GENETICS AND MOLECULAR BIOLOGY FOR ENVIRONMENTAL ANALYSIS

MOLECULAR ECOLOGY
LESSON 15: GENOME ANNOTATION

Prof. Alberto Pallavicini

pallavic@units.it

CATEGORIES OF GENE PREDICTION PROGRAMS

- The current gene prediction methods can be classified into two major categories,
  - ***ab initio-based*** and
  - ***homology-based*** approaches.

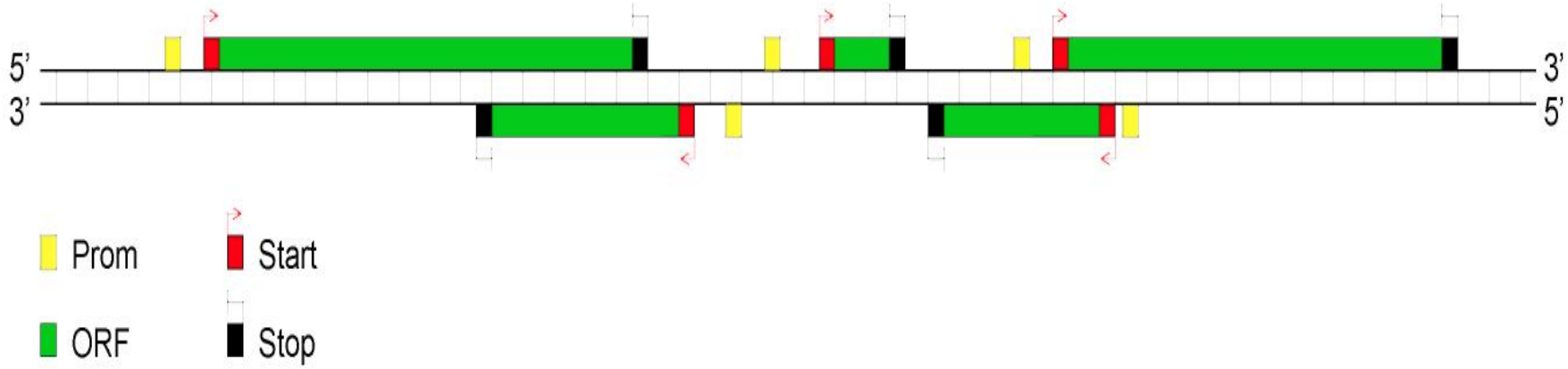
PROKARYOTIC GENE PREDICTION

Through decades of research and development, much progress has been made in prediction of prokaryotic genes.

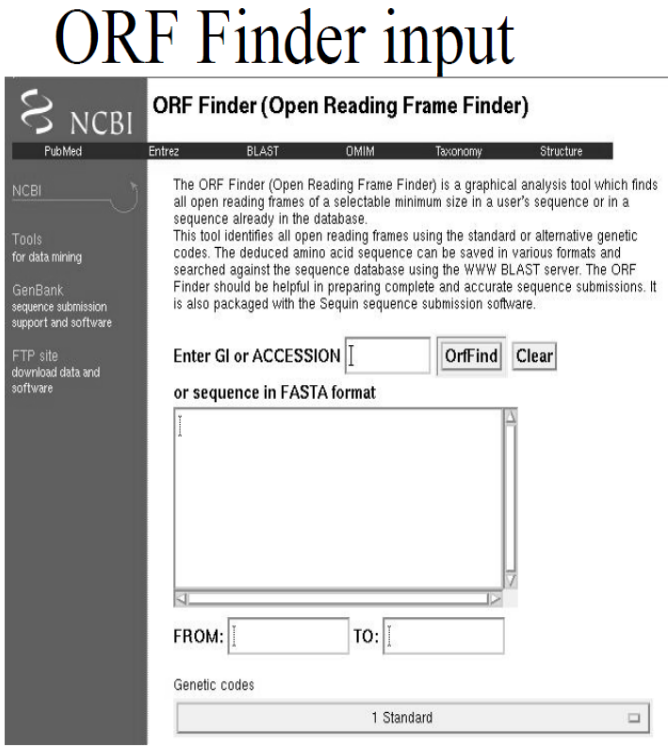
A number of gene prediction algorithms for prokaryotic genomes have been developed with varying degrees of success. Algorithms for eukarytotic gene prediction, however, are still yet to reach satisfactory results.

- How to pick a threshold length that will provide us with believable ORFs as candidate genes?

