : matches favored in exons

$e_i(x-) > e_E(x-)$ : gaps (and mismatches) favored in introns

## TWINSKAN Performance

- Slightly more sensitive than GENSCAN, much more specific
  - Exon sensitivity/specificity about 75%
- Much better at the gene level
  - Most genes are mostly right, about 25% exactly right

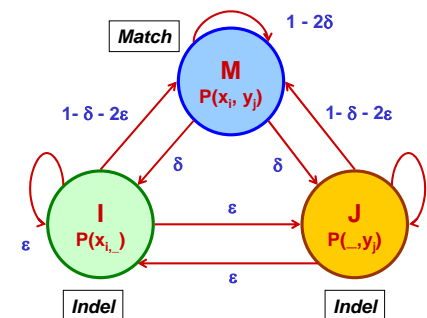
## SLAM

Alexanderson, Cawley, Pachter, 2003

- HMMs for simultaneous alignment and gene finding
- Generalized Pair HMMs
  - Generalized HMMs, aka Semi-HMMs
  - Each state has
    - a length distribution
    - a sequence generating model

## A Pair HMM for alignments

Given unaligned input sequences,  $x$  and  $y$





### Exon modules

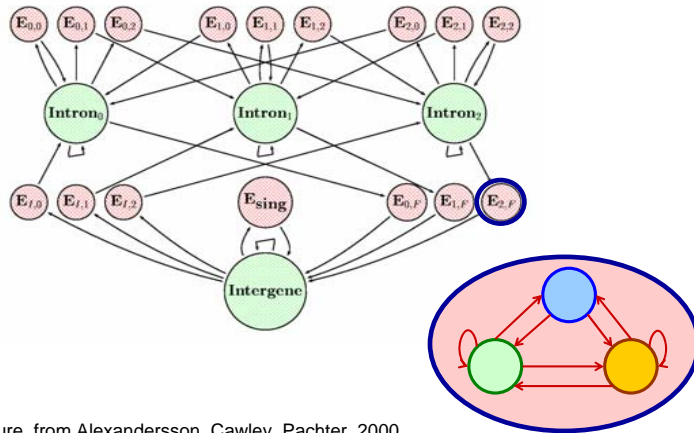
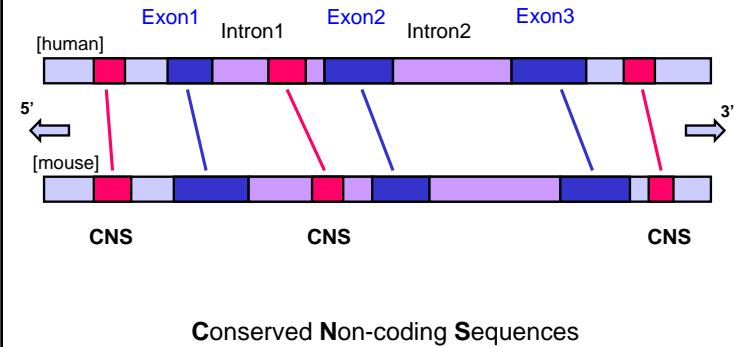


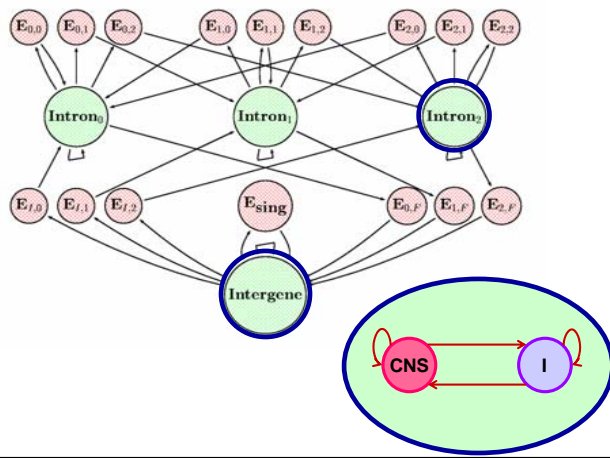
Figure from Alexandersson, Cawley, Pachter, 2000

