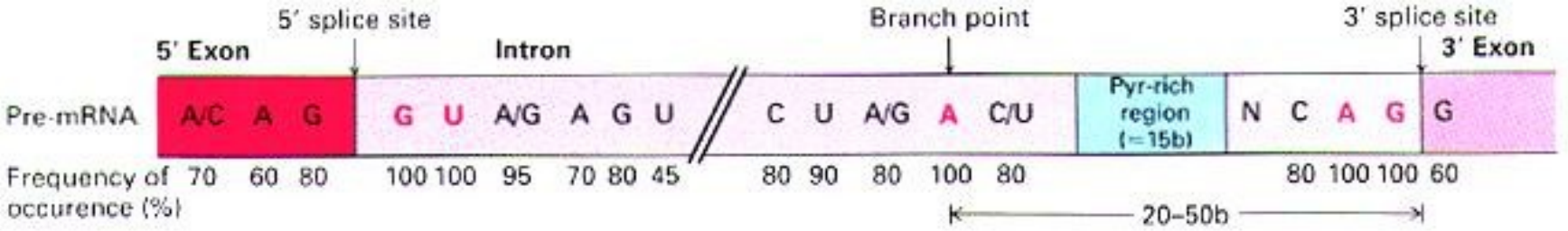
Sequence Motifs



Genomes do not store the coding information of a protein consecutively

AFF3 has 23 exons

Genes consist of coding segments (exons) and non-coding segments (introns)

Introns are removed from mRNA during splicing

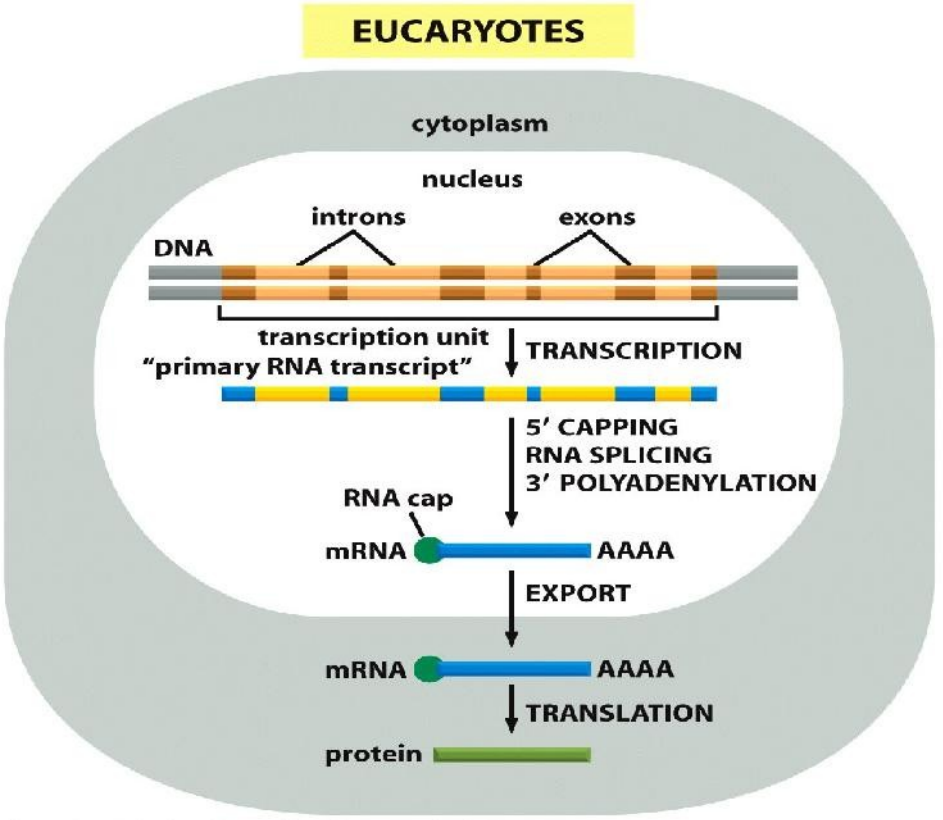


Figure 6-21a Molecular Biology of the Cell 5/e (© Garland Science 2008)

Spliced mRNA is capped at the 5' end (beginning) and gets a poly(A) tail at the 3' end (end)

```
begin {
   mRNA
}
end
```

Different cells combine the exons of a gene in different ways (alternative splicing)