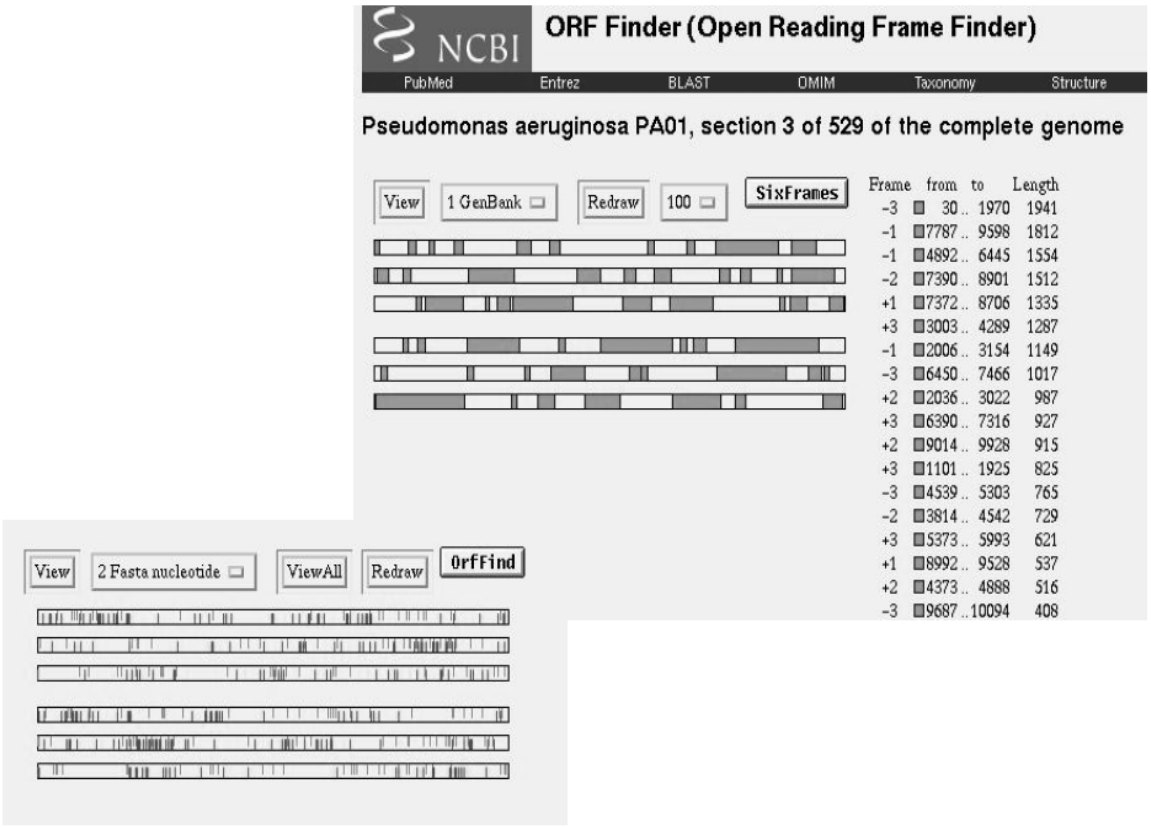
- Simple gene structure
- Overlapping genes



ORF FINDER


ORF Finder input
NCBI
ORF Finder (Open Reading Frame Finder)
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
NCBI
Tools
for data mining
GenBank
sequence submission
support and software
FTP site
download data and
software
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.
Enter GI or ACCESSION
OrfFind
Clear
or sequence in FASTA format
FROM:
TO:
Genetic codes
1 Standard

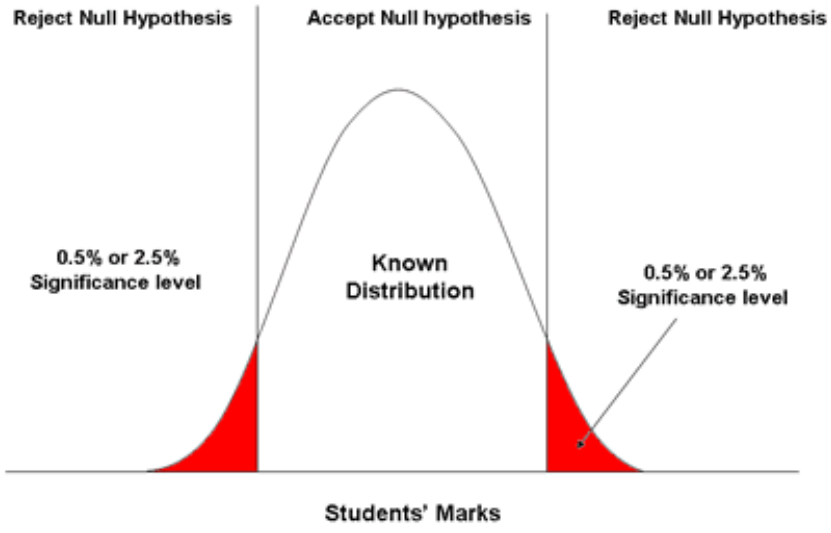
ORF finder results



DETECTING SPURIOUS SIGNALS: HYPOTHESIS TESTING

- We consider the data (e.g. an ORF of a certain length) to be ***significant*** when it is highly unlikely under the null model.

