

# Gene Recognition

Credits for slides:

- Marina Alexandersson
- Lior Pachter
- Serge Saxonov

# Reading

---

- GENSCAN
- SLAM
- Twinscan

Optional:  
Chris Burge's Thesis

Lecture 13, Thursday May 15, 2003

- GENSCAN
- SLAM
- Twinscan

Optional:  
Chris Burge's Thesis

# Gene expression

Diagram illustrating the process of gene expression:

**DNA** (represented by a double helix) undergoes **transcription** to produce **RNA** (represented by a single strand). The RNA sequence shown is: CCTGAGCCAACTATTGATGAA.

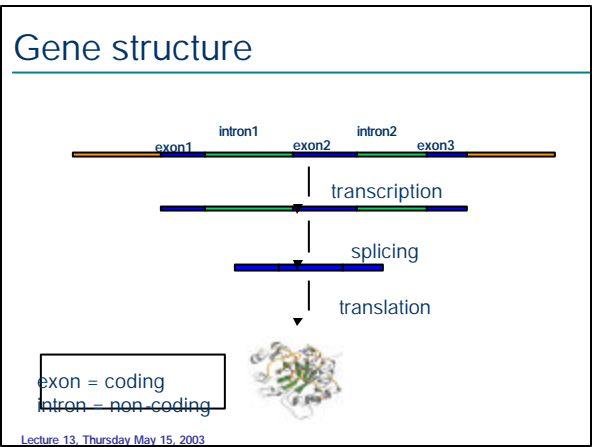
The RNA then undergoes **translation** to produce a **Protein** (represented by a folded chain). The protein sequence shown is: CCUGAGCCAAUUAUUGAUGAA.

The resulting protein is shown as a complex, folded structure.

**PEPTIDE**

Lecture 13, Thursday May 15, 2003

## Lecture 13, Thursday May 15, 2003



# Finding genes

The diagram illustrates a gene structure on a horizontal line. It consists of three orange segments labeled 'Exon 1', 'Exon 2', and 'Exon 3' from left to right. Between these exons are two green segments labeled 'Intron 1' and 'Intron 2'. Arrows at both ends of the line indicate the direction of transcription. Below the line, three vertical arrows point to specific locations: the first arrow points to the start of Exon 1, labeled 'Start codon ATG'; the second arrow points to the junction between Exon 1 and Intron 1; the third arrow points to the junction between Exon 2 and Intron 2; and the fourth arrow points to the end of Exon 3, labeled 'Stop codon TAG/TGA/TAA'. The text 'Splice sites' is positioned between the two splice junctions.

## Lecture 13, Thursday May 15, 2003

[illegible]

[illegible]

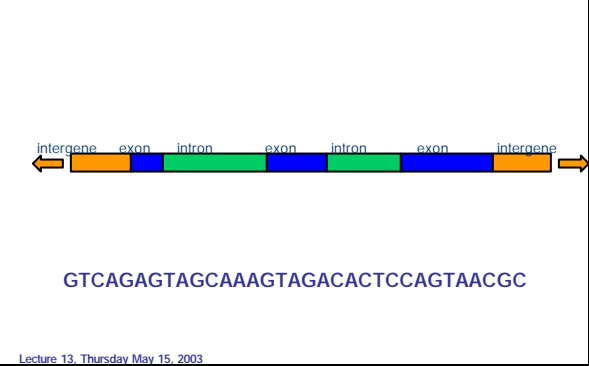
## Approaches to gene finding

- Homology
  - BLAST, Procrustes.
- Ab initio
  - Genscan, Genie, GeneID.
- Hybrids
  - GenomeScan, GenieEST, Twinscan, SGP, ROSETTA, CEM, TBLASTX, SLAM.

