Bioinformatics 1 -- lecture 22

Gene finding in eukaryotes

- intron/exon boundaries
- splicing
- alternative splicing

Gene Prediction: Computational Challenge

- Gene: A sequence of nucleotides coding for protein
- Gene Prediction Problem: Determine the beginning and end positions of genes in a genome

Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatg
ctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatg
ctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatg
caagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgat
gacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgct
aagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgc
ggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctc
atgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggcta
tgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcgg
ctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagc
tgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggct
atgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
gg

Gene Prediction: Computational Challenge

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatg
ctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatg
ctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatg
caagctgggatcctgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgat
gacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgct
aagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgc
ggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctc
atgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggcta
tgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcgg
ctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagc
tgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggct
atgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgc
gg

Gene!

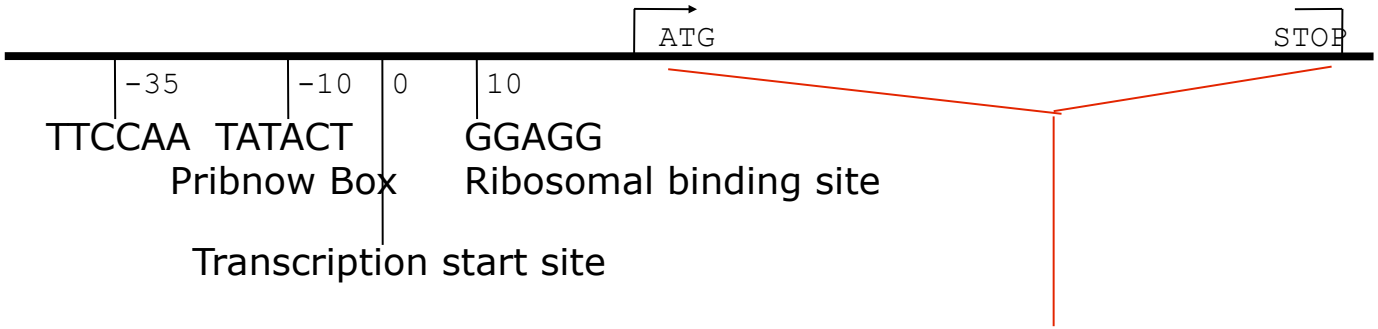
Two Approaches to Gene Prediction

- **Statistical**: coding segments (ORFs or exons) have typical sequences on either end and use different subwords than non-coding segments (introns or intergenic regions).

- **Similarity-based**: many human genes are similar to genes in mice, chicken, or even bacteria. Therefore, already known mouse, chicken, and bacterial genes may help to find human genes.

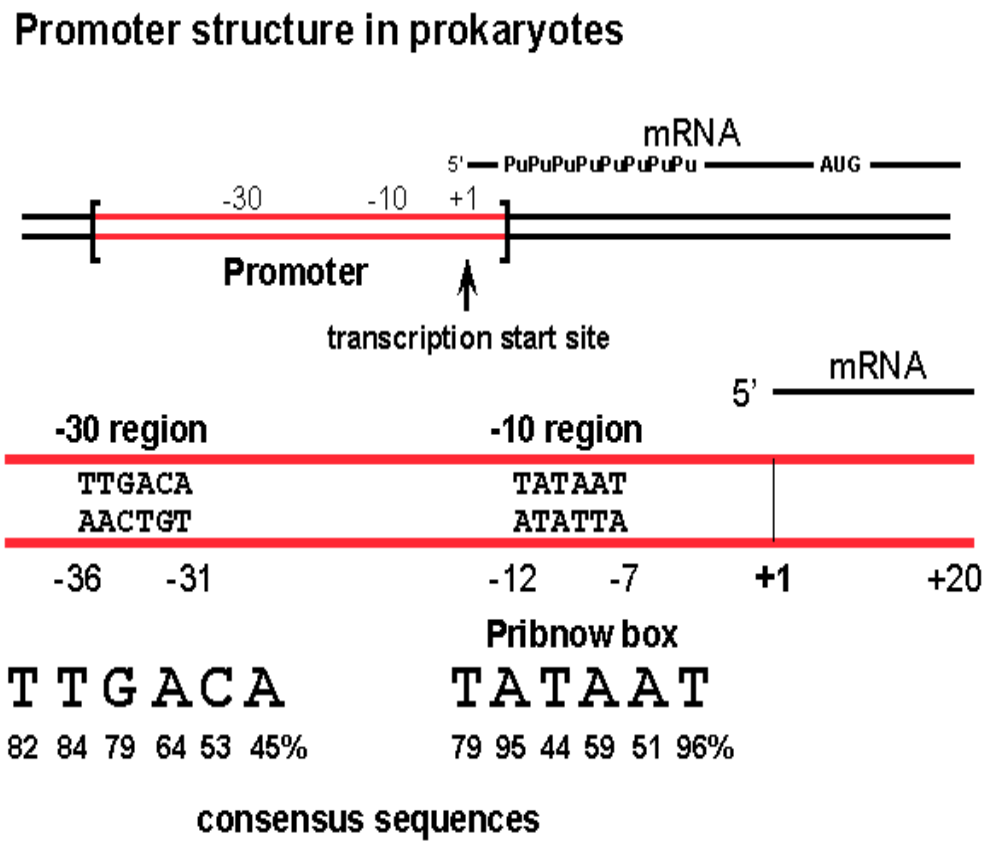
Gene Prediction and Motifs

- Upstream regions of genes often contain motifs that can be used for gene prediction



- Reading frame should be long enough.

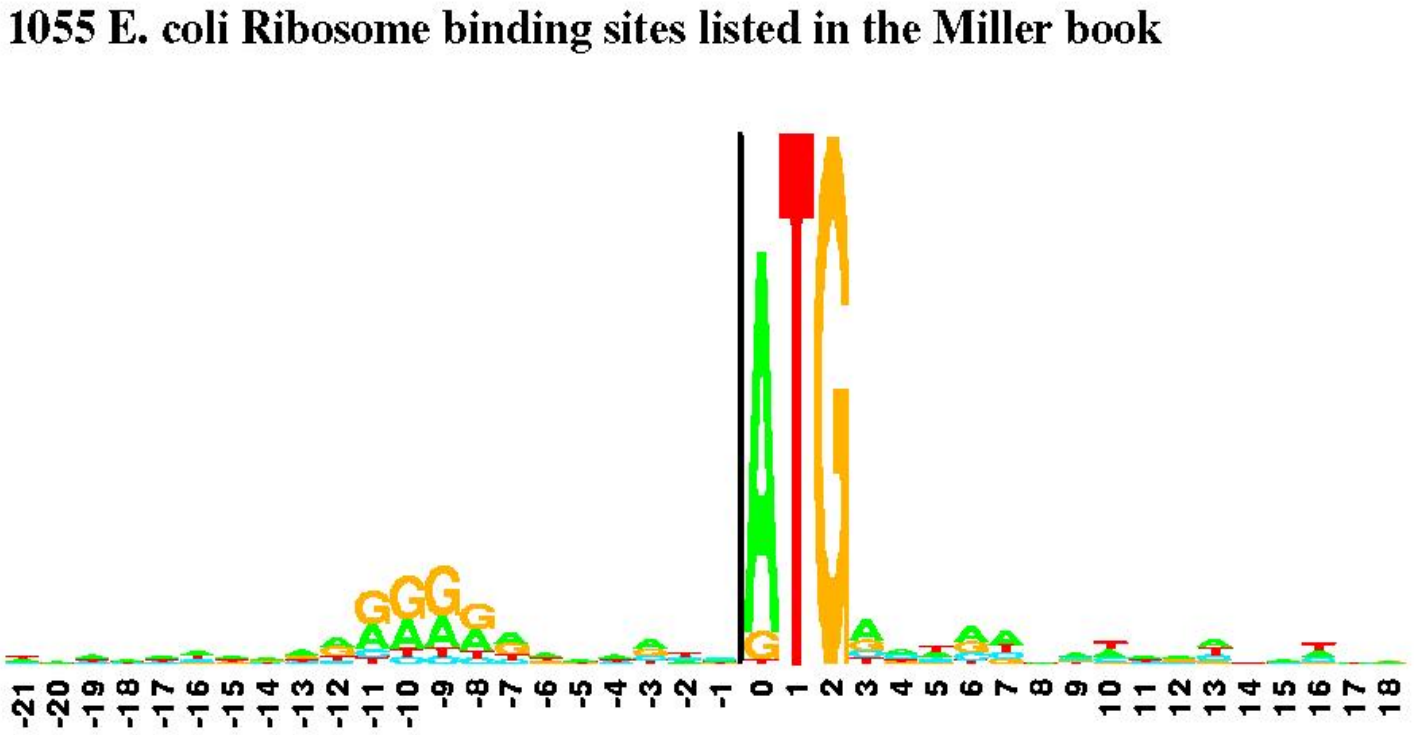
Promoter Structure in Prokaryotes (E.Coli)



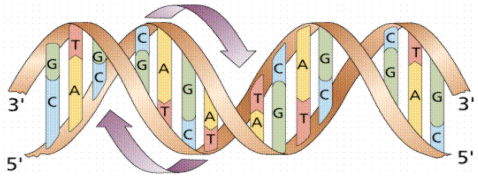
Transcription starts at offset 0.