- We can never guarantee that the data are not consistent with the null, but we can make a statement about the probability of the observed result arising by chance (called *a. p-value).*



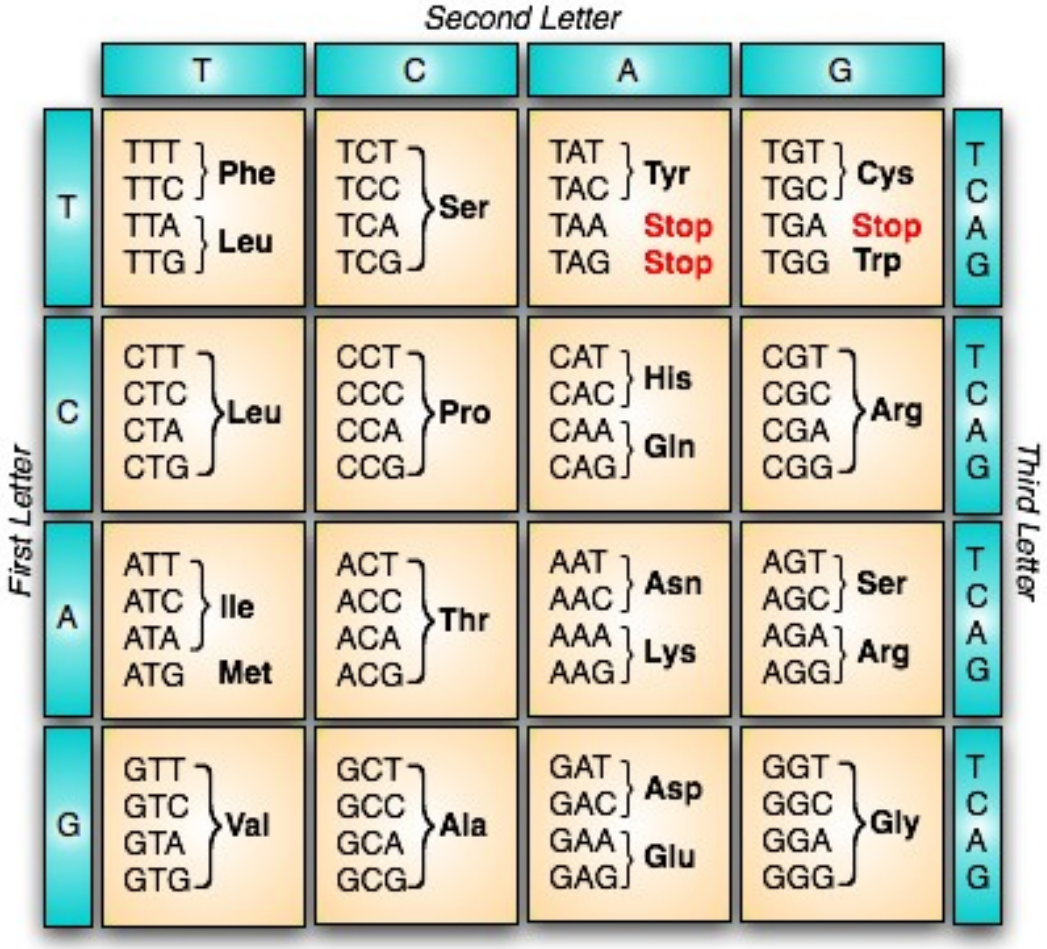
COMPUTING A P-VALUE FOR ORFS.

- Imagine a random process generating a sequence of DNA, and let us ask the probability of this process generating an ORF by chance.

- **If all nucleotides are emitted with the same probability we can easily estimate the probability of observing a stop codon in a given reading frame.**

- We are really computing the probability of seeing a stop codon, conditioned on seeing a start codon in the same frame, by chance.

THE GENETIC CODE



COMPUTING A P-VALUE FOR ORFS.

- **What is the probability of picking one of the stop codons?**
- It is the sum of their probabilities: if the distribution of codons is uniform, it will be 3/64, versus a 61/64 probability of picking a non-stop codon (since there are 64 possible codons, three of which are stop codons).
- So the probability of a run of k non-stop codons following a start codon is

$$P(\text{run of } k \textit{ non-stop codons}) = (61/64)^k.$$

- Setting $\alpha=0.05$, we can easily estimate the minimum acceptable ORF length. Since

$$(61/64)^{62} = 0.051$$

- discarding all ORFs of length $k < 62$, we will remove 95% of the spurious ORFs.

AB INITIO-BASE PROKARYOTES

- The *ab initio*-based approach predicts genes **based on the given sequence alone.**
- It does so by relying on **two major features** associated with genes.