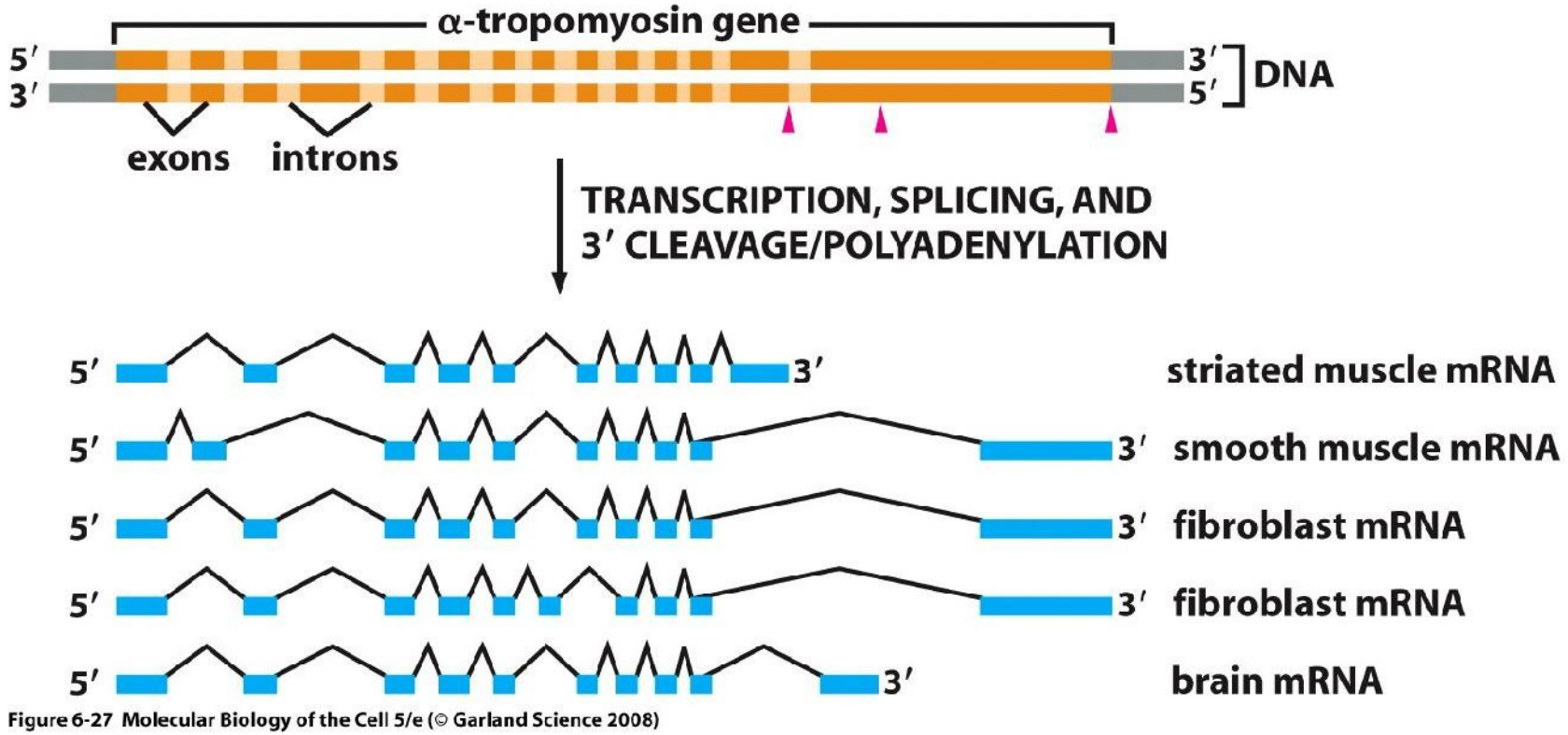


Figure 6-27 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Alternative splicing expands the human transcriptome by one order of magnitude

The cell recognizes sequence patterns at the splice junctions

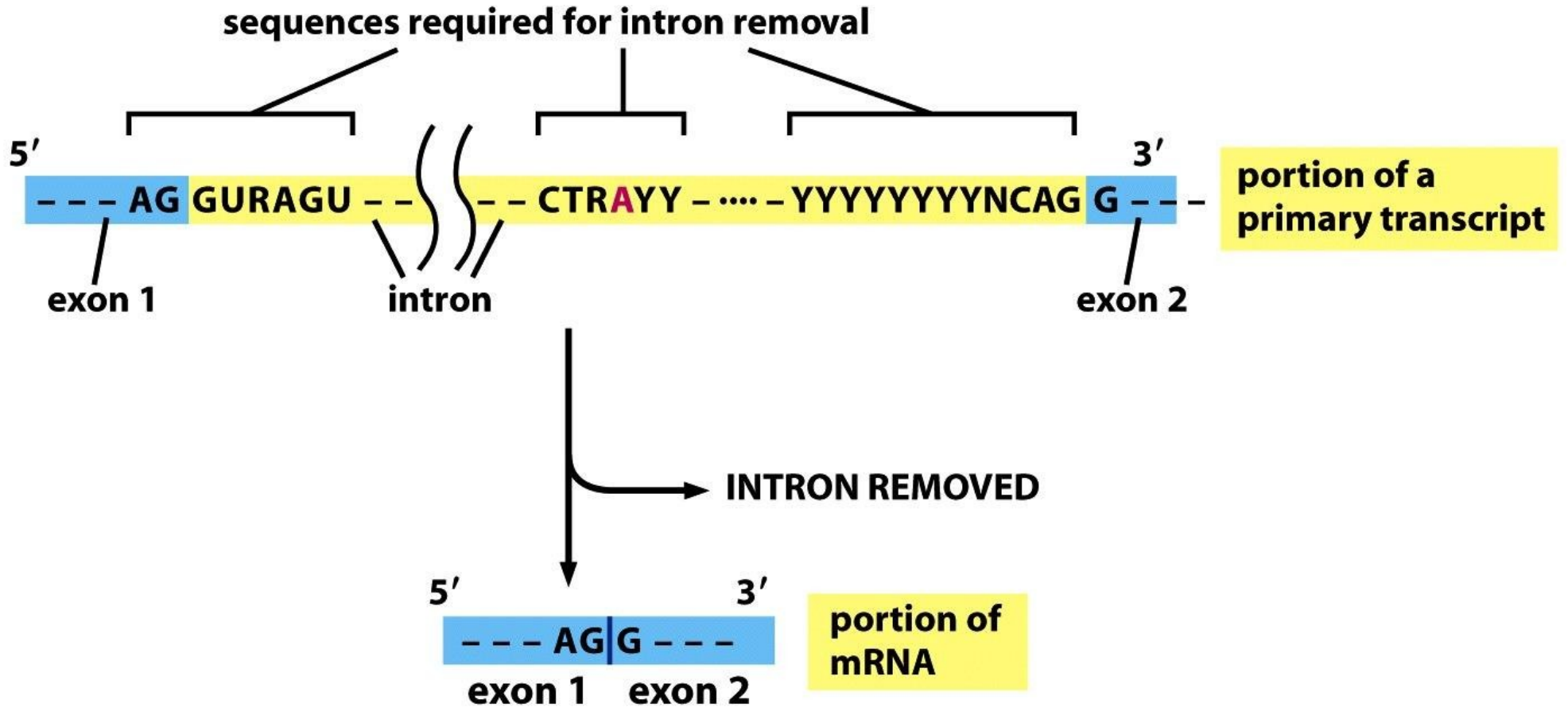


Figure 6-28 Molecular Biology of the Cell 5/e (© Garland Science 2008)

What are these patterns?

The splice machinery finds the splice junctions. How do we find them?

tcacccccttctccaggcgtgcagcccttcggcgtgccgctgtccatgccaccagtgatg

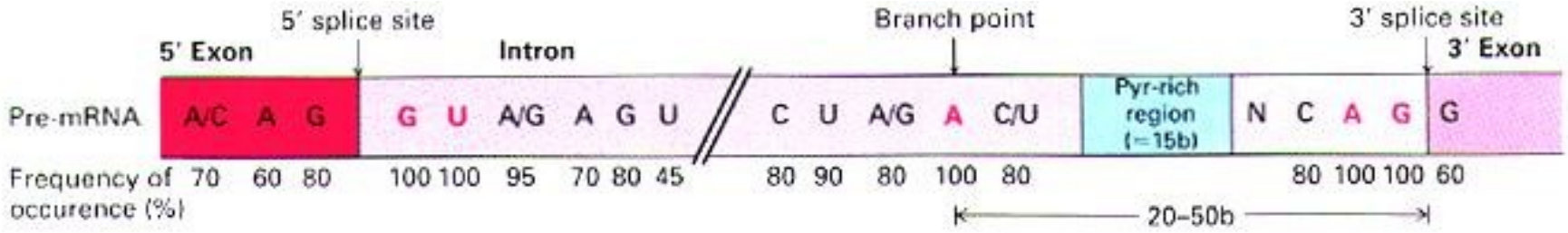
gcagctgccctctcgcggcatggaatacggagcccggggatcctgcccgtcatccagccg

gtggtggtgcagcccgtccccttttatgtacacaagtcacctccagcagcctctcatggtc

tccttatcggaggagatggaaaattccagtagtagcatgcaagtacctgtaattgaatca

tatgagaagcctatatcacagaaaaaaattaaaatagaacctgggatcgaaccacagagg

Many introns begin with GT and end with AG



Donor **Acceptor**

The splice site at the beginning of an intron is the donor sequence
The splice site at the end of an intron is the acceptor sequence