### Intron modules



Conserved Non-coding Sequences

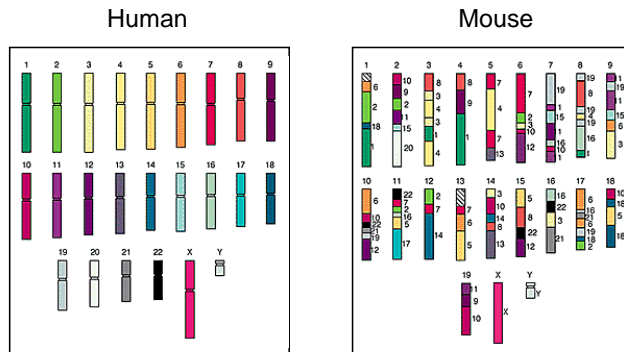
### Intron modules



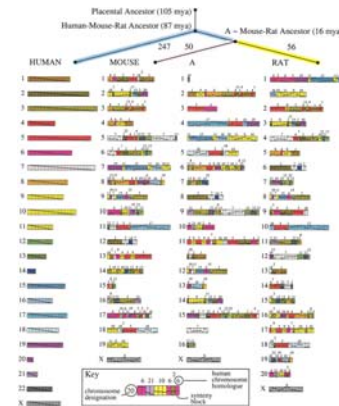
### Innovations in gene prediction since 2000

- Spliced alignment with proteins or ESTs
  - Genewise, Procrustes
- Dual-genome predictors
  - SLAM, TWINSCAN, SGP2
- Multi-genome predictors
  - PhyloHMMs (Exoniphy), NSCAN

## Two genomes can be hard to align



## Three genomes may provide more information



Guillaume Bourque et al. Genome Res. 2004; 14: 507-516

Cold Spring Harbor Laboratory Press

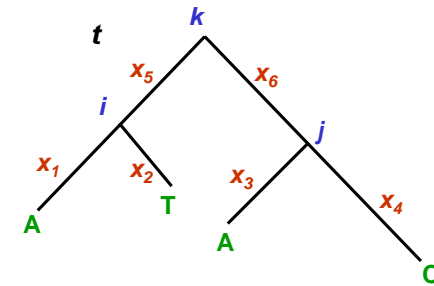


...AATTATCACGAT\_TAAGGT\_\_\_GACTGG...  
 ...ACCTTTCGAT\_TAAGGTAGCGATGTC...  
 ...AATTTGACGATATAAGGTACGAGCGT...  
 ...AATTTTCAGTTATAAGGTAGCTC\_GGT...

- Rates of mutation differ in introns, exons, UTRs, intergenic regions, etc.
- Idea:
  - Given an MSA, estimate the mutation rate at each site.
  - Classify the site based on the predicted rate.

## Probability of site given rate model, Q:

...TCAGG...  
 ...TGTCG...  
 ...TGACG...  
 ...TCCGA...



$$P(\{A, T, A, C\}^T | t, Q) =$$

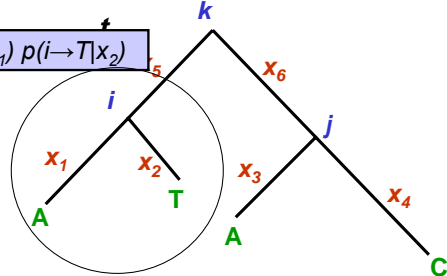
$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_5, Q) p(k \rightarrow j | x_6, Q)$$

$$\times p(i \rightarrow A | x_1, Q) p(i \rightarrow T | x_2, Q) p(j \rightarrow A | x_3, Q) p(j \rightarrow C | x_4, Q)$$

### Probability of site given rate model, Q:

Calculating  $p(i \rightarrow A | x_1)$   $p(i \rightarrow T | x_2)$

...TGTCG...  
 ...TGACG...  
 ...TCCGA...