- The **first** is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites, and polyadenylation (poly-A) sites.
- In addition, the triplet codon structure limits the coding frame length to multiples of three, which can be used as a condition for gene prediction.

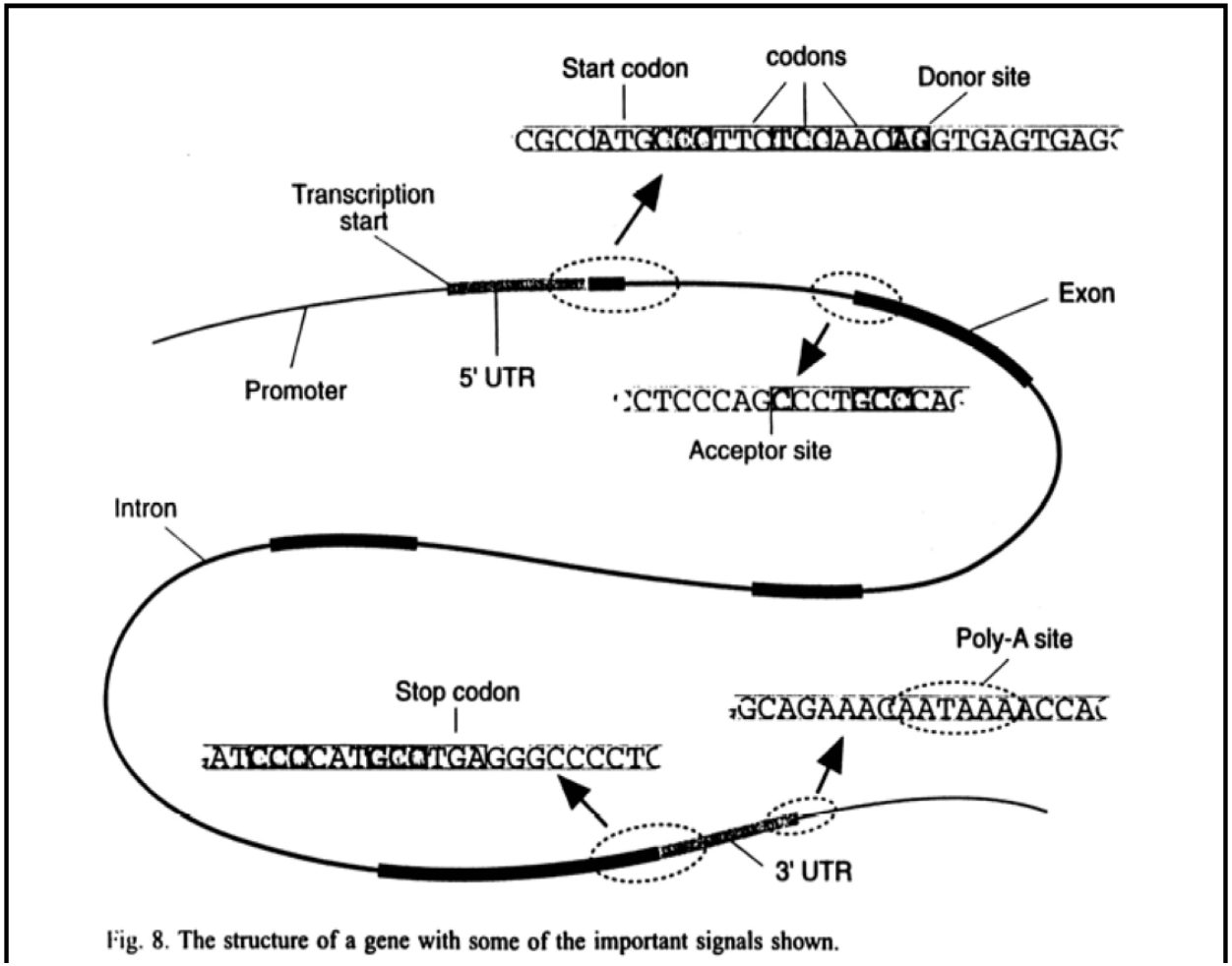
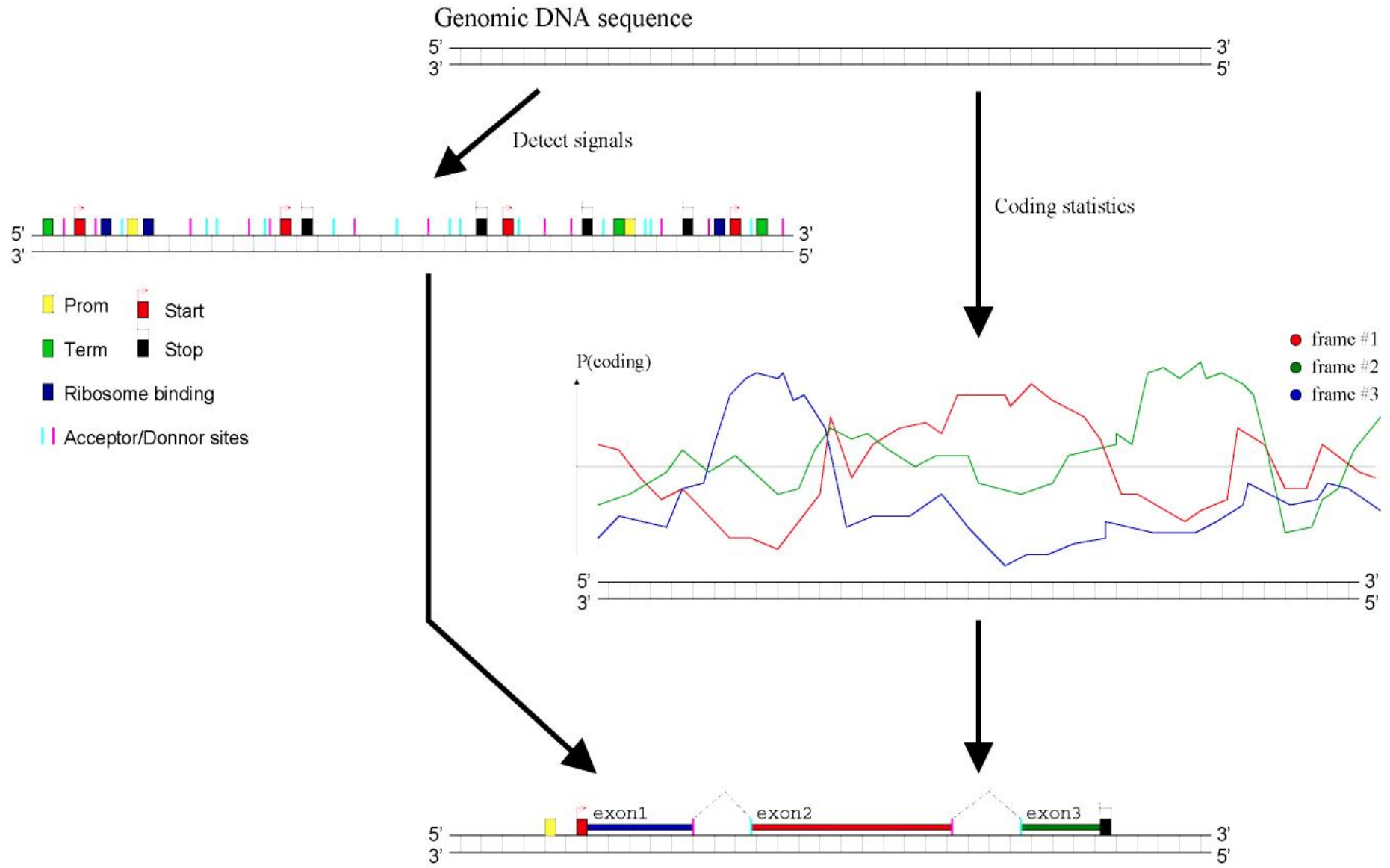