If all GT sites in the genome were donor sites the splice machinery would shredder the genome

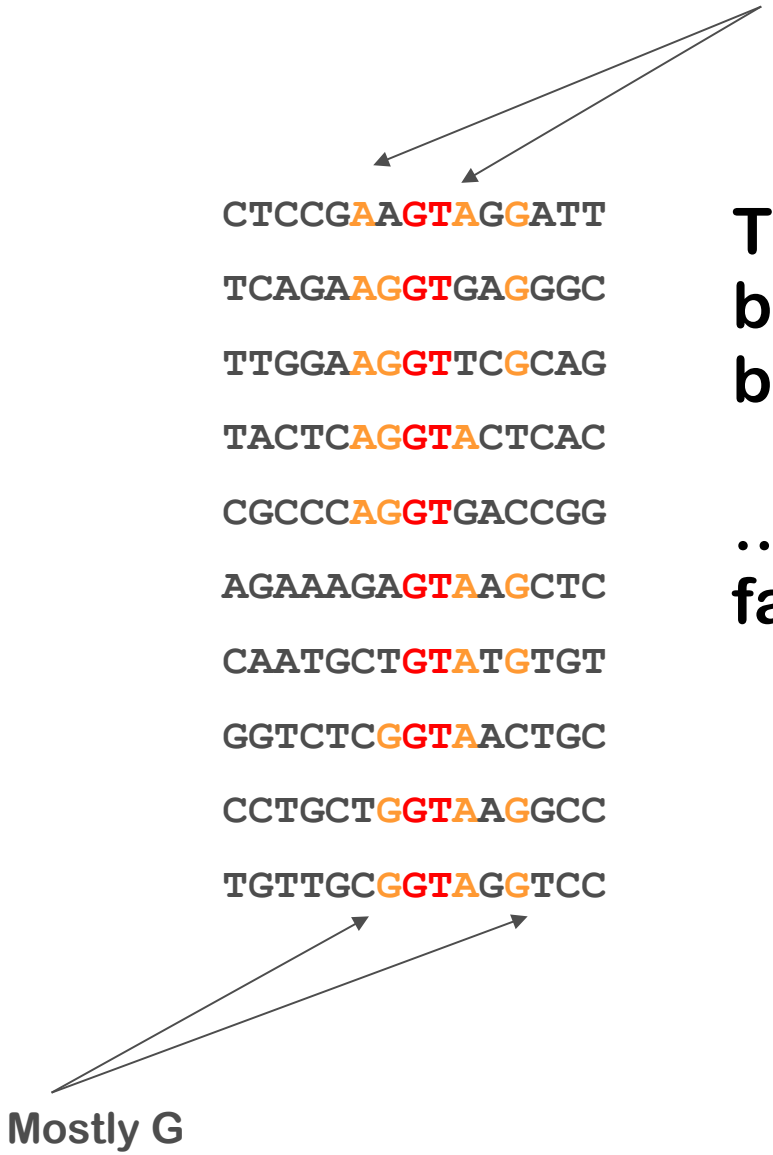
GGATCCCTGGAAAATGGAGAAGCTGTGCTAATAGAGGGGGGCCAGAAATCCCCACTCTAGAATGCT
GTAGAATGTTGGGAGACACCCAGGATGTGAGCCAGGGACTTTCTGGAAGTGTTTGTTCTGGCCCCA
CCCGACCCCAGGCAGTCCCCAGCTGTCTGCACAGTCGGATGGGGAGGGGGCTTGCACAGAGTTGGA
GCCAGAGGAGAGAGCTGGCTCATCCCCTACGGTAGGATGGGGAAACCTCACAGACCACATTGTCAC
CCGGCCTCAGCTCTCCGCCCCGGCGCTCAGAGGGTAACTCTCACCCACCTCGTCCGCTTCTCTGAA
CCAGAGTGACCCAGGCTGCCGTCCGCCCCGCTCTCCTACCCCGAGTTGGCACGGAGGTATAGCGCC
AGAGGGGGGCCCCAGGCGCCCCGAGGTCTTGCCCTTGCGGCTTTCCCTTTGCGGGGTGGGCGCCT
TCTTCCGGGTAGGGGCCACGTGGCCCTGGCCGGGCGGGGGGCTCGGCCCACCCCGCGCCGGGCCCA
GTGACTCAGGCCGCAGCTGTTACCGCGTCACATGAGGGAGGCCGGCGGCCACTCGGCGGGGGAGGG

Introns have a length between 100 and 50,000 bases

Consecutive GT sites are spaced much closer

The context of the GT position matters

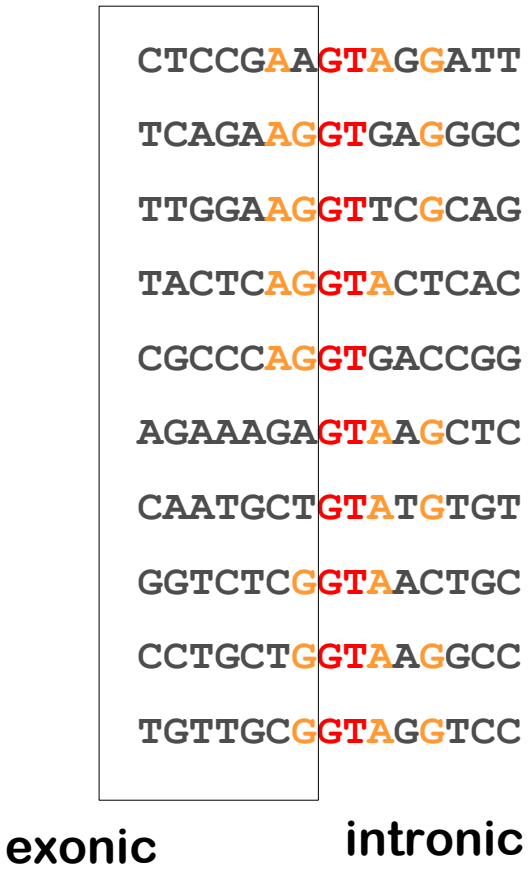
The distribution of bases up- and downstream of the GT is not uniform