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)

Ribosomal Binding Site



Central Dogma of biology



DNA

transcription

RNA

translation

Protein

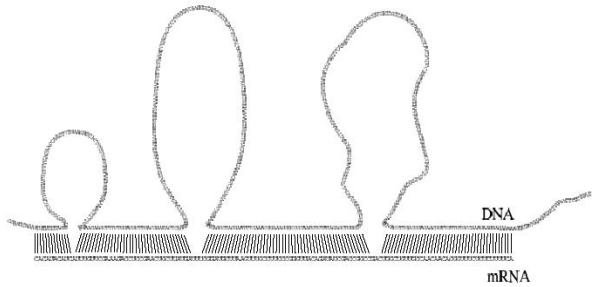
CCTGAGCCAACTATTGATGAA

CCUGAGCCAACUAUUGAUGAA

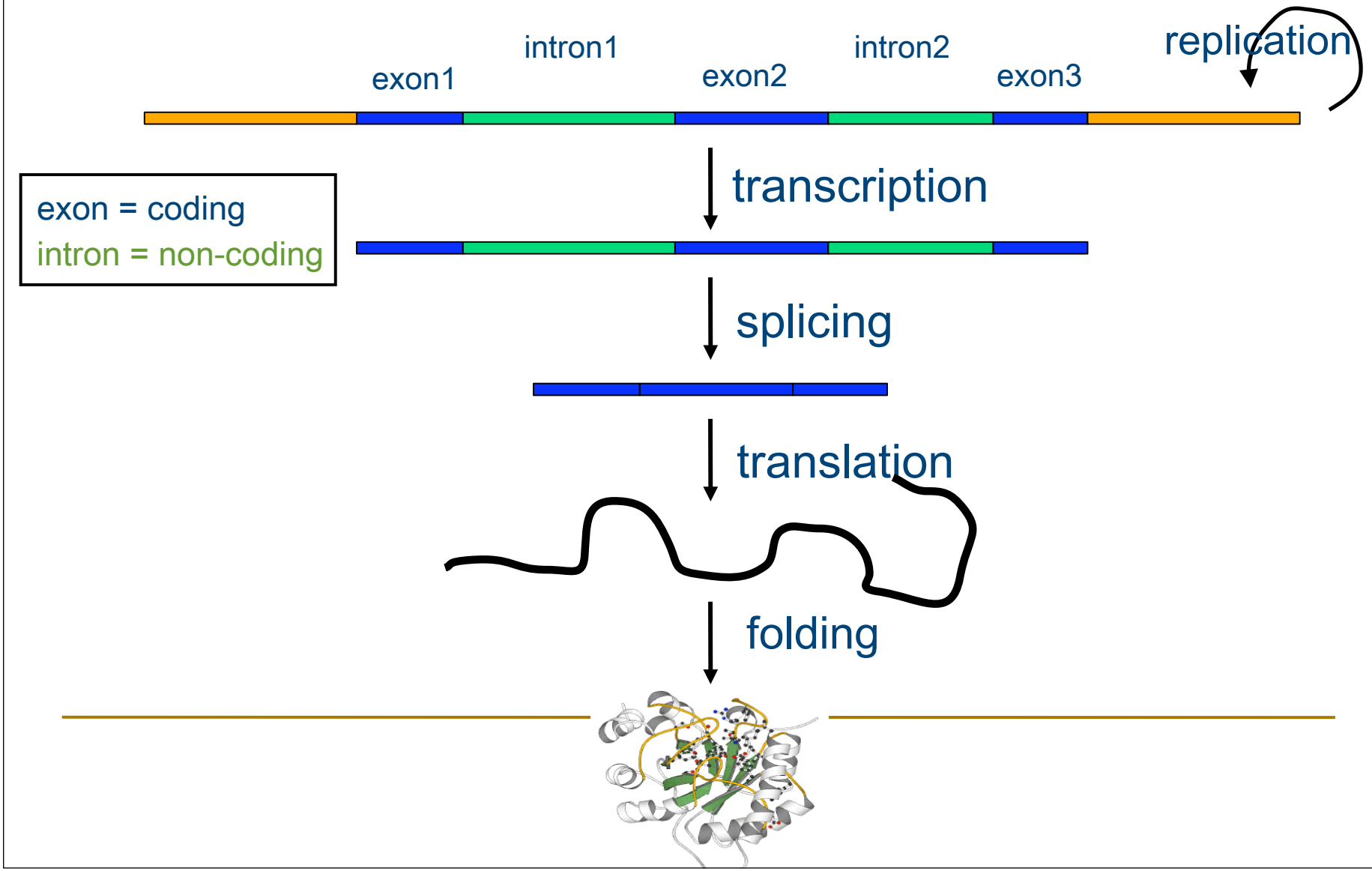
PEPTIDE

Discovery of Split Genes

- In 1977, Phillip Sharp and Richard Roberts experimented with mRNA of *hexon*, a viral protein.
  - Map hexon mRNA in viral genome by hybridization to adenovirus DNA and electron microscopy
  - mRNA-DNA hybrids formed three curious loop structures instead of contiguous duplex segments

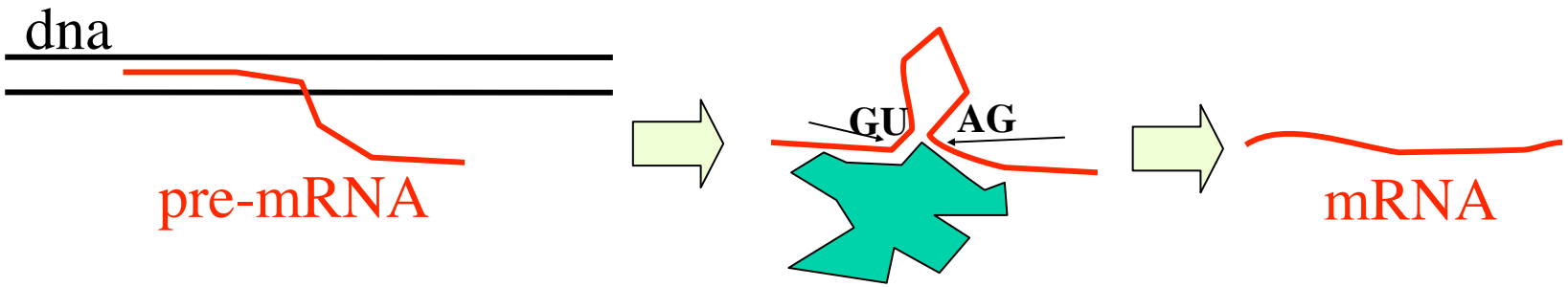


Expanded Central Dogma

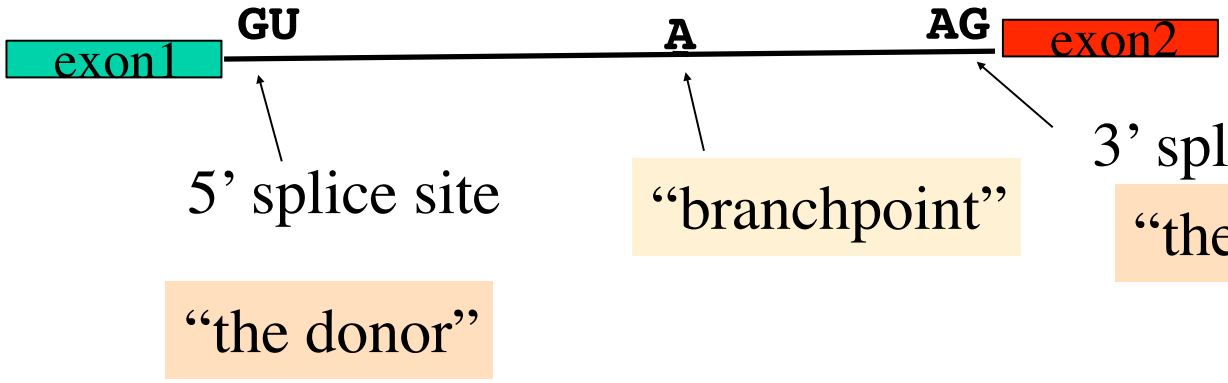


Finding genes in eukaryotes is harder.

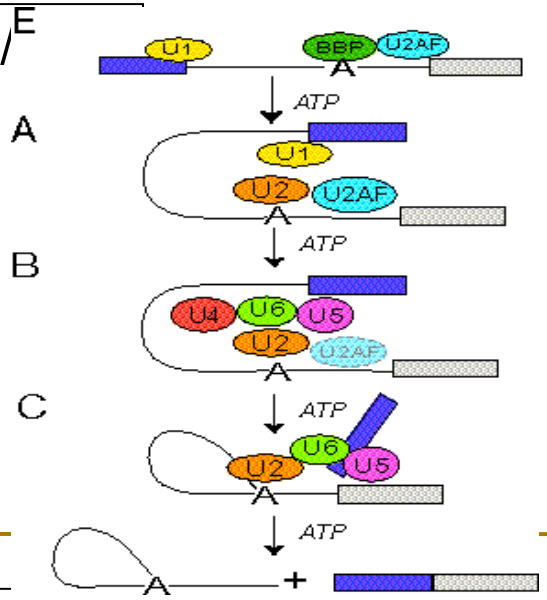
- Genes are composed of coding regions (exons) and internal non-coding regions (introns).
- Genes are transcribed to pre-mRNA.
- Introns are removed from pre-mRNA by the *spliceosome* (a ribozyme)
- Proteins are translated from the mRNA after splicing.
- Different tissues may splice pre-mRNA differently!