Fig. 8. The structure of a gene with some of the important signals shown.

AB INITIO-BASED

- The **second feature** used by ab initio algorithms is gene content, which is statistical description of coding regions.
- It has been observed that nucleotide composition and statistical patterns of the coding regions tend to vary significantly from those of the noncoding regions. The unique features can be detected by **employing probabilistic models** such as Markov models or hidden Markov models to help distinguish coding from noncoding regions.


Genomic DNA sequence
5'
3'
3'
5'
Detect signals
Coding statistics
5'
3'
3'
5'
Prom
Start
Term
Stop
Ribosome binding
Acceptor/Donnor sites
P(coding)
frame #1
frame #2
frame #3
5'
3'
3'
5'
exon1
exon2
exon3
5'
3'
3'
5'

AB INITIO-BASED

- Inter-genic regions, introns, and exons have differer nucleotide contents
- Example: observed stop codons (TAG,TAA,TGA)
  - assuming an uniform random distribution, we expect stop codons every 64/3 codons ($\sim$ 21 codons) in average
  - in coding regions the occurrence of stop codons decreases
  - ... but, this measure is sensitive to frame shift errors and can't detect short coding regions

AB INITIO BASED

- Dimer frequencies observed in proteins in Shewanel (avg ~ 5%):

	ala	asn	arg	asp	cys	glu	gly	glu	his	ile	leu	met	lys	phe	pro	ser	trp	thr	tyr	val	
ala	9.5	4.3	4.1	5.3	1.2	6.0	6.5	4.8	2.0	6.5	11.5	2.6	6.0	3.7	3.5	6.2	1.1	5.0	2.7	6.5	= 100
arg	7.9	3.9	5.5	5.3	1.1	6.0	5.9	5.5	2.6	6.5	11.4	2.2	5.0	4.7	3.6	5.5	1.4	4.4	4.0	6.6	= 100
asn	9.6	4.2	4.9	4.9	1.0	5.3	7.4	5.6	2.3	6.0	10.0	2.0	4.9	3.5	5.1	6.1	1.5	5.5	3.1	6.1	= 100
asp	9.3	4.7	4.0	5.1	1.0	6.7	7.0	2.9	1.8	7.1	9.6	2.3	6.3	4.3	3.9	5.9	1.6	5.1	3.6	6.6	= 100
cys	8.4	3.3	4.8	5.4	1.7	5.6	8.1	5.2	4.3	5.4	10.2	1.8	3.8	4.1	4.5	6.3	1.6	4.3	3.4	6.8	= 100
glu	9.4	3.6	5.8	4.5	0.8	4.9	5.8	7.0	2.6	5.9	12.7	2.4	5.0	4.0	3.5	5.4	1.1	5.0	2.8	6.8	= 100
gln	10.3	3.0	4.9	4.4	0.9	4.5	7.0	6.8	2.7	5.5	12.8	2.0	4.1	3.9	3.8	5.8	1.4	5.3	3.0	6.9	= 100
gly	8.1	3.9	4.8	5.1	1.2	6.0	6.4	4.6	2.4	6.8	10.5	2.7	5.8	4.8	2.4	5.8	1.4	5.1	3.7	7.5	= 100
his	7.3	4.0	4.7	4.8	1.5	4.9	6.9	5.6	3.0	6.2	10.8	1.6	4.8	5.0	5.2	6.8	1.7	4.9	4.2	5.1	= 100
ile	11.0	4.9	4.7	6.5	1.1	6.9	7.2	3.6	2.1	5.3	8.6	1.8	5.3	3.2	4.2	7.0	0.9	5.6	2.9	6.1	= 100
leu	10.4	4.3	4.2	5.2	1.1	5.2	6.8	3.7	2.0	5.6	10.6	2.3	5.3	3.8	4.5	7.4	1.0	6.2	2.6	6.6	= 100
lys	10.6	3.8	5.2	5.2	0.5	5.3	6.6	5.9	2.6	5.2	11.3	1.9	4.7	2.8	4.6	6.0	1.2	5.5	2.6	7.6	= 100
met	10.8	3.8	4.8	4.6	0.7	4.6	7.0	4.9	1.7	4.7	11.4	2.8	5.2	3.3	5.1	7.4	0.9	6.3	2.0	6.8	= 100
phe	9.6	5.2	3.7	6.5	1.2	6.4	7.9	2.7	1.9	6.7	7.4	2.5	5.0	3.9	3.6	8.0	1.3	5.8	3.3	6.3	= 100
pro	8.4	4.6	3.6	5.4	0.7	7.6	5.4	5.2	2.3	6.1	11.2	2.4	5.5	4.2	2.8	6.5	1.4	5.4	2.9	7.5	= 100
ser	9.1	3.7	4.6	5.0	1.0	5.4	7.2	5.2	2.6	6.0	11.6	2.2	4.5	4.1	4.1	6.5	1.2	5.0	3.2	6.8	= 100
thr	9.1	3.7	4.2	5.6	0.9	5.7	7.5	5.7	2.2	5.5	12.0	2.0	4.2	3.5	5.5	6.2	1.1	5.3	2.6	6.7	= 100
trp	7.1	3.2	6.3	4.8	1.3	3.9	6.6	8.5	3.6	5.0	14.2	2.4	3.2	4.6	3.9	5.8	1.3	4.3	3.0	6.1	= 100
tyr	7.9	3.6	6.5	4.9	1.2	4.5	7.1	7.0	2.6	5.0	11.7	1.6	4.0	4.7	4.9	6.4	1.5	4.6	3.4	5.7	= 100
val	9.6	4.4	4.1	5.9	1.0	6.2	6.4	3.4	1.8	6.5	10.2	2.5	5.2	3.7	3.8	7.2	1.1	6.1	2.7	7.1	= 100

AB INITIO-BASED

- A bias is observed for dimers in proteins
- The dimer bias is reflected in dicodons: coding and non-coding regions have different dicodons bias.
- The bias in the observed dicodon (hexamer) frequencies can be used to predict coding regions in genomic sequences.
- Most of the *ab initio* methods for gene structure prediction use this information!

AB INITIO-BASED

Let $f^{c}_{abc,a'b'c'}$ denote the observed frequency for dicodon $abc, a'b'c'$ in a set of known coding regions, and let $f^{n}_{abc,a'b'c'}$ denote the observed frequency for the same dicodon in non-coding regions.