There is no unique binding sequence but a binding motif

… like for transcription factor binding sites

The exonic part of the binding sequence "codes twice"



1. It codes for a protein

2. It codes for binding of the splicing machinery

Both functions of the sequence are under selective pressure

Three candidate donor sites

CTCCGAAGTAGGATT

TCAGAAGGTGAGGGC

TTGGAAGGTTCGCAG

TACTCAGGTACTCAC

CGCCCAGGTGACCGG

AGAAAGAGTAAGCTC

CAATGCTGTATGTGT

GGTCTCGGTAACTGC

CCTGCTGGTAAGGCC

TGTTGCGGTAGGTCC

3 candidates

AGGTACG

CCGTCCC

TGGTCCG

Which one is most likely to be a real donor?

We can calculate relative base frequencies for all positions

```
123456789
CTCCGAAGTAGGATT
TCAGAAGGTGAGGGC
TTGGAAGGTTCGCAG
TACTCAGGTACTCAC
CGCCCAGGTGACCGG
AGAAAGAGTAAGCTC
CAATGCTGTATGTGT
GGTCTCGGTAACTGC
CCTGCTGGTAAGGCC
TGTTGCGGTAGGTCC
```

Position 3:	Position 5:	Position 7:
A: 20%	A: 0%	A: 50%
C: 0%	C: 0%	C: 20%
G: 70%	G: 0%	G: 20%
T: 10%	T: 100%	T: 10%

Or with more than just 10 real donors …

```
# DONOR FREQUENCY MATRIX from http://genomic.sanger.ac.uk/spldb/SpliceDB.html
   1       2       3       4       5       6       7       8       9
A  34.08   60.36    9.14    0.00    0.00   52.57   71.26    7.08   15.98
C  36.24   12.90    3.27    0.00    0.00    2.82    7.56    5.50   16.46
G  18.31   12.48   80.34  100.00    0.00   41.94   11.76   81.35   20.90
T  11.38   14.25    7.24    0.00  100.00    2.55    9.29    5.88   46.16
```

Such a matrix is also called a profile of the sequence motif

We can use it to find more donor sites in the genome

For each position we can associate the relative frequencies with a discrete distribution

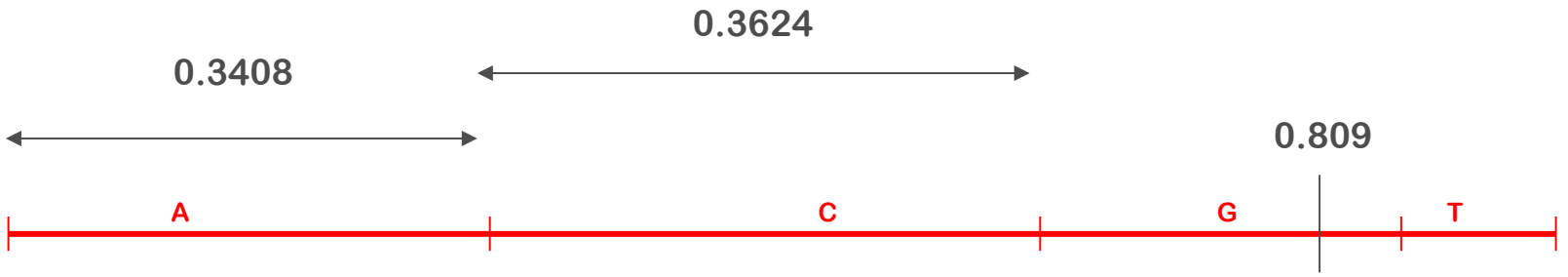
Position 1 of the donor profile

A	34.08
C	36.24
G	18.31
T	11.38

We can associate a random experiment with this distribution:

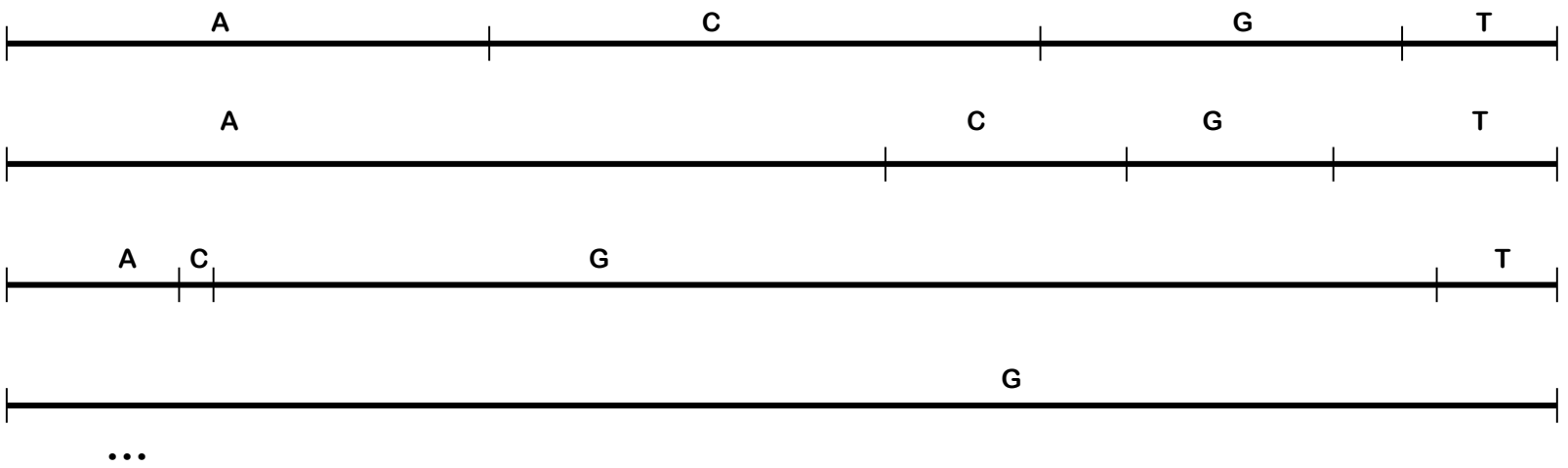
Generate a uniform random number
U=0.809