---

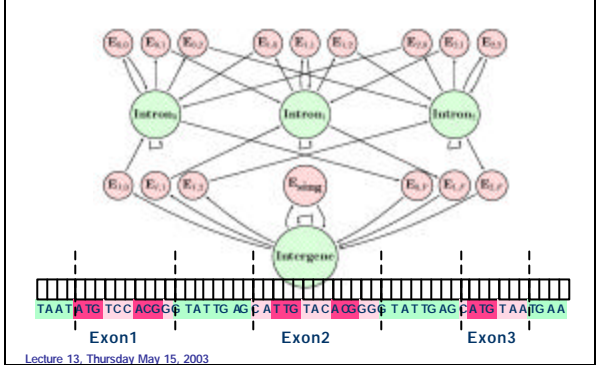
HMMs for single species gene finding:  
Generalized HMMs

## HMMs for gene finding



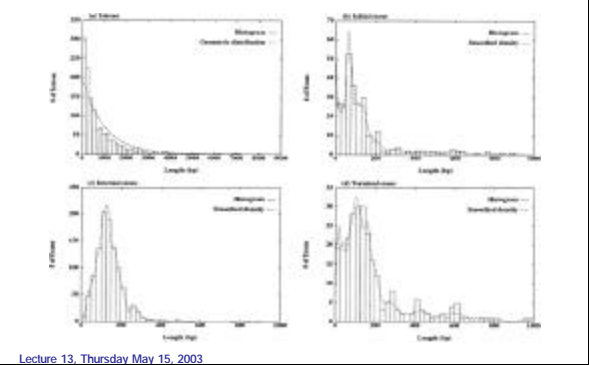
Lecture 13, Thursday May 15, 2003

## GHMM for gene finding



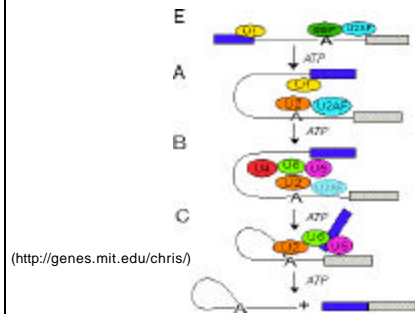
---

### Observed duration times



---

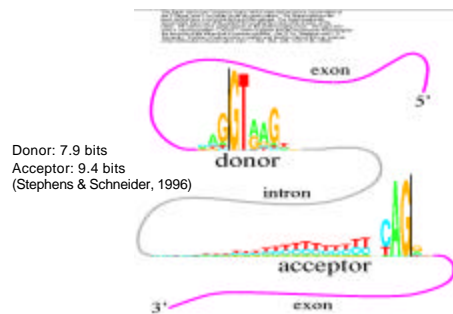
## Biology of Splicing



(<http://genes.mit.edu/chris/>)

Lecture 13, Thursday May 15, 2003

## Consensus splice sites



Lecture 13, Thursday May 15, 2003

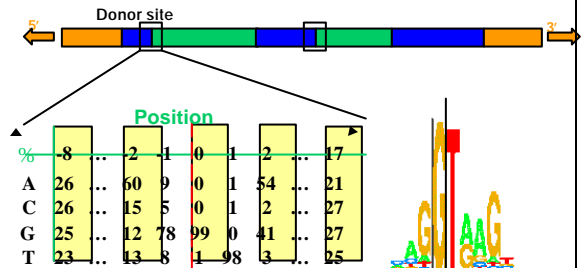
(<http://www-lmmb.ncicrf.gov/~toms/sequencelogo.html>)

## Splice Site Models

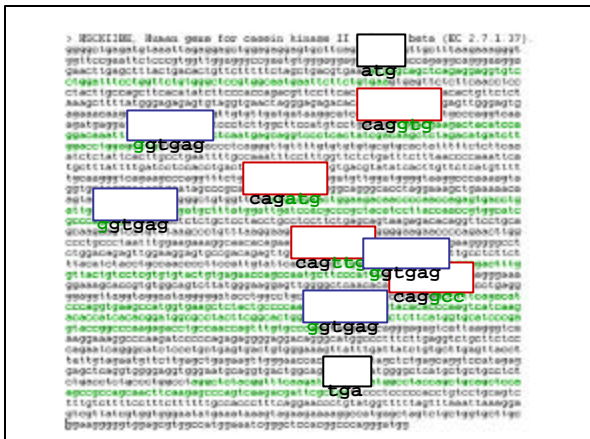
- WMM: weight matrix model = PSSM (Staden 1984)
- WAM: weight array model = 1<sup>st</sup> order Markov (Zhang & Marr 1993)
- MDD: maximal dependence decomposition (Burge & Karlin 1997)  
decision-tree like algorithm to take significant pairwise dependencies into account