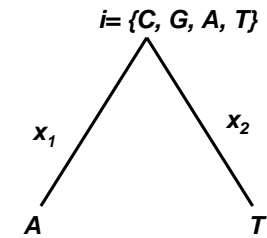


$$P(\{A, T, A, C\}^T | t, Q) =$$

$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_5, Q) p(k \rightarrow j | x_6, Q)$$

$$\times p(i \rightarrow A | x_1, Q) p(i \rightarrow T | x_2, Q) p(j \rightarrow A | x_3, Q) p(j \rightarrow C | x_4, Q)$$

Calculating  $p(i \rightarrow A | x_1)$   $p(i \rightarrow T | x_2)$



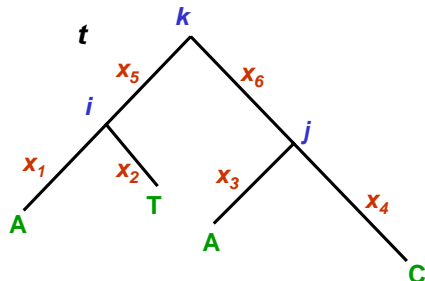
$$P(i=A)P(x_1)_{AA}P(x_2)_{AT} + P(i=T)P(x_1)_{TA}P(x_2)_{TT} \\ + P(i=C)P(x_1)_{CA}P(x_2)_{CT} + P(i=G)P(x_1)_{GA}P(x_2)_{GT}$$

Probabilities given by, e.g., Jukes Cantor model:

$$P(x_1)_{CC} = (1/4 + 3/4 e^{-4x_1}), P(x_1)_{CG} = (1/4 - 1/4 e^{-4x_1}), \text{ etc.}$$

### Probability of site given rate model, Q:

...TCAGG...  
 ...TGTCG...  
 ...TGACG...  
 ...TCCGA...



$$P(\{A, T, A, C\}^T | t, Q) =$$

$$\sum_{i \in \{A, C, G, T\}} \sum_{j \in \{A, C, G, T\}} \sum_{k \in \{A, C, G, T\}} p(k) p(k \rightarrow i | x_5, Q) p(k \rightarrow j | x_6, Q)$$

$$\times p(i \rightarrow A | x_1, Q) p(i \rightarrow T | x_2, Q) p(j \rightarrow A | x_3, Q) p(j \rightarrow C | x_4, Q)$$

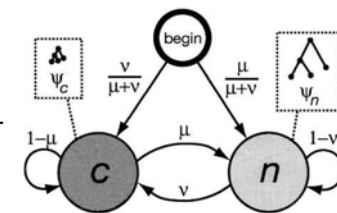
### A two-state phylo-HMM

Siepel and Haussler, 2004

Construct an HMM where each state emits a column vector,  $v$ , with  $P(v | t, Q)$

Exonphy is an exon predictor by Siepel and Haussler based on this idea.

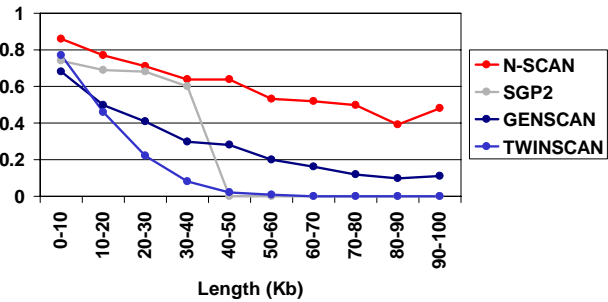
Doesn't use gene signals



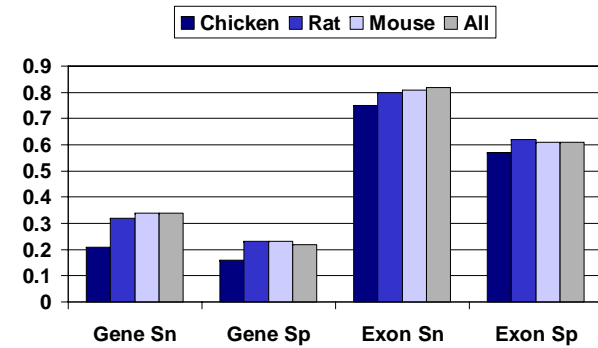
$x =$   
 TCGCGACATATACGA...  
 TTGGGGCATGTGGGT...  
 AGCAGACGTCGCAA...  $\gg$