Outcome: G



We can associate such an experiment with every position in the profile

	1	2	3	4	5	6	7	8	9
A	34.08	60.36	9.14	0.00	0.00	52.57	71.26	7.08	15.98
C	36.24	12.90	3.27	0.00	0.00	2.82	7.56	5.50	16.46
G	18.31	12.48	80.34	100.00	0.00	41.94	11.76	81.35	20.90
T	11.38	14.25	7.24	0.00	100.00	2.55	9.29	5.88	46.16



Consider every candidate sequence to be an outcome of the experiment

	1	2	3	4	5	6	7	8	9
A	34.08	60.36	9.14	0.00	0.00	52.57	71.26	7.08	15.98
C	36.24	12.90	3.27	0.00	0.00	2.82	7.56	5.50	16.46
G	18.31	12.48	80.34	100.00	0.00	41.94	11.76	81.35	20.90
T	11.38	14.25	7.24	0.00	100.00	2.55	9.29	5.88	46.16

P (AAGGTACGT) ≈ 0.34*0.6*0.8*1*1*0.53*0.08*0.81*0.46 = 0.0026

P (CCCGTCCCC) ≈ 0.36*0.13*0.03*1*1*0.03*0.08*0.06*0.16 = 3.23e-08

P (CTGGTCCGA) ≈ 0.36*0.14*0.8*1*1*0.03*0.08*0.81*0.16 = 1.25e-05

P (TACCTCCGT) = 0

The probability of a sequence as outcome of the experiment scores candidate sequences

The experiment can generate any sequence with a GT in the middle

But the sequences do not have the same probability

Those that look like donor sites have higher probabilities than those that do not

If we screen a genome, real donor sites should generate high probabilities

0.0026 is still a small probability

	1	2	3	4	5	6	7	8	9
A	34.08	60.36	9.14	0.00	0.00	52.57	71.26	7.08	15.98
C	36.24	12.90	3.27	0.00	0.00	2.82	7.56	5.50	16.46
G	18.31	12.48	80.34	100.00	0.00	41.94	11.76	81.35	20.90
T	11.38	14.25	7.24	0.00	100.00	2.55	9.29	5.88	46.16

P (AAG**GT**ACGT) ≈ 0.34*0.6*0.8*1*1*0.53*0.08*0.81*0.46 = 0.0026

P (CCC**GT**CCCC) ≈ 0.36*0.13*0.03*1*1*0.03*0.08*0.06*0.16 = 3.23e-08

P (CTG**GT**CCGA) ≈ 0.36*0.14*0.8*1*1*0.03*0.08*0.81*0.16 = 1.25e-05

P (TAC**CT**CCGT) = 0

Does this mean that the probability that sequence 1 is a donor is low ?

No. It means that the probability that a donor is identical to our sequence is low. There are many other sequences that can be functional donors.

A background model can give us an idea whether the sequence is likely to be a donor or not

We have a model for donors

Let's build a similar model for non-donors

If the sequence is just regular genome every base should occur on all positions with the same probability of 0.25 (Background Model (Q))

P (AAGGTACGT) = 0.34*0.6*0.8*1*1*0.53*0.08*0.81*0.46 = 0.0026

Q (AAGGTACGT) = $(0.25)^9$ = 3.815e-06
The odds are 700:1 in favor of the donor model

The back ground model is less likely

P (AAGGTACGT) ≈ 0.34*0.6*0.8*1*1*0.53*0.08*0.81*0.46 = 0.0026

Q (AAGGTACGT) = $(0.25)^9$ ≈ 3.815e-06

A sequence is either a donor or it is not. However, the two numbers do not add up to one.

The probabilities must refer to non-complementary events.

In fact they refer to the same event: Seq=AAGGTACGT

… but to different models: P vs Q.

We can use likelihood ratios to compare two competing models

For any sequence s we can calculate the likelihood ratio:

$$\rho(s) = \frac{P(s)}{Q(s)}$$

For AAGGTACGT the odds are 700:1 in favor of the donor model

For CCCGTCCCC the odds are 120:1 in favor of the background model

For CTGGTCCGA the odds are 3:1 in favor of the donor model

We compared two models of sequence 1

P (AAGGTACGT) ≈ 0.34*0.6*0.8*1*1*0.53*0.08*0.81*0.46 = 0.0026

Q (AAGGTACGT) = $(0.25)^9$ ≈ 3.815e-06

We used two different models and decided which one can explain the sequence better

The two numbers 0.0026 and 3.815e-06 are properties of the models P and Q

We say:

Model P has likelihood 0.0026
Model Q has likelihood 3.815e-06

Probabilities refer to outcomes of random experiments

P(AAGGTACGT) = 0.0026

P(CCCGTCCCC) = 3.23e-08

The model (random experiment) is always the same.

The events vary.

Probabilities are properties of events.

Likelihoods refer to models

P(AAGGTACGT) = 0.0026

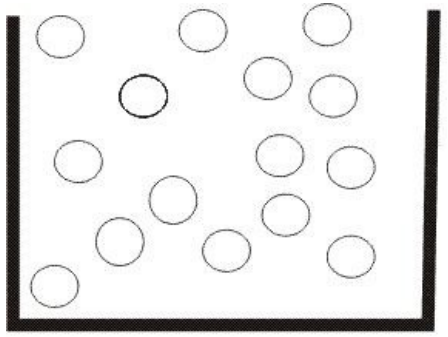
Q(AAGGTACGT) = 3.815e-06

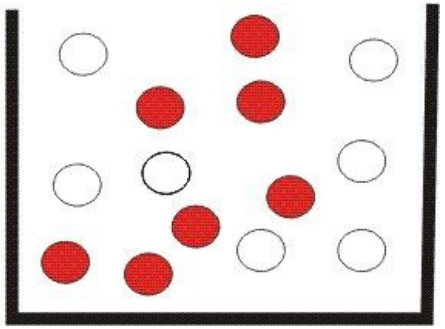
The data (sequence) is always the same

The models vary

Likelihoods are properties of models

A canonical experiment helps us understand likelihood ratios





There are two boxes with balls:

Box 1: all balls are white

Box 2: 50% white balls, 50% red balls

Someone gives you *n* balls sampled with replacement from one of the boxes.

They are all white.

Which box did he sample from?

We are in a modeling context

The data is fixed: *n* white balls

We have two competing models: Box 1 and Box 2

The likelihood ratio is:

$$\frac{L(Box1)}{L(Box2)} = \frac{1}{(1/2)^n} = 2^n$$

The more white balls we observe the more evidence we have in favor of the Box that has only white balls.

We can translate likelihood ratios into numbers of white balls

White balls = $\log_2(\rho)$

For AAGGTACGT the odds are 700:1 in favor of the donor model.