Splicing mechanism



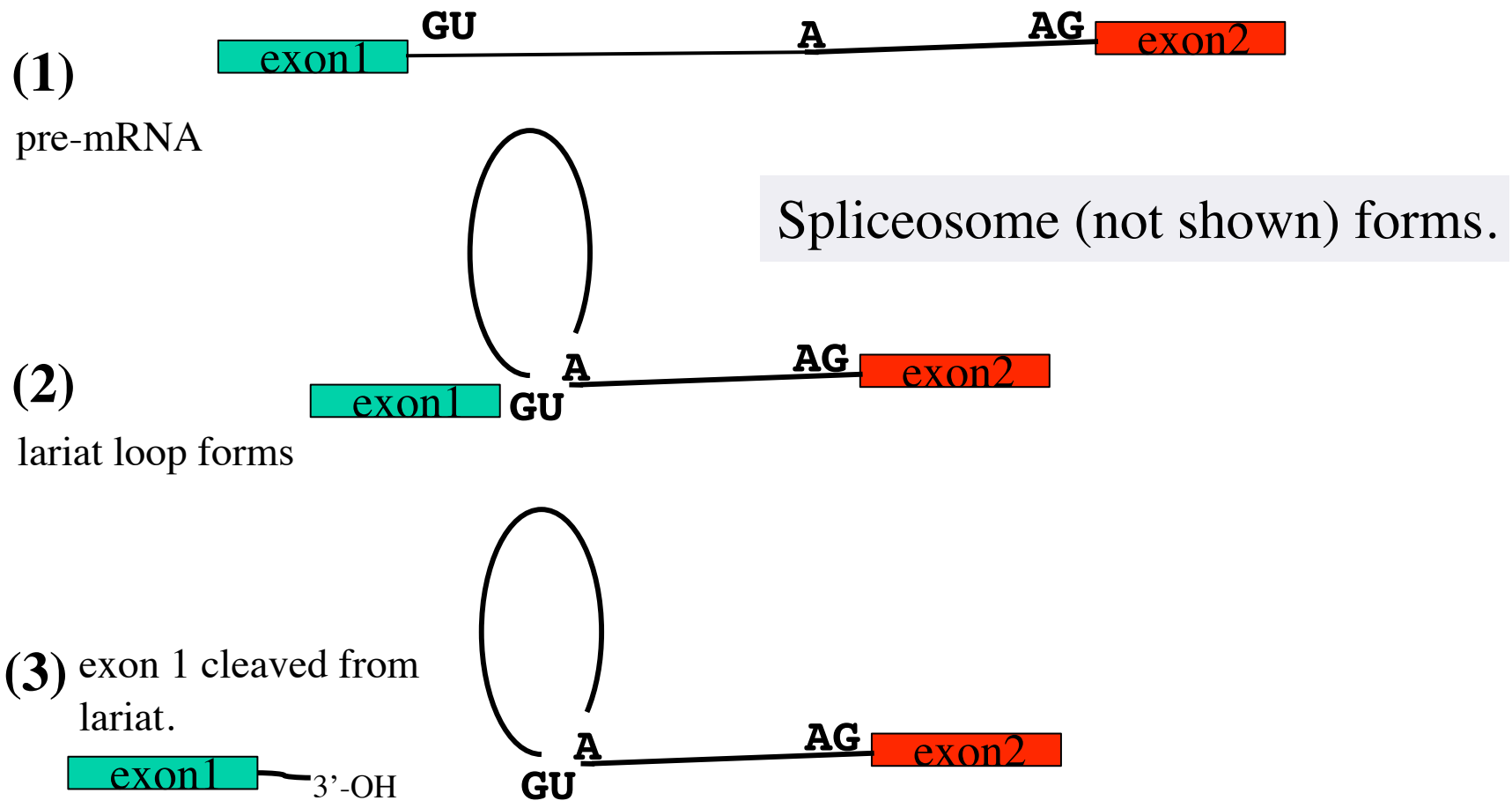
- snRNPs bind donor, acceptor, branchpoint sites/
- The *spliceosome* forms
- Spliceosome excises introns in the mRNA



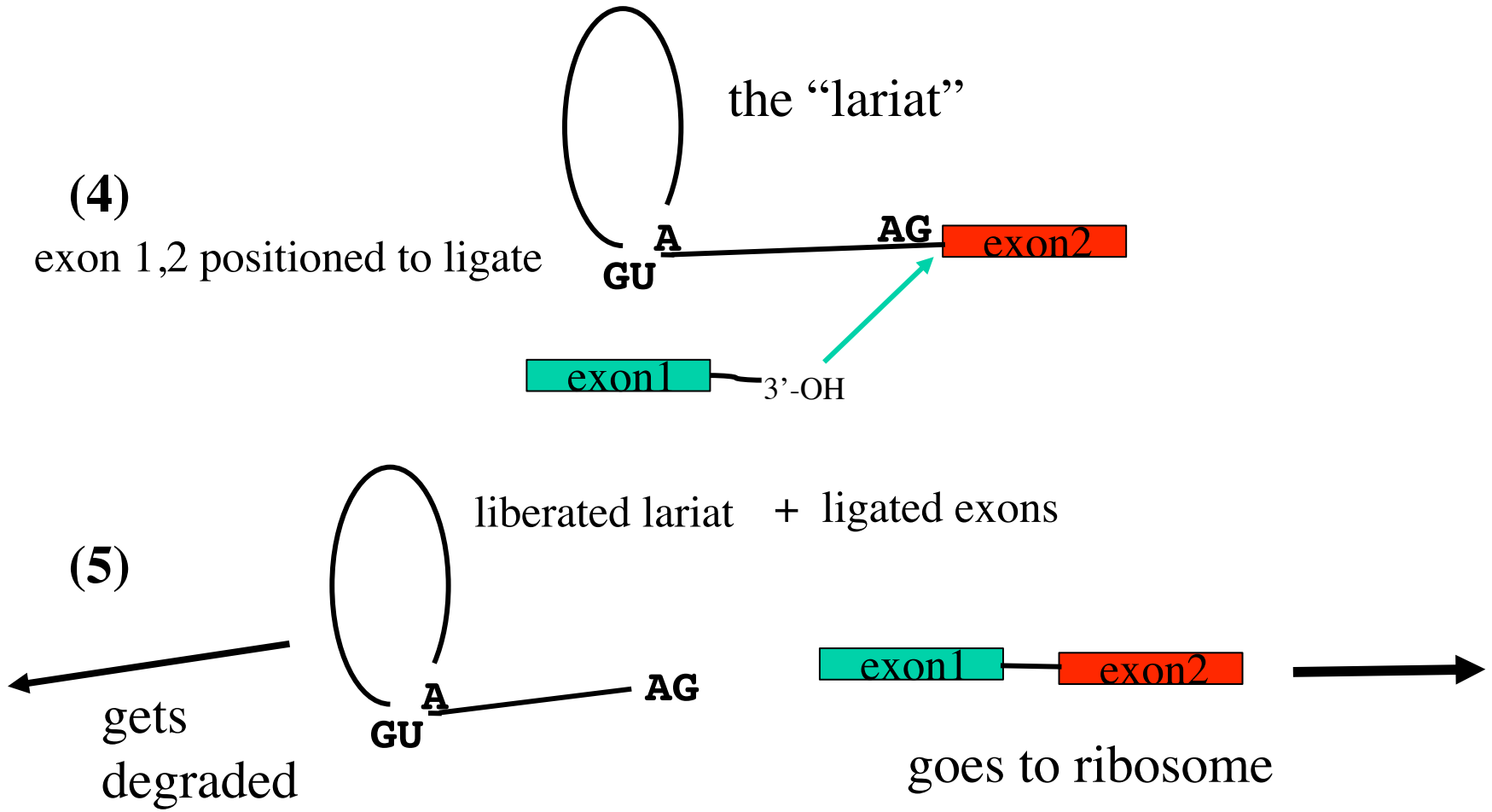
Spliceosome: Def: A ribonucleoprotein complex, containing RNA and small nuclear ribonucleoproteins (snRNPs) that is assembled during the splicing of messenger RNA primary transcript to excise an intron.

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

Splicing mechanism



Splicing mechanism



Spliceosome (not shown) disassociates.