The score of dicodon $abc, a'b'c'$ in being coding is defined as:

$$P(abc, a'b'c') = log(\frac{f^{c}_{abc,a'b'c'}}{f^{n}_{abc,a'b'c'}})$$

AB INITIO-BASED

- Properties of $P(abc, a'b'c')$:
  - if $P(abc, a'b'c') = 0$: dicodon $abc, a'b'c'$ has the same frequencies in coding and non-coding regions
  - if $P(abc, a'b'c') > 0$: dicodon $abc, a'b'c'$ is observed more frequently in coding regions
  - if $P(abc, a'b'c') < 0$: dicodon $abc, a'b'c'$ is observed more frequently in non-coding regions

AB INITIO-BASED

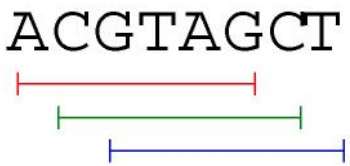
- Assume $S = a_1b_1c_1, a_2b_2c_2, ..., a_{n+1}b_{n+1}c_{n+1}$ is a genimic coding region with unknown reading frame.
- We can calculate the score of each frame of being coding:

$$P_1 = P(a_1b_1c_1, a_2b_2c_2) + P(a_3b_3c_3, a_4b_4c_4) + ... + P(a_{n-1}b_{n-1}c_{n-1}, a_nb_nc_n)$$

$$P_2 = P(b_1c_1a_2, b_2c_2a_3) + P(b_3c_3a_4, b_4c_4a_5) + ... + P(b_{n-1}c_{n-1}a_n, b_nc_na_{n+1})$$

$$P_3 = P(c_1a_2b_2, c_2a_3b_3) + P(c_3a_4b_4, c_4a_5b_5) + ... + P(c_{n-1}a_nb_n, c_na_{n+1}, b_{n+1})$$

AB INITIO-BASED

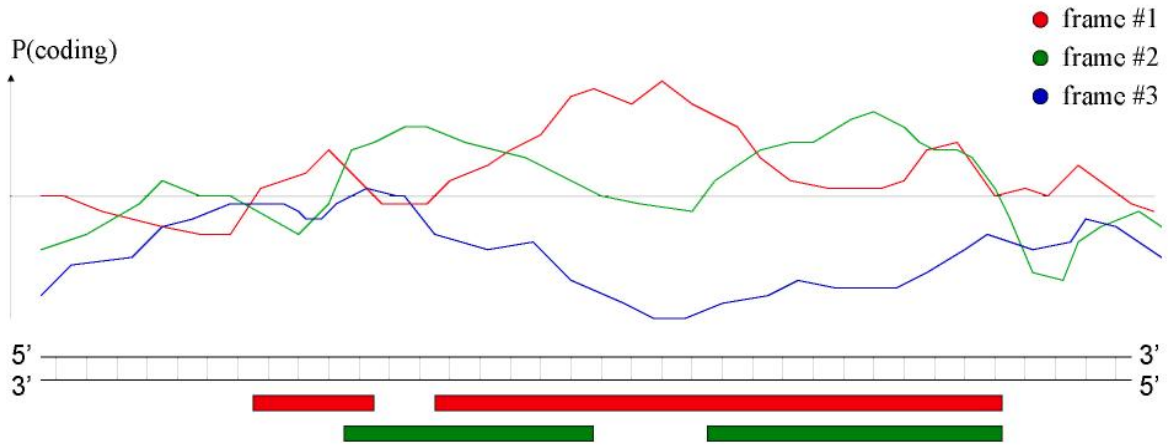


- f^{c}(ACG,TAG) = 0.000, f^{n}(ACG,TAG) = 0.062
- f^{c}(CGT,AGC) = 0.068, f^{n}(CGT,AGC) = 0.019
- f^{c}(GTA,GCT) = 0.021, f^{n}(GTA,GCT) = 0.026

- P(ACG,TAG) = - infinity (special case STOP codon)
- P(CGT,AGC) = log(0.068 / 0.019) = 1.3
- P(GTA,GCT) = log(0.021 / 0.026) = -0.2

AB INITIO-BASED

- Procedure for predicting coding regions using coding statistics:
  - find all ORFs of the sequence (start/stop regions)
  - slide through the ORFs with a window of 60bp and find good scoring regions



PERFORMANCE EVALUATION

- The accuracy of a prediction program can be evaluated using parameters such as **sensitivity** and **specificity**.

- To describe the concept of sensitivity and specificity accurately, four features are used:
  - true positive (TP), which is a correctly predicted feature;
  - false positive (FP), which is an incorrectly predicted feature;
  - false negative (FN), which is a missed feature; and
  - true negative (TN), which is the correctly predicted absence of a feature.