A log ratio of 700 corresponds to a little more than 9 white balls.

Likelihoods can be small

P (CCCGTCCCC) ≈ 0.36*0.13*0.03*1*1*0.03*0.08*0.06*0.16 = 3.23e-08

The likelihood is the product of numbers smaller than 1

If there are many, the computer will round to 0 and we have LR=0/0.

The problem is just a technical problem due to limited CPU memory used to store a real number.

Log-Likelihoods do not suffer from numerical problems

$$LLR(S) = \log \prod_{i=1}^{n} \frac{P(s_i)}{Q(s_i)} = \sum_{i=1}^{n} \log \frac{P(s_i)}{Q(s_i)}$$

Profile scores and weight matrices

Score for position i:

$$\text{Score}_i(s_i) = \log \frac{P(s_i)}{Q(s_i)}$$

Score for the full sequence:

$$\text{Score}(S) = \sum_i \text{Score}_i(s_i)$$

We call the score of a motif profile score.

We can define the profile by a 4 x 9 weight matrix W.

W_{ij} = Score for observing letter *i* at position *j*

Log-likelihood scores are a standard search strategy in bioinformatics

They are used for:

Protein alignments
(searching for the right alignment)

Genomic database searches

Search for protein domains

Search for members of protein families

Search for transcription factor binding sites

Gene finding

Pattern finding in general

The general setup is:

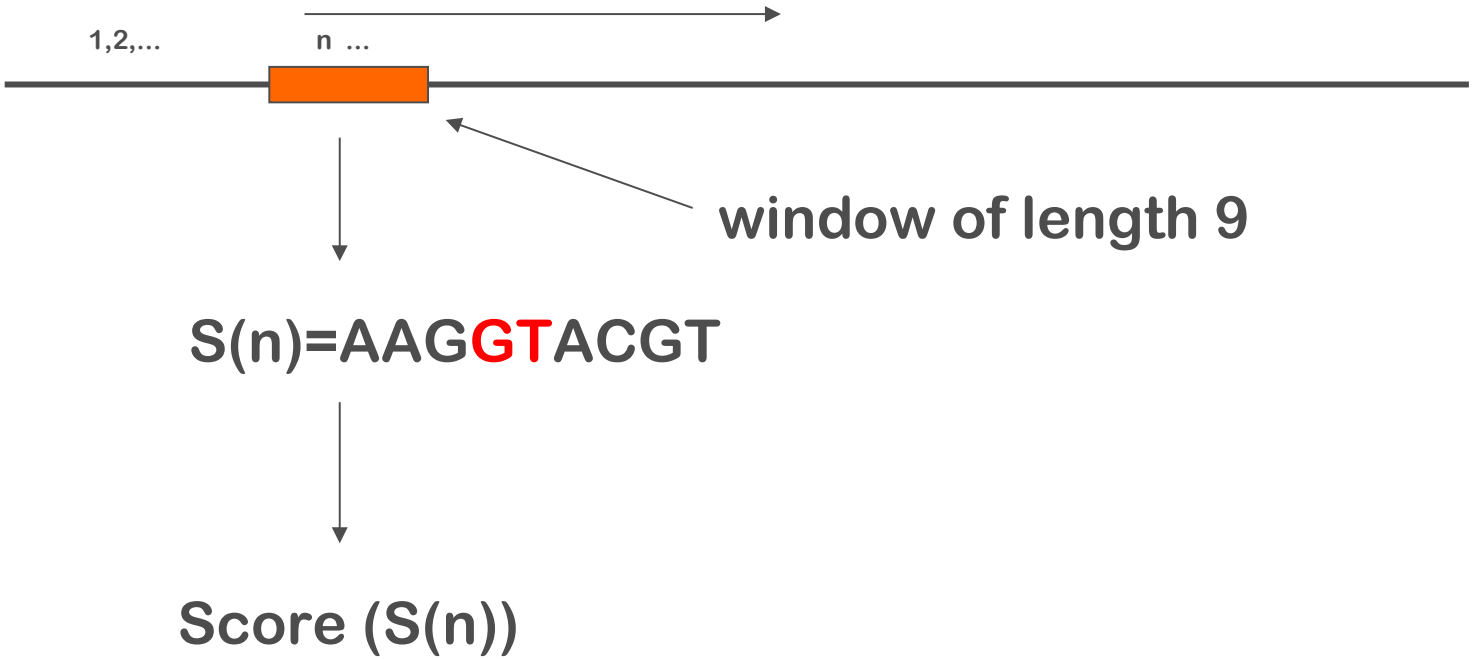
P is a model for what you are searching

Q is a background model

Search for high scoring sequences using

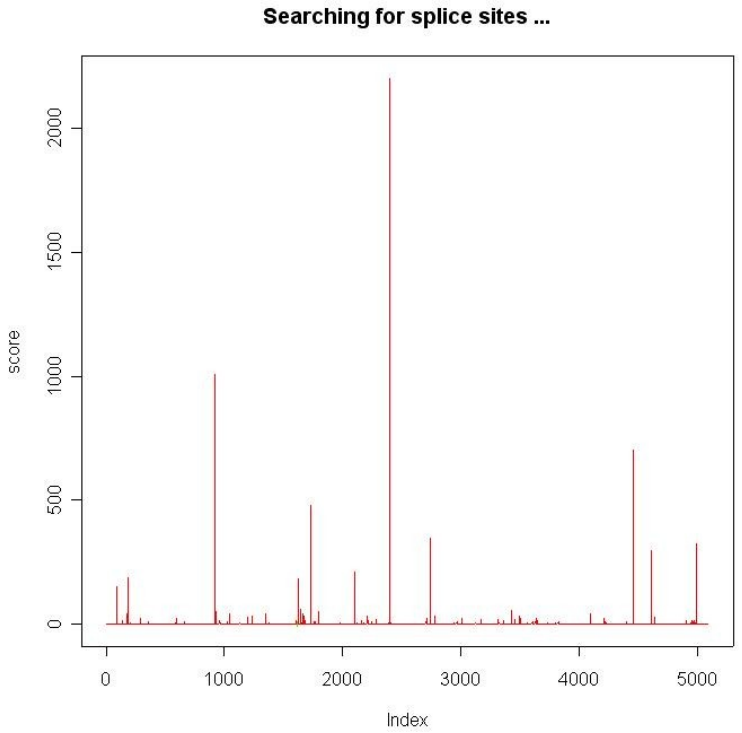
$$S(\cdot) = \sum \log \frac{P(\cdot)}{Q(\cdot)}$$