Splicing mechanism on the web

http://neuromuscular.wustl.edu/pathol/diagrams/splicefunct.html

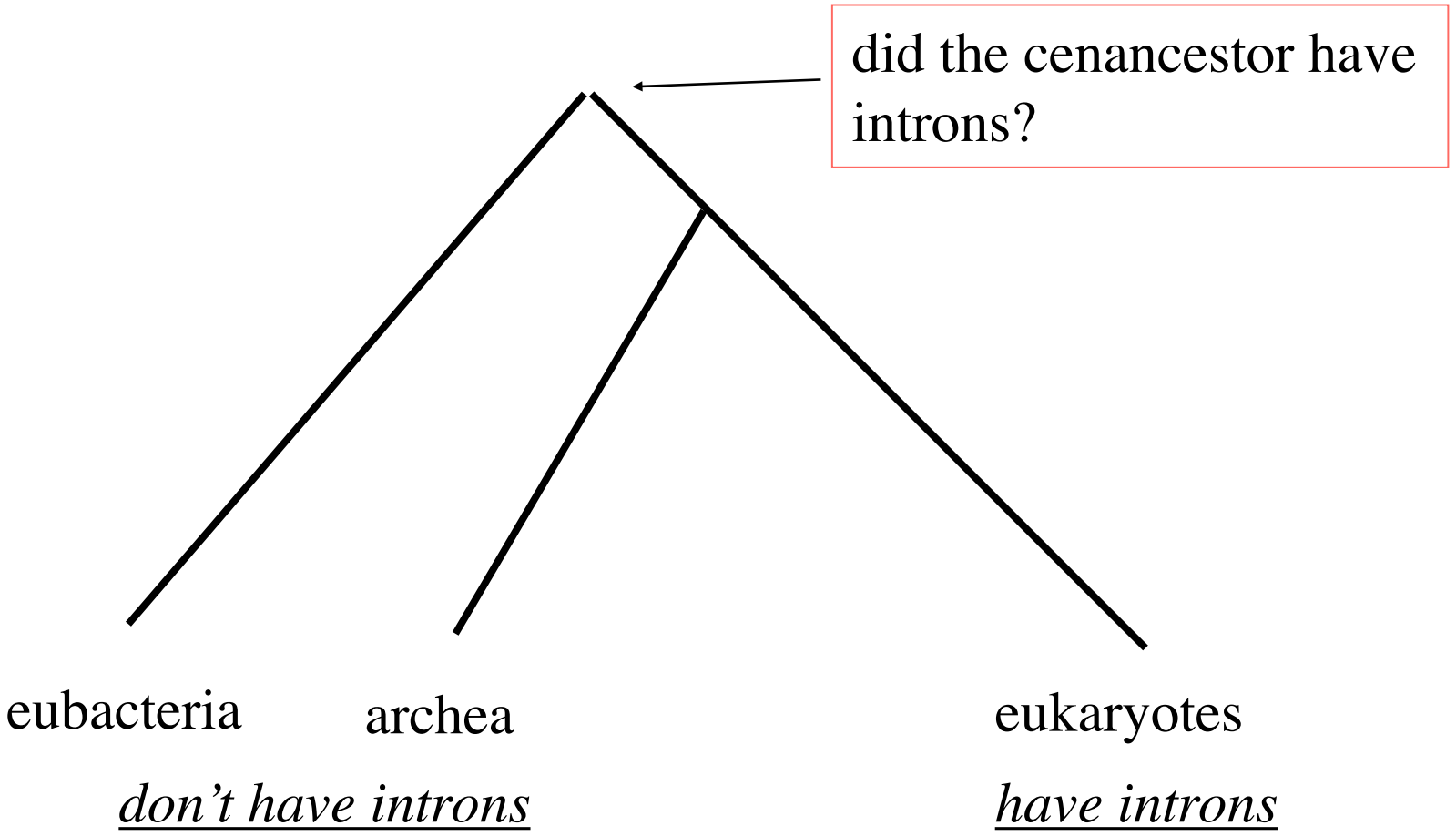
google: wustl neuromuscular splicefunct

http://neuromuscular.wustl.edu/pathol/diagrams/splicemech.html

google: wustl neuromuscular splicemech

Much thanks to T. Wilson, UCSC!

Introns-early? Introns-late?



How to find splice points, using a BLAST search.

(1) Translate the DNA in all 6 frames.

(2) Search the database of **protein sequences** using the **translations** (blastx).

(3) Using the complete protein sequence, align it to the translation and find the regions of (near) *perfect identity*. These will abruptly end at the intron start site.

(4) Find the 5'-GT or 3'-AG signal at the point where the identity matches abruptly end.

(5) If your translation has an insertion with nearly perfect matches on either side, you have an alternative splicing.

In Class exercise: find the alternative spliced variants

Go to NCBI, search nucleotides for AKAP9 (you should get the sequence with accession number NM_005751.4, GI:197245395)

Select “BLAST sequence”

Select **blastx** (not **tblastx**)

Select the nr/nt database. Organism: homo sapiens

Submit.

While waiting, do exercise on the next page....

A sure sign of alternative splicing in *blastx* output:

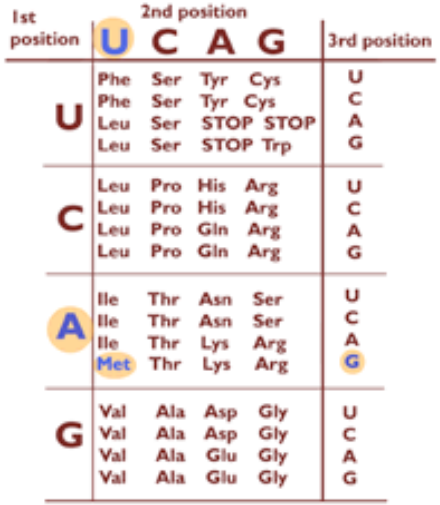
```
 Score =  160 bits (404), Expect = 8e-37
 Identities = 85/116 (73%), Positives = 86/116 (74%)
 Frame = +2

Query: 76820 RSHENGFMEDLDKTWVRYQECDSRSNAPATLTFENMAGAFSFIHSRVGSPWXXXXXXXXX 76999
              +SHENGFMEDLDKTWVRYQECDSRSNAPATLTFENMA
Sbjct: 778   KSHENGFMEDLDKTWVRYQECDSRSNAPATLTFENMA---------------------- 814

Query: 77000 XXXXRHTGVFMLVAGGIVAGIFLIFIEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ 77167
                     GVFMLVAGGIVAGIFLIFIEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ
Sbjct: 815   -------GVFMLVAGGIVAGIFLIFIEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ 863
```

Identical up to the insertion. Identical after the insertion. These must be the *same gene*.

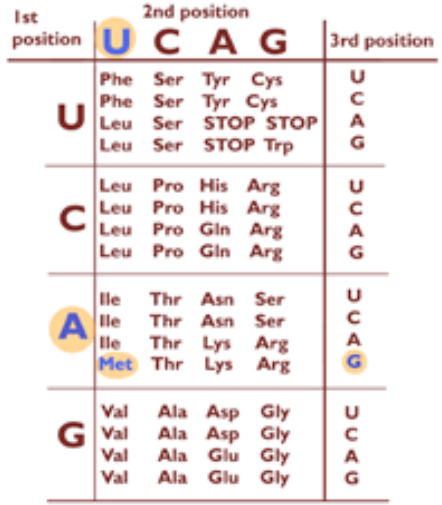
Which **codons** can come at the start/end of an alternative exon?



Frame :	0	1	2
The un-spliced intron starts with GU.	|GU*i*	*e*/GU	*ee*/G,U*ii*
The un-spliced intron ends with AG.	*i*AG|	*ii*A,G|*ee*	AG|*e*

e=a base within the exon.
i =a base within the intron.
| = intron/exon boundary.

Which **amino acids** can come at the start/end of an alternative exon?



Frame :	0	1	2
The un-spliced intron starts with GU.	/*V*	[CRSG]	{FINHYCD} [FLSYCW]
The un-spliced intron ends with AG.	[QKE]I	{FMNHYWCD} [VADEG]	[SR]

e=a base within the exon.
i =a base within the intron.
I = intron/exon boundary.

What frame is the intron in the earlier slide?

Exon,GU..AG,Exon
Is that all there is to it?

GU occurs on average every 16 nucleotides. AG, too.

If this were the only information, there would be too many splice sites.

GU..AG is necessary, not sufficient, for splicing.

What else is needed?

GU..AG

Spliceosome cuts before GU and after AG.
This is a **constraint**.

Frame of intron

Frame 0: intron starts at codon boundary

AGU CUU AUC UUU UCA GUU GGG ... CCG UAG AAC CAC UCG UAA

Frame 1: intron starts one after codon boundary

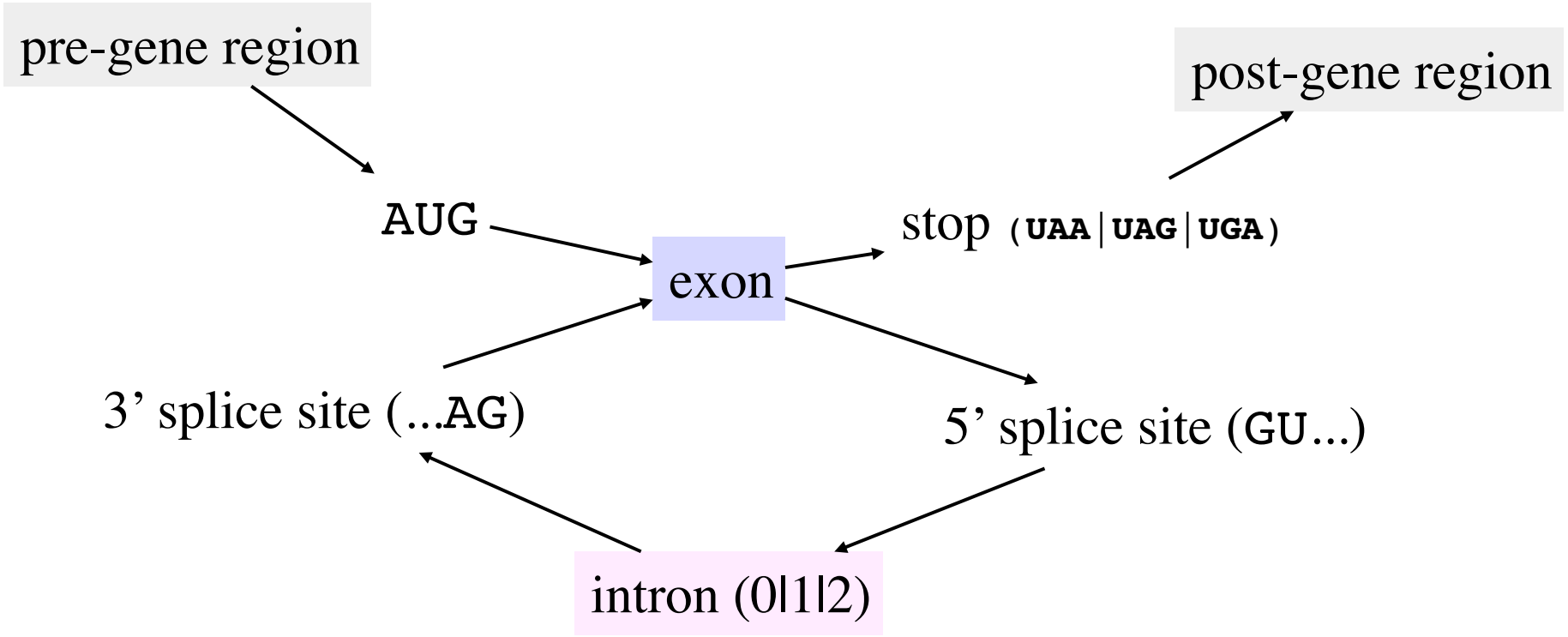
AGU CUU AUC UUU UCA UGU GGG ... CCG UAA GAC CAC UCG UAA

Frame 2: intron starts two after codon boundary

AGU CUU AUC UUU UCA GGG UGG ... CCG UAG AGC CAC UCG UAA

This must be multiple of 3 if the intron is alternatively spliced.