Lecture 13, Thursday May 15, 2003

## Splice site detection



Lecture 13, Thursday May 15, 2003



## Coding potential

Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

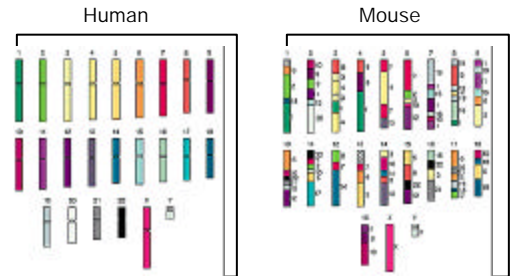
Lecture 13, Thursday May 15, 2003

## Comparison of 1196 orthologous genes (Makalowski et al., 1996)

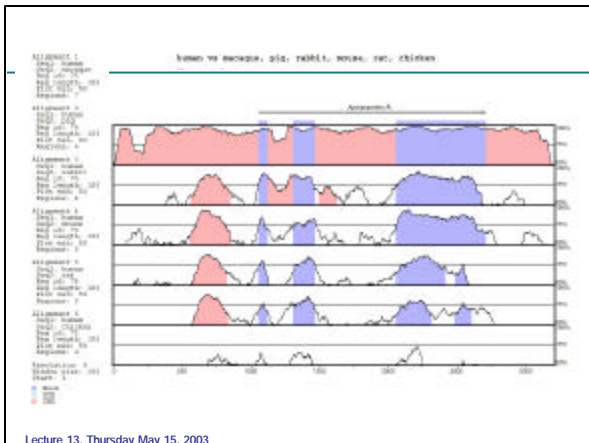
- Sequence identity:
  - exons: 84.6%
  - protein: 85.4%
  - introns: 35%
  - 5' UTRs: 67%
  - 3' UTRs: 69%
- 27 proteins were 100% identical.

Lecture 13, Thursday May 15, 2003

## Human-mouse homology

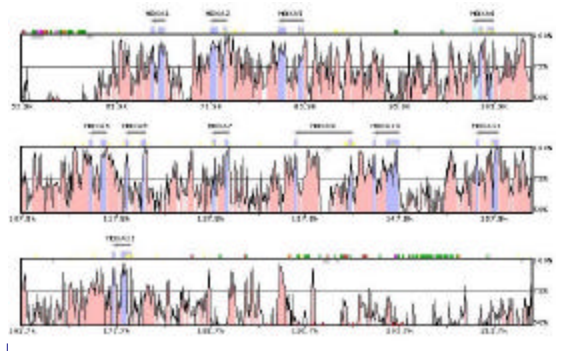


Lecture 13, Thursday May 15, 2003



Lecture 13, Thursday May 15, 2003

## Not always: HoxA human-mouse



## Alignment

```

50      .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
247 GGTGAGGTCGAGGACCTGCA  CGGAGCTGTATGGAGGCA  AGAGC
    |:  ||  ||||:  ||||  --:|  ||  ||||  ||||  ||||  ||||
368 GAGTCGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCAITGAGAGG

100      .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
292 TTC      CTACAGAAAAGTCCAGCAAGGAGCCACACTTCACTG
    |||-----|||  |||  |||  |||  |||  |||  |||  |||  |||
418 TTCTGGCTAAGCTCTCCCTTAGGACTGAGCAGAGGGCT  CAGGTGCGG

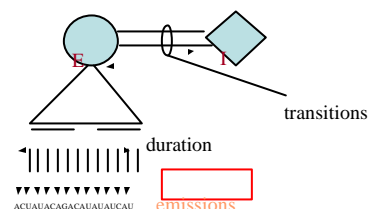
150      .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
332      ATGTGAGGGGAAGACATCATTTGGGATGTCAGTG
    -----|||  |||  |||  |||  |||  |||  |||  |||  |||
467 TGGGAGATGAGGCCAATGTGAGGGGAAGACATCATTTGGGATGTCAGTG

200      .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
367 TTCAACCTCAGCAATGCCATCATGGGCAGCGGCATCTGGGACTGCGCTA
    ||||:|||||||:|||||||:|||||||:|||||||:|||||||:|||||||:|||||||
517 TTCAATCTCAGCAAGCCATCATGGGCAGTGAATTCTGGGGCTGCGCTA
    
```

Lecture 13, Thursday May 15, 2003

## Twinscan

- Twinscan is an augmented version of the Gencscan HMM.