Back to donors in genomic sequences



To search a long RNA sequence for donor splice sites, we can slide a window of length 9 along the sequence and score every window using the profile score

The donor profile score applied to the RNA of human Ribosomal-Protein S6



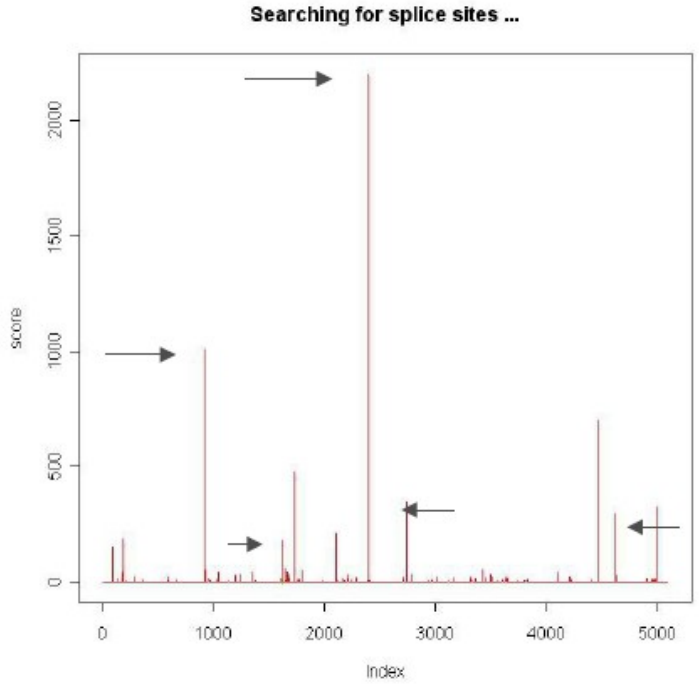
Every GT position generates a peak

Some peaks are negative (not shown) meaning that the context does not look like a donor site

Other peaks are positive and large

Are they real donors?

The RNA of human Ribosomal-Protein S6 has 5 splice sites



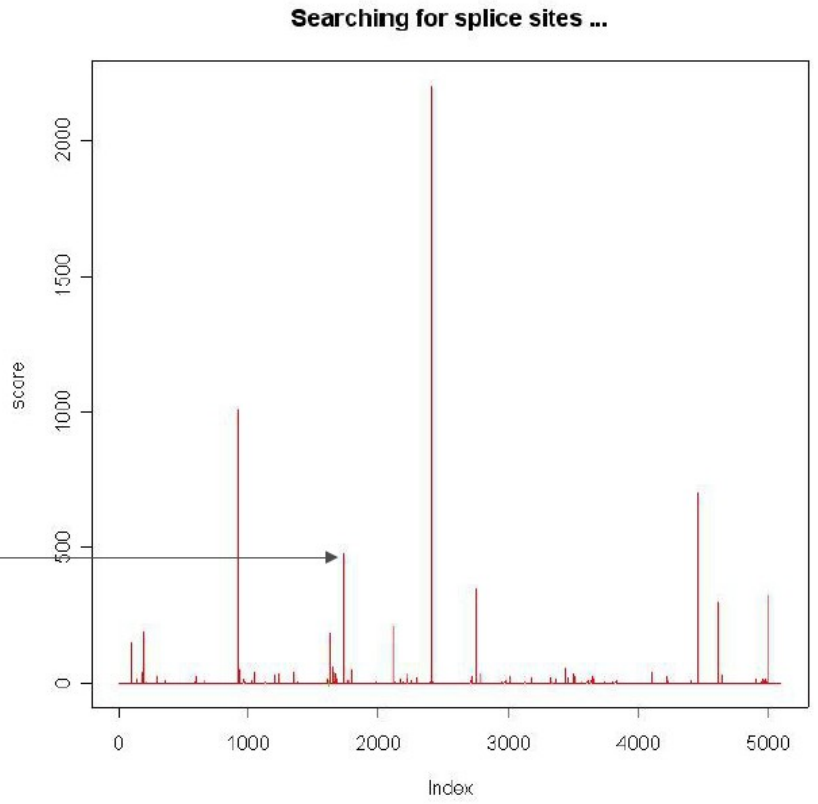
They all produce notable peaks

There are also high peaks that do not correspond to real donor sites

This position corresponds to

$\log_2 500 \approx 9$ white balls

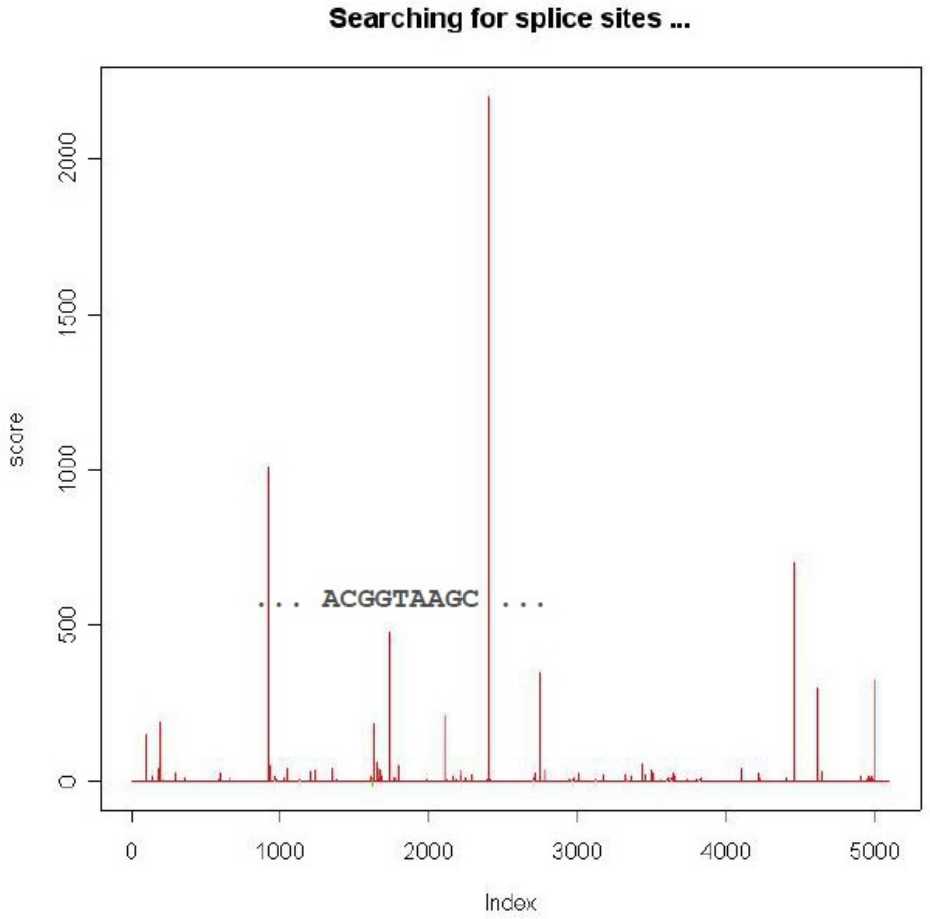
But it is no donor



All models are wrong. Some are useful.