A generic gene sequence model for pre-mRNA

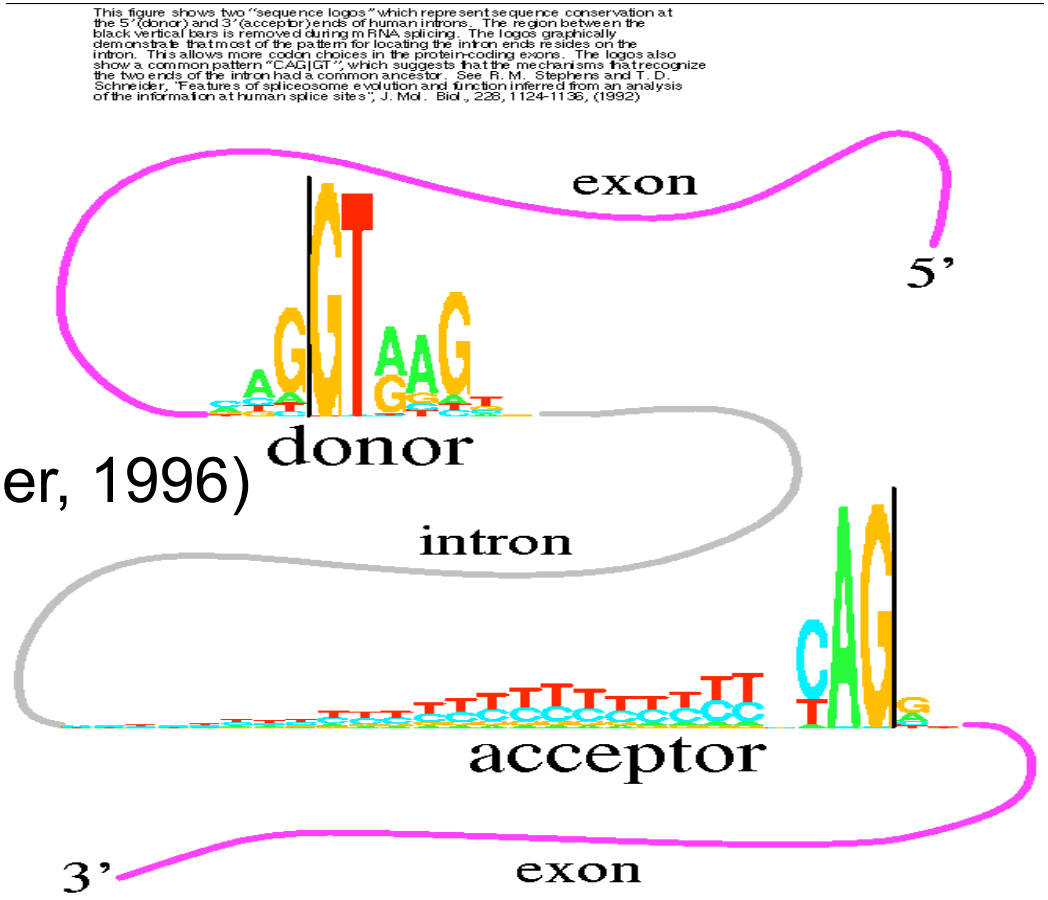


exon intron exon intron ...

XXXXXXATG...XXXGUX...XXAGXX...XXGUX...XXAG
XX...XXTAAXXX

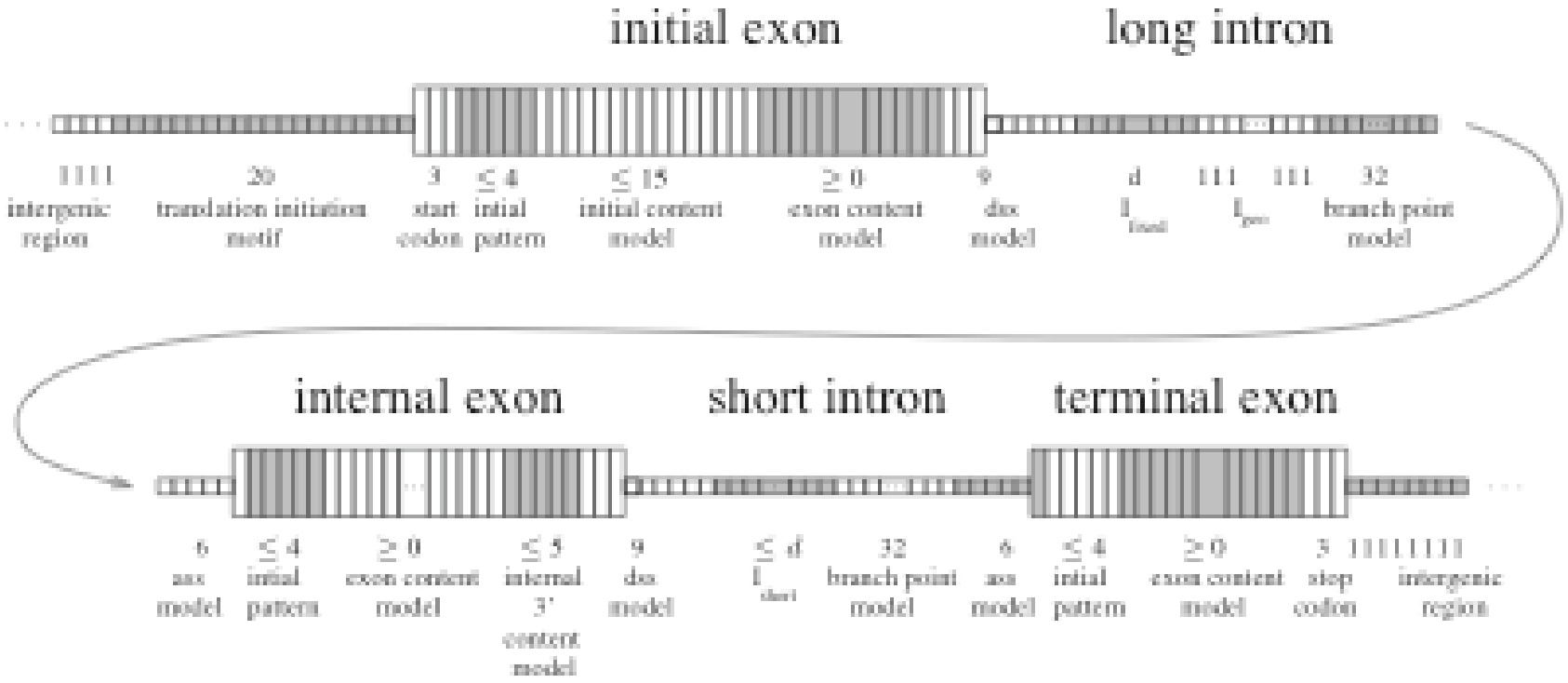
Donor and Acceptor Sites: Motif Logos

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



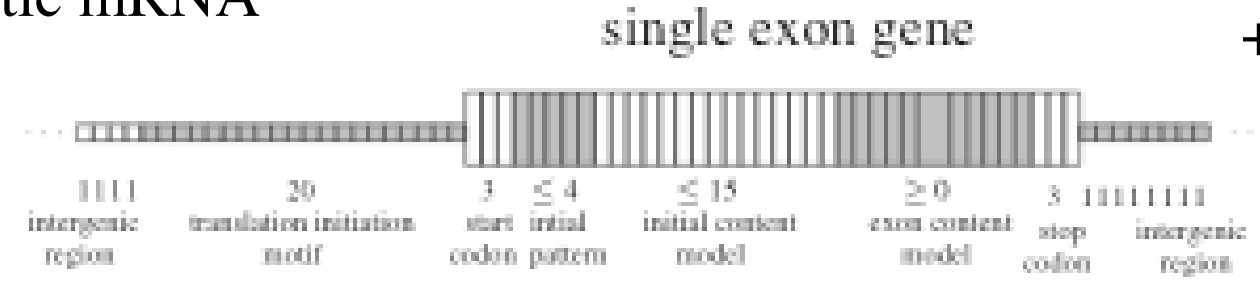
(http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html)

pre-mRNA structure



---

prokaryotic mRNA



Popular Gene Prediction Algorithms

- **GENSCAN**: uses Hidden Markov Models (HMMs)
- **TWINSCAN**
  - Uses both HMM and similarity (e.g., between human and mouse genomes)

The GENSCAN Algorithm

- Algorithm is based on probabilistic model of gene structure similar to *Hidden Markov Models (HMMs).*
- GENSCAN uses a training set in order to estimate the *HMM parameters*, then the algorithm returns the exon structure using maximum likelihood approach standard to many HMM algorithms (*Viterbi* algorithm).
  - Biological input: Codon bias in coding regions, gene structure (start and stop codons, typical exon and intron length, presence of promoters, presence of genes on both strands, etc)
  - Covers cases where input sequence contains no gene, partial gene, complete gene, multiple genes.