Lecture 13, Thursday May 15, 2003

## Twinscan Algorithm

1. Align the two sequences (eg. from human and mouse)
2. Mark each human base as gap ( - ), mismatch ( : ), match ( | )

New "alphabet":  $4 \times 3 = 12$  letters

$\Sigma = \{ A-, A:, A|, C-, C:, C|, G-, G:, G|, U-, U:, U| \}$

Lecture 13, Thursday May 15, 2003

## Twinscan Algorithm (cont'd)

3. 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

Lecture 13, Thursday May 15, 2003

## Example

Human: **ACGGCGACUGUGCACGU**

Mouse: **ACUGUGAC GUGCACUU**

Align: **| |: |: || - | | | | | : |**

Input to Twinscan HMM:

**A| C| G: G| C: G| A| C| U- G| U| G| C| A| C| G: U|**

Recall,  $e_E(A|) > e_I(A|)$

$e_E(A-) < e_I(A-)$

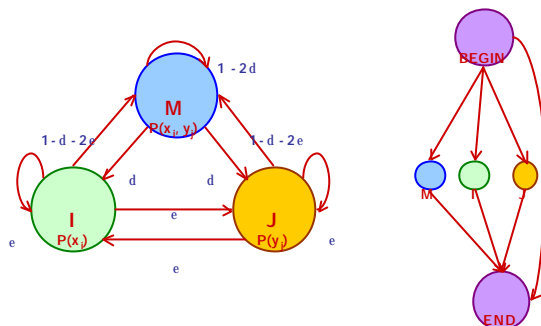
Likely exon

Lecture 13, Thursday May 15, 2003

HMMs for simultaneous alignment  
and gene finding:

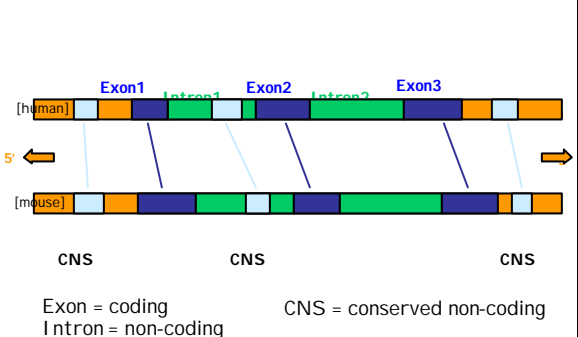
Generalized Pair HMMs

## A Pair HMM for alignments



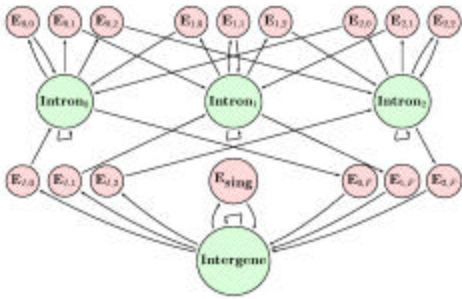
Lecture 13, Thursday May 15, 2003

## Cross-species gene finding



Lecture 13, Thursday May 15, 2003

## Generalized Pair HMMs



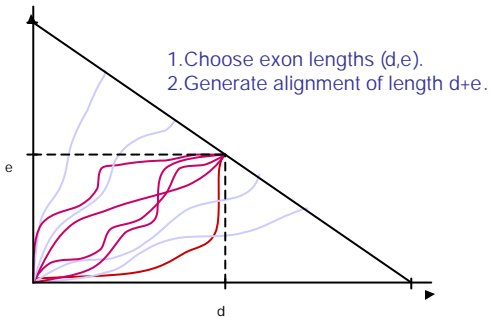
Lecture 13, Thursday May 15, 2003

## Ingredients in exon scores

- Splice site detection (VLMM)
- Length distribution (generalized)
- Coding potential (codon freq. tables)
- GC-stratification