George Box

The sequences of the false positive hits look like donor sites



The splice machinery can see that it is no donor site.

It has more information than just the RNA sequence.

RNA structure for example

Our model is not perfect but it is useful: It could detect all real splice sites by large peaks

The modeling approach filters GT positions

GGATCCCTGGAAAATGGAGAAGCTGTGCTAATAGAGGGGGGCCAGAAATCCCCACTCTAGAATGCTGTAGAATGT
TGGGAGACACCCAGGATGTGAGCCAGGGACTTTCTGGAAGTGTTTGTTCTGGCCCCACCCGACCCCAGGCAGTCC
CCAGCTGTCTGCACAGTCGGATGGGGAGGGGGCTTGCACAGAGTTGGAGCCAGAGGAGAGAGCTGGCTCATCCCC
TACGGTAGGATGGGGAAACCTCACAGACCACATTGTCACCCGGCCTCAGCTCTCCGCCCCGGCGCTCAGAGGGTA
ACTCTCACCCACCTCGTCCGCTTCTCTGAACCAGAGTGACCCAGGCTGCCGTCCGCCCCGCTCTCCTACCCCGAG
TTGGCACGGAGGTATAGCGCCAGAGGGGGGCCCCAGGCGCCCCGAGGTCTTGCCCTTGCGGCTTTCCCTTTGCGG
GGGTGGGCGCCTTCTTCCGGGTAGGGGCCACGTGGCCCTGGCCGGGCGGGGGGCTCGGCCCACCCCGCGCCGGGC
CCAGTGACTCAGGCCGCAGCTGTTACCGCGTCACATGAGGGAGGCCGGCGGCCACTCGGCGGGGGAGGG

The RNA of human Ribosomal-Protein S6 has 208 GT positions.

Only 11 of them score higher than 100.

These 11 include all 5 real donor sites.

The model generates hypotheses on potential splice sites

Genomics is an exploratory science

Scientists go out into the wilderness of genomes and dig for gold (drug targets).

But where to dig?

The log odds score makes suggestions (generates hypotheses).

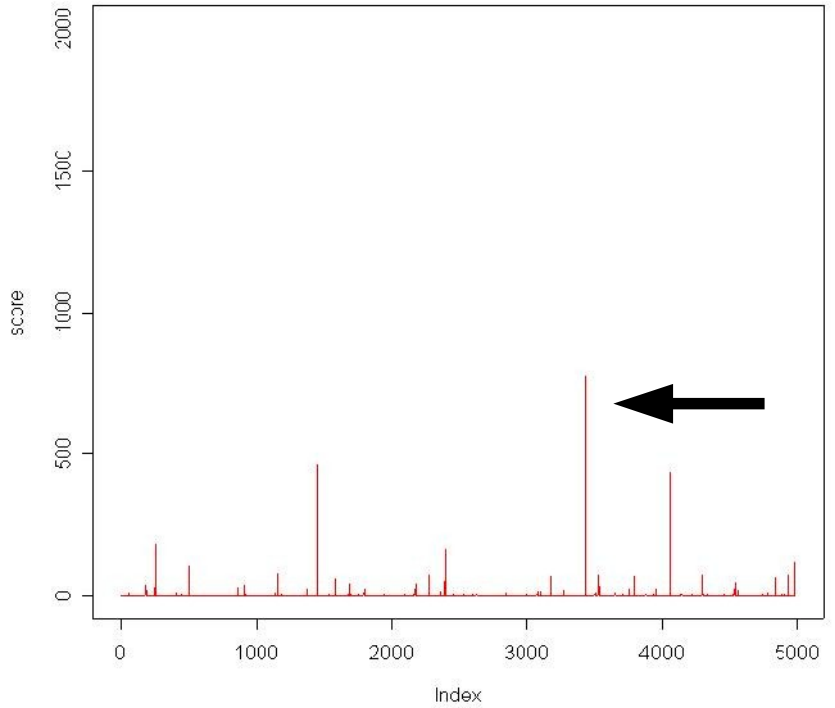
They will not all be correct.

We still have to prove the hypothesis.

But we know where to start.

Long genomes generate fata morganas

Random sequences include sequences that look like donors

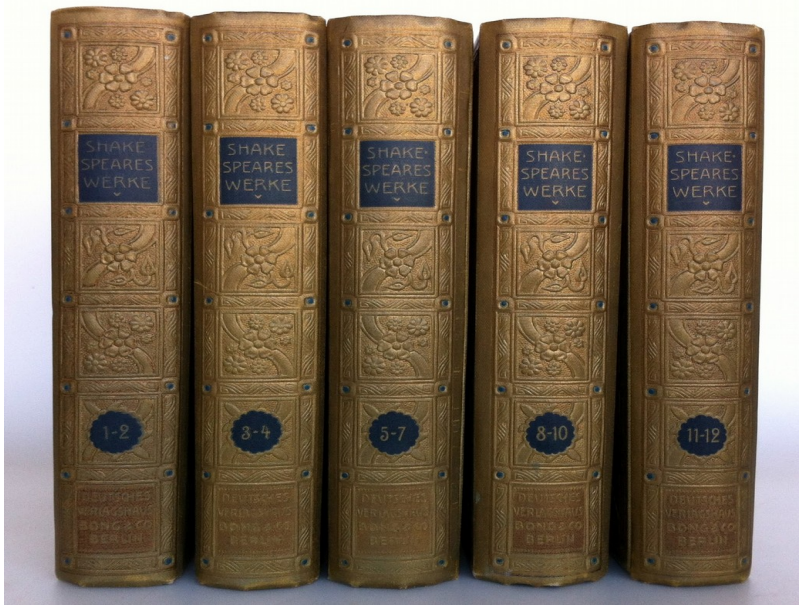


Your score see things that do not exist

The Borel Cantelli Lemma is responsible for fata morganas in random sequences

Lemma von Borel Cantelli:
Every infinitely long random sequence of letters contains every finite text infinitely often

This includes the complete works of William Shakespeare



End of Chapter 7