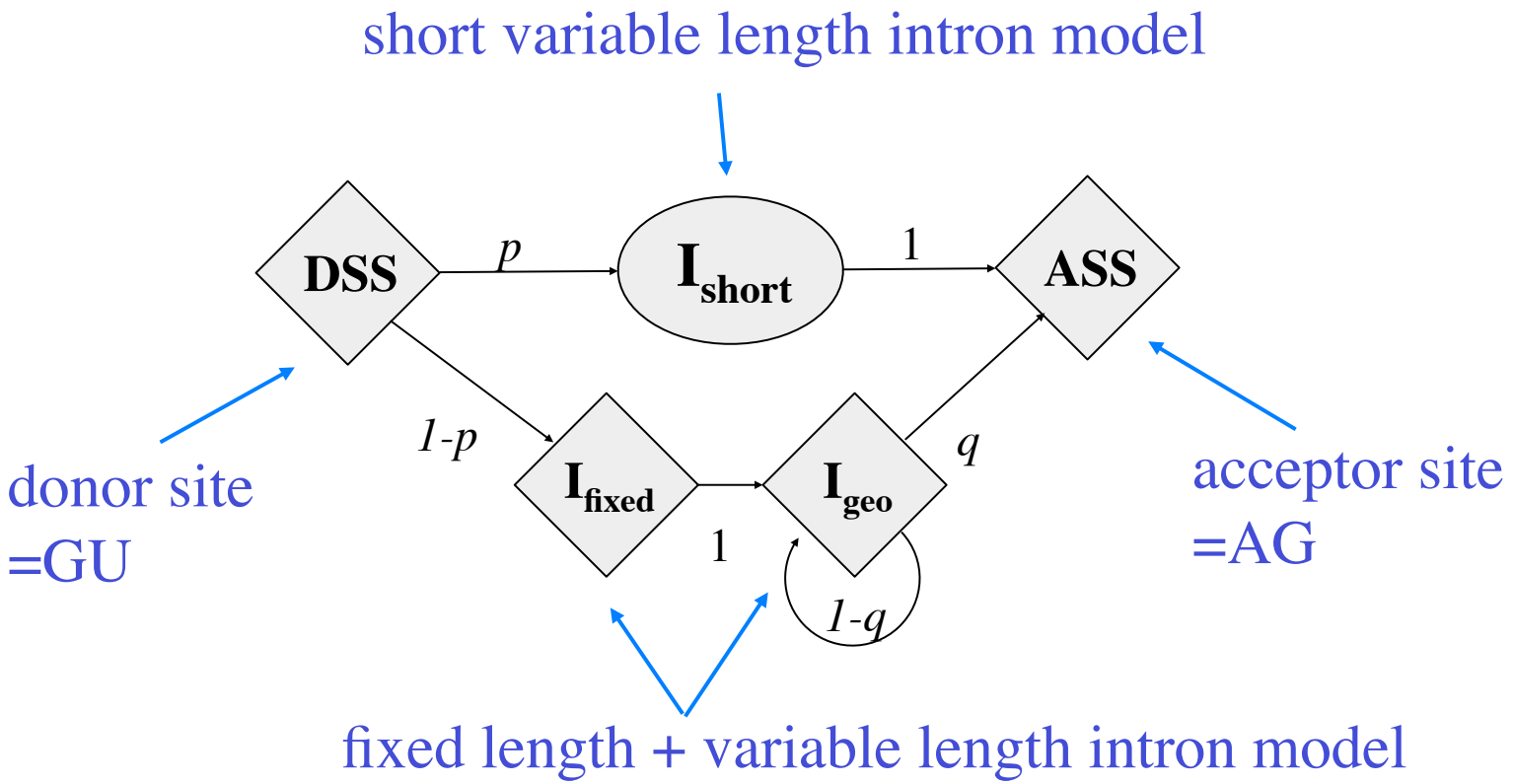
GENSCAN Limitations

- Does not use similarity search to predict genes.
- Does not address alternative splicing.
- Could combine two exons from consecutive genes together

GenomeScan

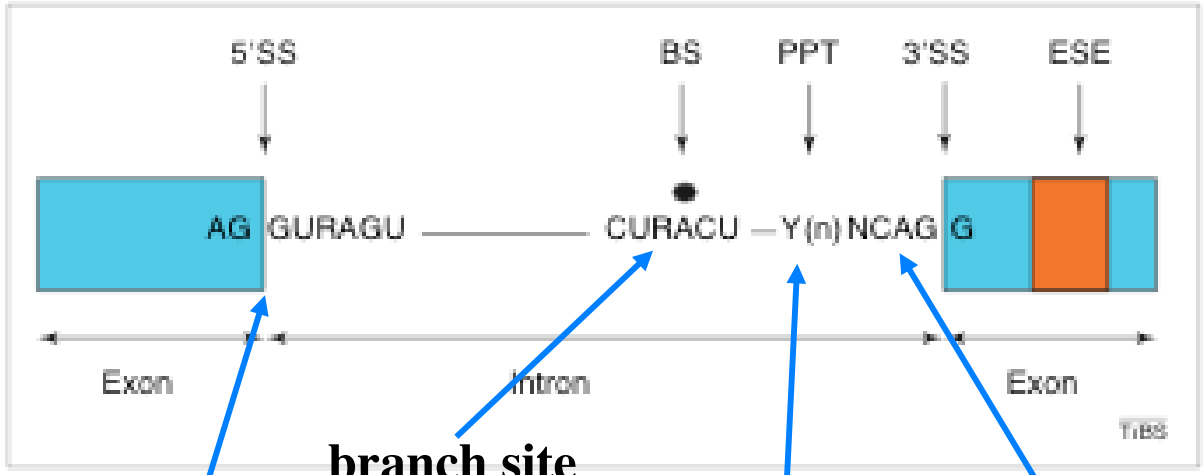
- Incorporates similarity information into GENSCAN: predicts gene structure which corresponds to maximum probability conditional on similarity information
- Algorithm is a combination of two sources of information
  - Probabilistic models of exons-introns
  - Sequence similarity information

A modular HMM for introns



Stanke M, Steinkamp R, Waack S, Morgenstern B. "AUGUSTUS: a web server for gene finding in eukaryotes. " Nucleic Acids Res. 2004 Jul 1;32:W309-12.

Intron model for mammals



branch site

poly-pyrimidine region

donor motif (contains GU)

acceptor motif (contains AG)

from: Blencowe, BJ. "Exonic splicing enhancers: mechanism of ction, diversity and role in human genetic diseases. TIBS 25:106 (2000)