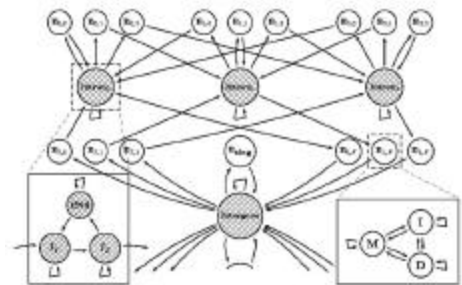
Lecture 13, Thursday May 15, 2003

## Exon GPHMM



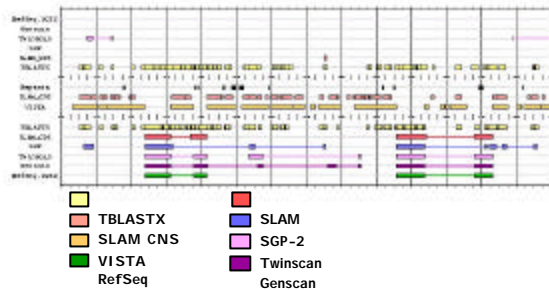
Lecture 13, Thursday May 15, 2003

## The SLAM hidden Markov model



Lecture 13, Thursday May 15, 2003

## Example: HoxA2 and HoxA3



Lecture 13, Thursday May 15, 2003

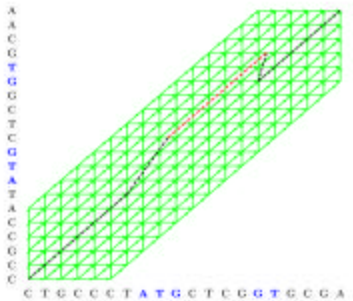
## Computational complexity

Model	Time	Space
HMM	$N^2T$	$NT$
PHMM	$N^2TU$	$NTU$
GHMM	$D^2N^2T$	$NT$
GPHMM	$D^4N^2TU$	$NTU$

$N$  = no. states       $T$  = length seq1  
 $D$  = max duration     $U$  = length seq2

Lecture 13, Thursday May 15, 2003

## Approximate alignment



Reduces  
TU -factor  
to  
 $hT$

Lecture 13, Thursday May 15, 2003

## Measuring Performance

- Definition:

- Sensitivity (SN):

(# correctly predicted)/(# true)

- Specificity (SP):

(# correctly predicted)/(# total predicted)

Lecture 13, Thursday May 15, 2003

## Measuring Performance

Test set	Nucleotide level			Exon level			
	SN	SP	AC	SN	SP	(SN+SP)/2	ME
<b>The ARSETTA sets</b>							
ARSETTA	0.935	0.978	0.949	0.833	0.829	0.831	0.047
SP=1	0.940	0.960	0.940	0.708	0.700	0.700	0.040
SLAM	0.951	0.981	0.960	0.753	0.755	0.750	0.057
TVINSCAN-p	0.900	0.941	0.940	0.855	0.824	0.840	0.061
TVINSCAN	0.964	0.989	0.981	0.889	0.789	0.800	0.118
GENSCAN	0.975	0.988	0.982	0.887	0.770	0.790	0.102
<b>Brca</b>							
SLAM	0.852	0.896	0.864	0.727	0.533	0.630	0.000
TVINSCAN-p	0.976	0.980	0.986	0.779	0.521	0.652	0.000
TVINSCAN	0.945	0.931	0.934	0.691	0.573	0.632	0.000
SP=2	0.940	0.936	0.939	0.609	0.573	0.591	0.001
GENSCAN	0.932	0.980	0.956	0.745	0.205	0.490	0.000
<b>ExonIn</b>							
SLAM	0.876	0.981	0.906	0.802	0.809	0.831	0.121
TVINSCAN-p	0.942	0.980	0.945	0.809	0.809	0.809	0.096
TVINSCAN	0.933	0.977	0.903	0.835	0.826	0.831	0.110
SP=2	0.755	0.998	0.853	0.533	0.900	0.719	0.152
GENSCAN	0.947	0.980	0.962	0.869	0.751	0.789	0.121

Lecture 13, Thursday May 15, 2003