## Intron Sensitivity By Length



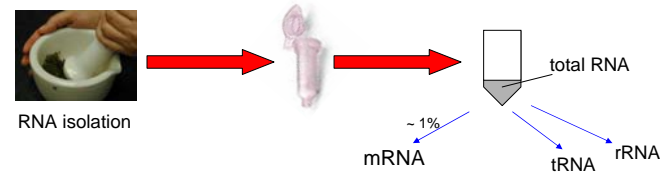
## Human Informant Effectiveness N-Scan



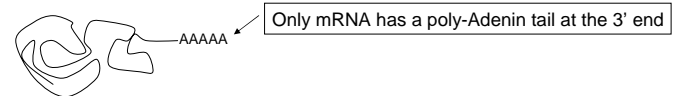
## Verifying gene predictions via RT-PCR

- Obtain mRNAs from cells
- Reverse transcribe to cDNA
- Create primers for predicted genes
- Amplify using PCR
- Sequence amplified PCR products
- Compare with predictions

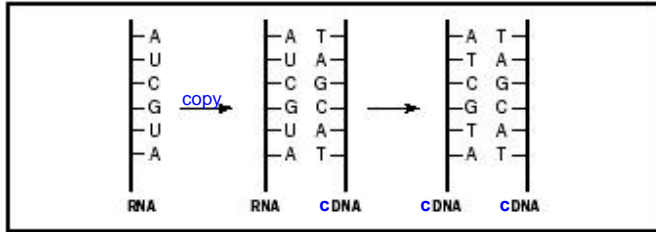
### Obtain mRNAs



- Most of the RNA is not of interest (tRNA, rRNA)

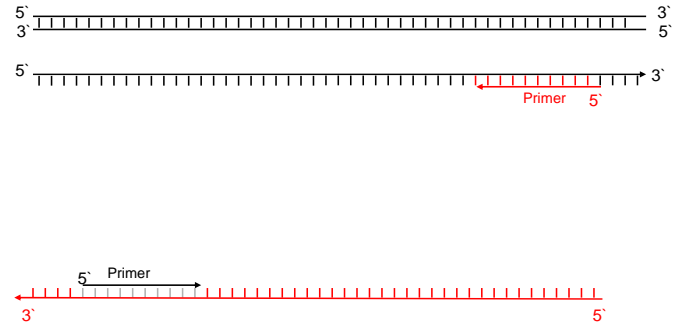


### Reverse Transcription



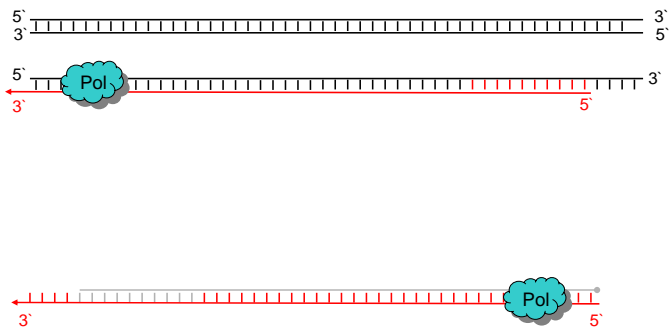
Reverse transcriptase uses a single-stranded RNA template to create a double-stranded DNA.

### Polymerase Chain Reaction

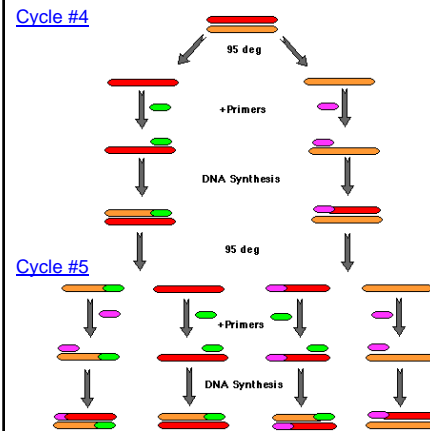


Primers designed from predicted genes

### Polymerase Chain Reaction



### Polymerase Chain Reaction (PCR)



## Verifying gene predictions via RT-PCR

- Obtain mRNAs from cells
- Reverse transcribe to cDNA
- Create primers for predicted genes
- Amplify using PCR
- Sequence amplified PCR products
- Compare with predictions