A genefinding HMM: Genescan

internal exon model

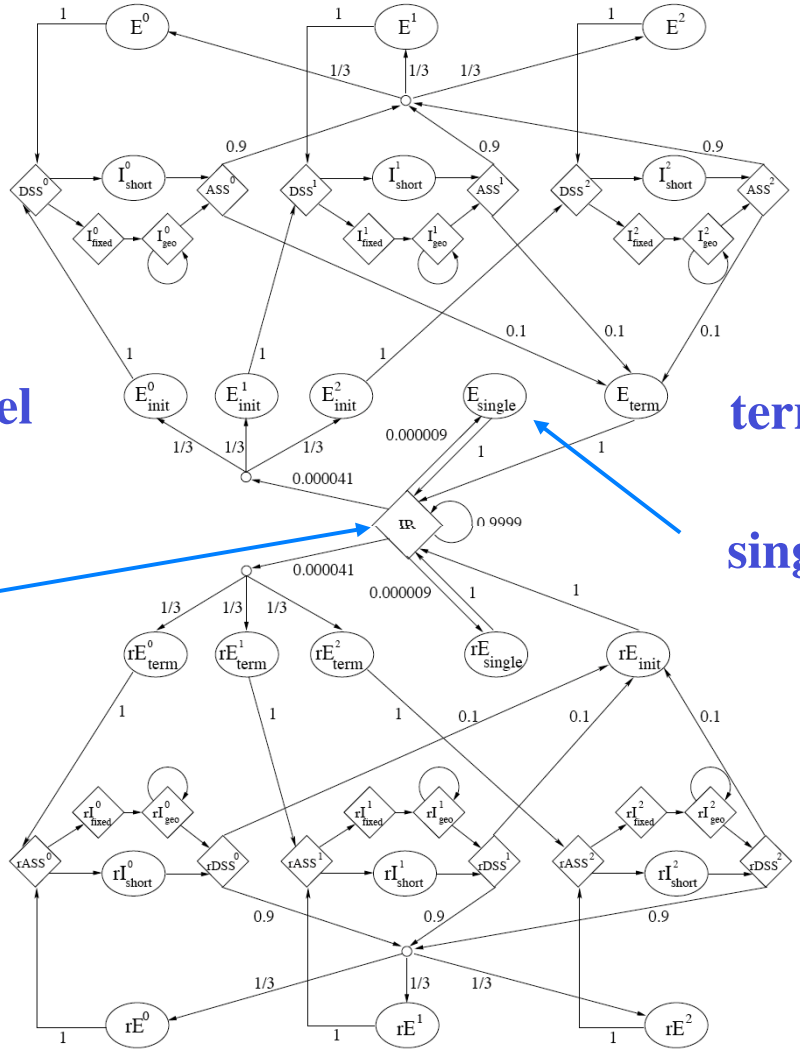
intron models

initial exon model

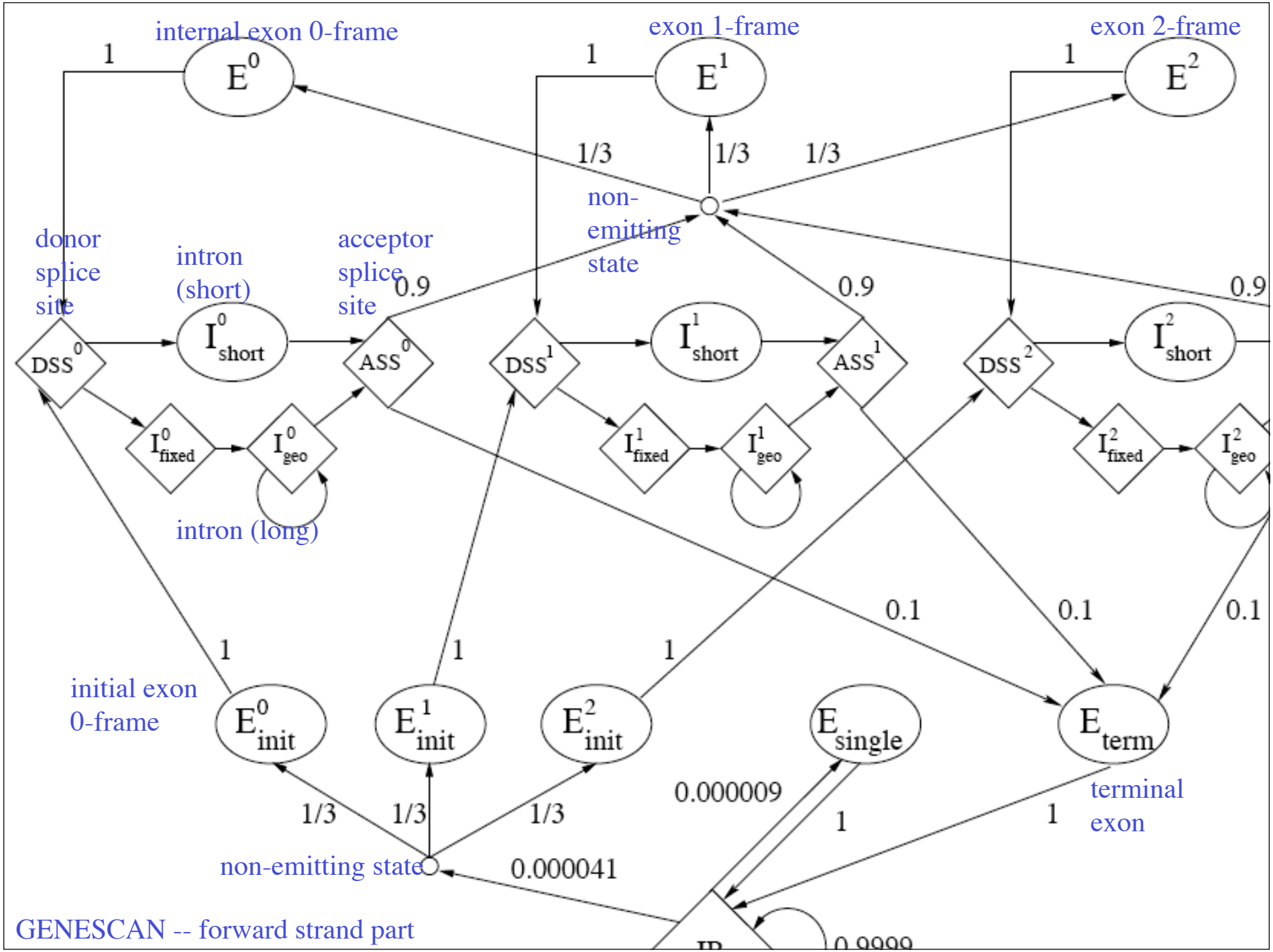
terminal exon model

single exon model

Intergentic Regions



Mirrored models for reverse complement strand


internal exon 0-frame
exon 1-frame
exon 2-frame
1
1
1
E^0
E^1
E^2
1/3
1/3
1/3
non-
emitting
state
donor
splice
site
intron
(short)
acceptor
splice
site
0.9
0.9
0.9
DSS^0
I^0_{short}
ASS^0
DSS^1
I^1_{short}
ASS^1
DSS^2
I^2_{short}
I^0_{fixed}
I^0_{geo}
I^1_{fixed}
I^1_{geo}
I^2_{fixed}
I^2_{geo}
intron (long)
0.1
0.1
0.1
1
1
1
initial exon
0-frame
E^0_{init}
E^1_{init}
E^2_{init}
E_{single}
E_{term}
0.000009
1
1
terminal
exon
1/3
1/3
1/3
non-emitting state
0.000041
0.9999
GENESCAN -- forward strand part

TwinScan

- Aligns two sequences and marks each base as gap (-), mismatch (:), match (|), resulting in a new alphabet of 12 letters: Σ {A-, A:, A |, C-, C:, C |, G-, G:, G |, T-, T:, T|}.
- Run Viterbi algorithm using emissions $e_k(b)$ where $b \in$ {A-, A:, A|, …, T|}.

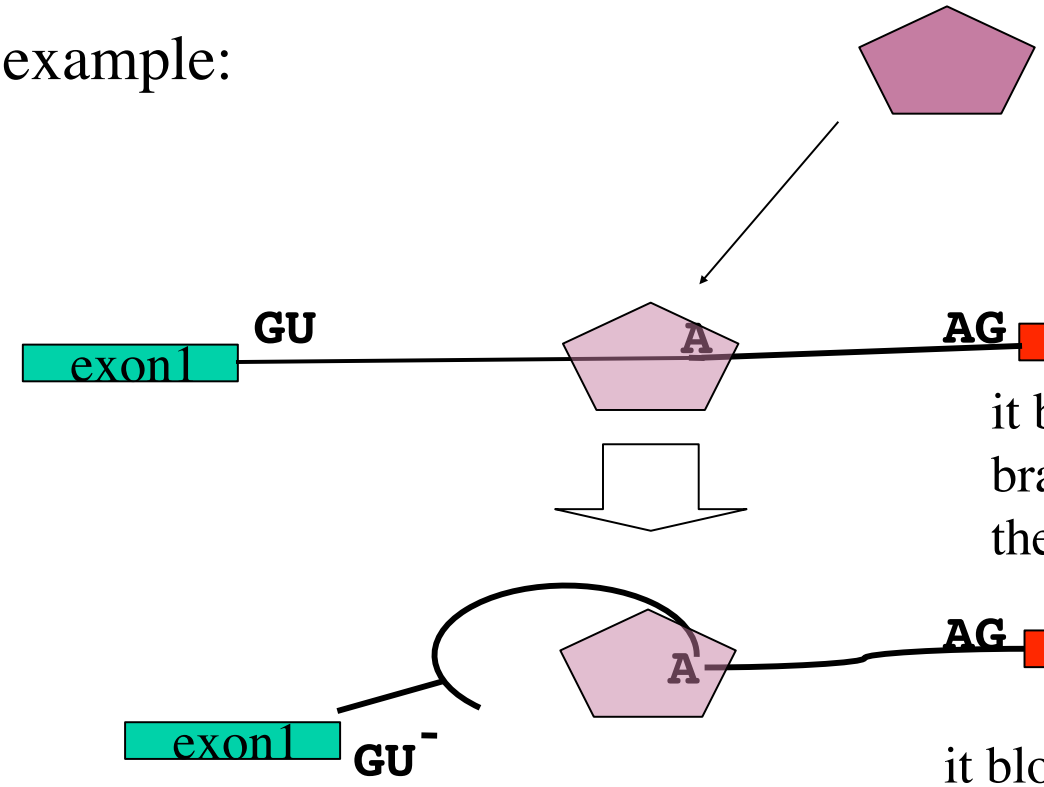
http://www.standford.edu/class/cs262/Spring2003/Notes/ln10.pdf

TwinScan (cont'd)

- The emission probabilities are estimated from from human/mouse gene pairs.
  - Ex. $e_I(x|) < e_E(x|)$ since matches are favored in exons, and $e_I(x-) > e_E(x-)$ since gaps (as well as mismatches) are favored in introns.
  - Compensates for dominant occurrence of poly-A region in introns

RNA binding proteins may selectively block splicing in some tissues.

For example:



an RNA binding protein is expressed in response to a stimulus.

it binds near the branchpoint (or one of the splicepoints)

it blocks, in this case, the cyclizing step.

What information is used to predict intron/exon boundaries?

- Introns always start with GU and end with AG (GT..AG in DNA)
- Introns can start in one of three "frames" (0|1|2) relative to the codon frame.
- Alternatively spliced introns (may be exons) must have a multiple of 3 nucleotides.
- 3' and 5' intron sequence motifs
- branchpoint sequence motif
- Enhancer/silencer sequence motifs (ESEs, ESSs, ISEs, ISSs)
- Base composition in exons/introns.
- Orthologs conserve intron/exon boundaries.

Sequence composition method for genefinding

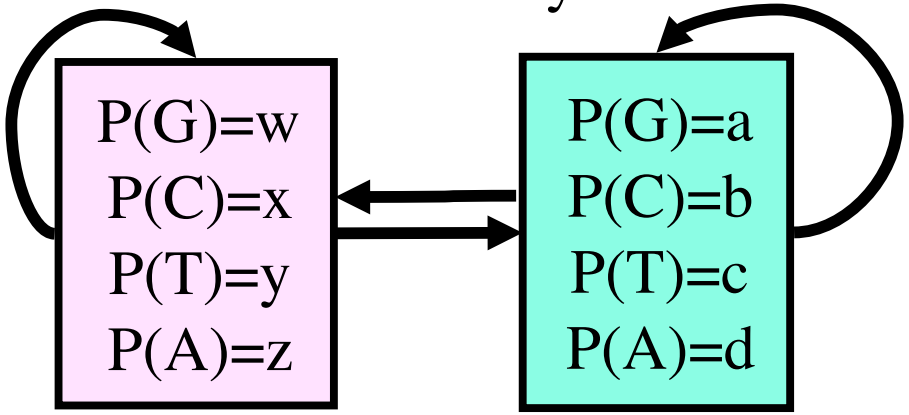
Most exons code for protein. Most introns do not.

Selective pressure on exons includes:

(1) species-specific codon preferences

(2) amino acid preferences

(3) selection for “foldability” and function.



A simple HMM for intron/exon base composition. Not so specific.

TestCode

- Statistical test described by James Fickett in 1982: tendency for nucleotides in coding regions to be repeated with periodicity of 3
  - Judges randomness instead of codon frequency
  - Finds "putative" coding regions, not introns, exons, or splice sites
- TestCode finds ORFs based on compositional bias with a periodicity of three

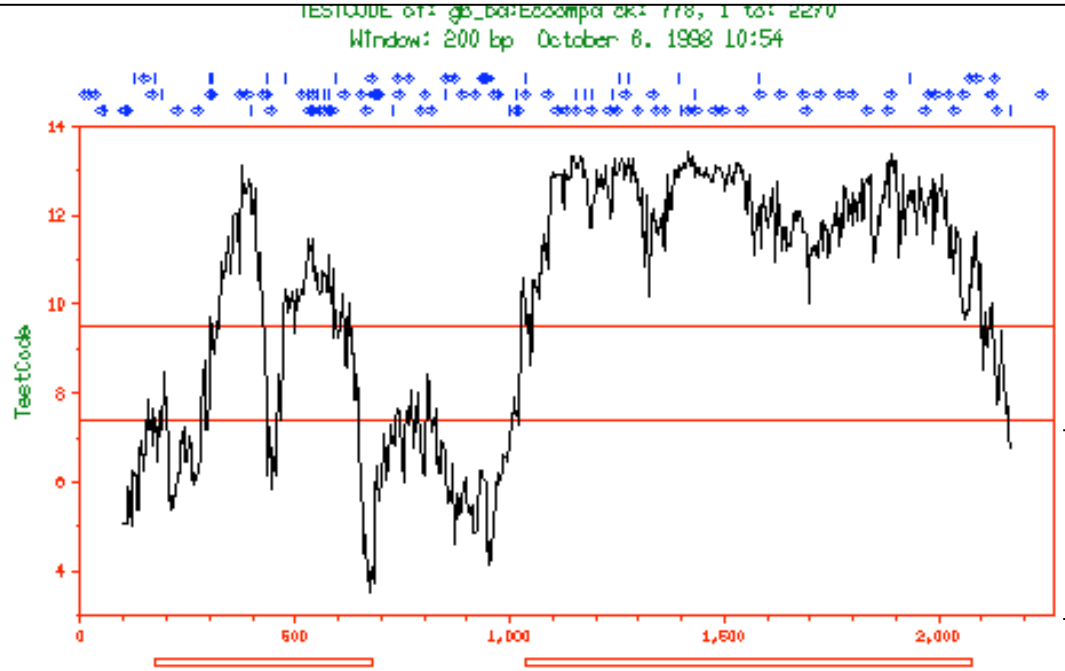
TestCode Statistics

- Define a window size no less than 200 bp, slide the window the sequence down 3 bases. In each window:
  - Calculate for each base {A, T, G, C}
    - max (n_{3k+1}, n_{3k+2}, n_{3k}) / min (n_{3k+1}, n_{3k+2}, n_{3k})
    - Use these values to obtain a probability from a lookup table (which was a previously defined and determined experimentally with known coding and noncoding sequences

TestCode Statistics (cont'd)

- Probabilities can be classified as indicative of " coding" or "noncoding" regions, or "no opinion" when it is unclear what level of randomization tolerance a sequence carries
- The resulting sequence of probabilities can be plotted

TestCode Sample Output



Coding

No opinion

Non-coding

Splicing-related motifs

ESEs, ESSs, ISEs, ISSs

ESE =Exonic Splicing enhancers: sequence in the *exons* that *promote* splicing

ESS =Exonic Splicing Silencers: sequence in the *exons* that *inhibit* splicing

ISE =Intronic Splicing Enhancers: sequence in the *introns* that *promote* splicing.

ISS =Intronic Splicing Silencers: sequence in the *introns* that *inhibit* splicing

How were ESEs found?

(1) Training database was constructed of exonic mRNA (post-spliced) that was (a) constitutively spliced (not alternatively spliced), and (b) from an *internal non-protein-coding exon*.

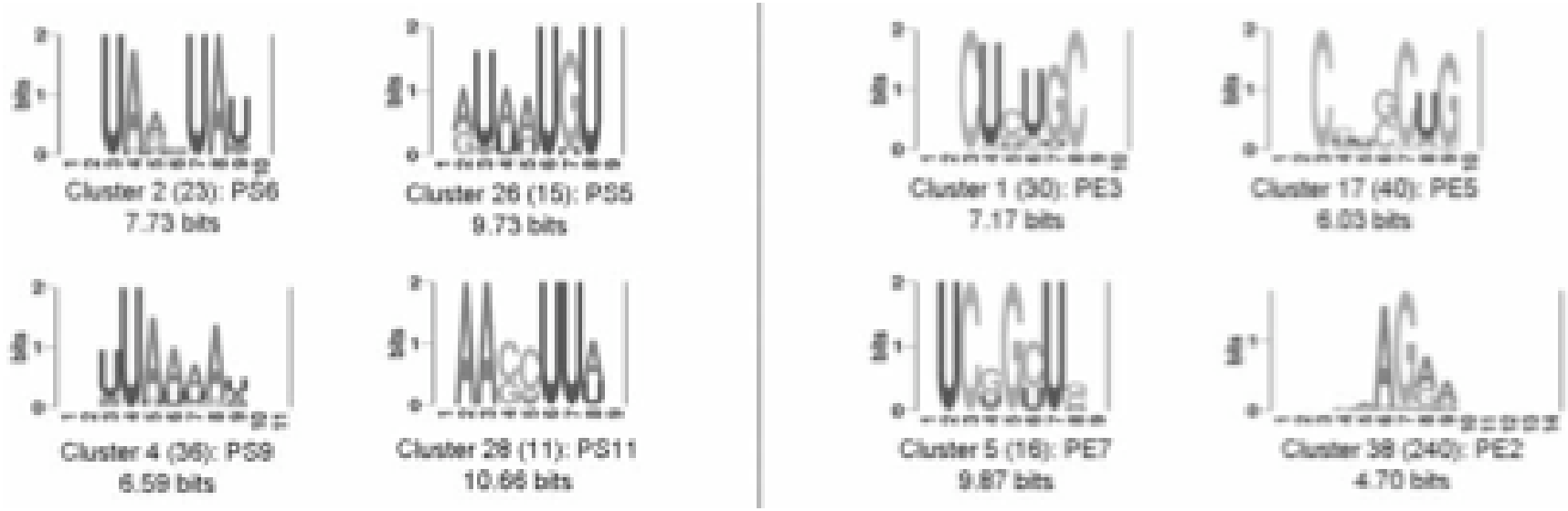
(2) Database of 'control', non-ESE sequences was constructed.

(3) The relative abundance of all "8-mers" was found.

(4) 8-mers with high relative abundance were tested by mutating the putative ESE 8-mers and determining the splicing efficiency by gel electrophoresis.

Zhang,XHF. and Chasin LA. "Computational definition of sequence motifs governing constitutive exon splicing." Genes & Development 18: 1241-1250 (2004)

Relative abundance ESE and ESS motifs



putative ESSs **putative ESEs**

Some of the motifs found by Zhang & Chasin using relative abundance analysis of 8-mers, after clustering.

Splicing facts

Exons average 145 nucleotides in length

- **Contain regulatory elements :**
  - **ESEs: Exonic splicing enhancers**
  - **ESSs: Exonic splicing silencers**

Introns average more than 10x longer than exons

- **Contain regulatory elements(bind regulatory complexes)**
  - **ISEs: Intronic splicing enhancers**
  - **ISSs: Intronic splicing silencers**

Splice sites

- **5' splice site**
  - **Sequence: AGguragu (r = purine)**
  - **U1 snRNP: Binds to 5' splice site**
- **3' splice site**
  - **Sequence: yyyyyyy nagG (y= pyrimidine)**
- **Branch site**
  - **Sequence: ynyuray (r = purine)**
  - **U2 snRNP: Binds to branch site via RNA:RNA interactions between snRNA and pre-mRNA**

Alternative splicing fact sheet

Alternative splicing

- **Definition: Joining of different 5' and 3' splice sites**
- **~80% of alternative splicing results in changes in the encoded protein**
- **Up to 59% of human genes express more than one mRNA by alternative splicing**
- **Functional effects: Generates several forms of mRNA from single gene**
- **Allows functionally diverse protein isoforms to be expressed according to different regulatory programs**

Structural effects:

- **Insert or remove amino acids**
- **Shift reading frame**
- **Introduce termination codon**
- **Gene expression effects**
- **Removes or inserts regulatory elements controlling translation, mRNA stability, or localization**

Regulation

- **Splicing pathways modulated according to:**
  - **Cell type**
  - **Developmental stage**
  - **Gender**
  - **External stimuli**