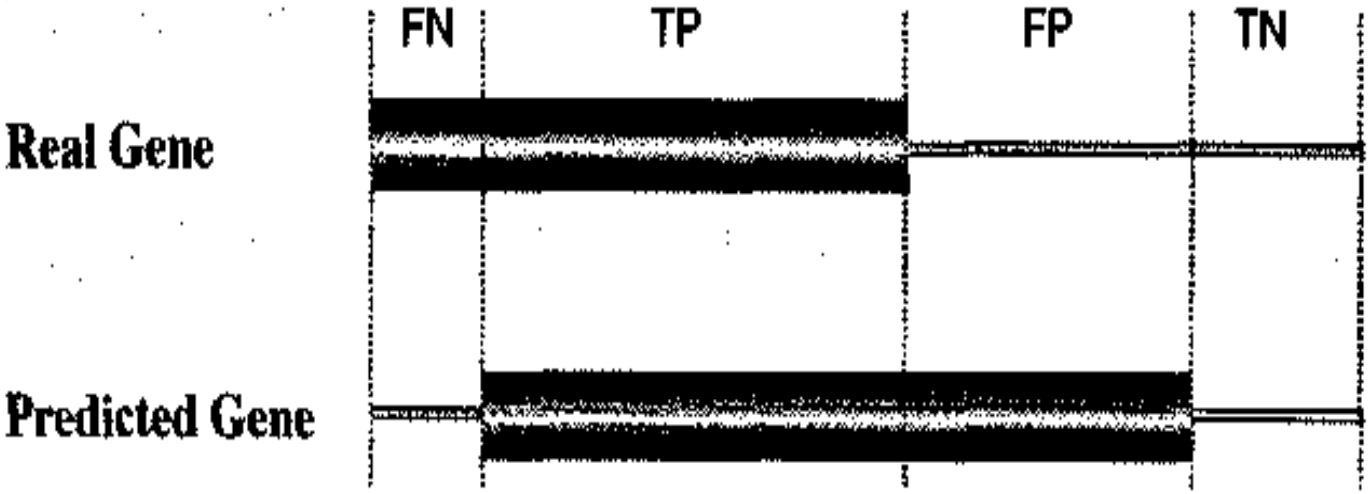


Figure 8.4: Definition of four basic measures of gene prediction accuracy at the nucleotide level. *Abbreviations:* FN, false negative; TP, true positive; FP, false positive; TN, true negative.

PERFORMANCE EVALUATION

- Using these four terms, sensitivity (Sn) and specificity (Sp) can be described by the following formulas:

$$Sn = TP/(TP + FN)$$

$$Sp = TP/(TP + FP)$$

- According to these formulas, ***sensitivity*** is the proportion of true signals predicted among all possible true signals. It can be considered as the ability to include correct predictions.

- In contrast, ***specificity*** is the proportion of true signals among all signals that are predicted. It represents the ability to exclude incorrect predictions.

PERFORMANCE EVALUATION

- Because neither sensitivity nor specificity alone can fully describe accuracy, it is desirable to use a single value to summarize both of them. In the field of gene finding, a single parameter known as the correlation coefficient (CC) is often used, which is defined by the following formula:

$$CC = \frac{TP \bullet TN - FP \bullet FN}{\sqrt{(TP + FP)(TN + FN)(FP + TN)}}$$

PERFORMANCE EVALUATION

- The value of the CC provides an overall measure of accuracy, which ranges from -1 to +1, with +1 meaning always correct prediction and —1 meaning always incorrect prediction.

TABLE 8.1. Performance Analysis of the Glimmer Program for Gene Prediction of Three Genomes

Species	GC (%)	FN	FP	Sensitivity	Specificity
Campylobacter jejuni	30.5	10	19	99.3	98.7
Haemophilus influenzae	38.2	3	54	99.8	96.1
Helicobacter pylori	38.9	6	39	99.5	97.2

Note: The data sets were from three bacterial genomes (Aggarwal and Ramaswamy, 2002).
Abbreviations: FN, false negative; FP, false positive.

GENE PREDICTION IN EUKARYOTES

- Eukaryotic nuclear genomes are much larger than prokaryotic ones, with sizes ranging from 10 Mbp to 670 Gbp (1 Gbp = 10^9 bp).

- They tend to have a very low gene density. In humans, for instance, only 3% of the genome codes for genes, with about 1 gene per 100 kbp on average. The space between genes is often very large and rich in repetitive sequences and transposable elements.

GENE PREDICTION PROGRAMS

- To date, numerous computer programs have been developed for identifying eukaryotic genes.

- They fall into all three categories of algorithms: **ab initio based, homology based, and consensus based.**

- **Most of these programs are organism specific** because training data sets for obtaining statistical parameters have to be derived from individual organisms.

- Some of the algorithms are able to predict the most probable exons as well as suboptimal exons providing information for possible alternative spliced transcription products.

GENE PREDICTION IN EUKARYOTES

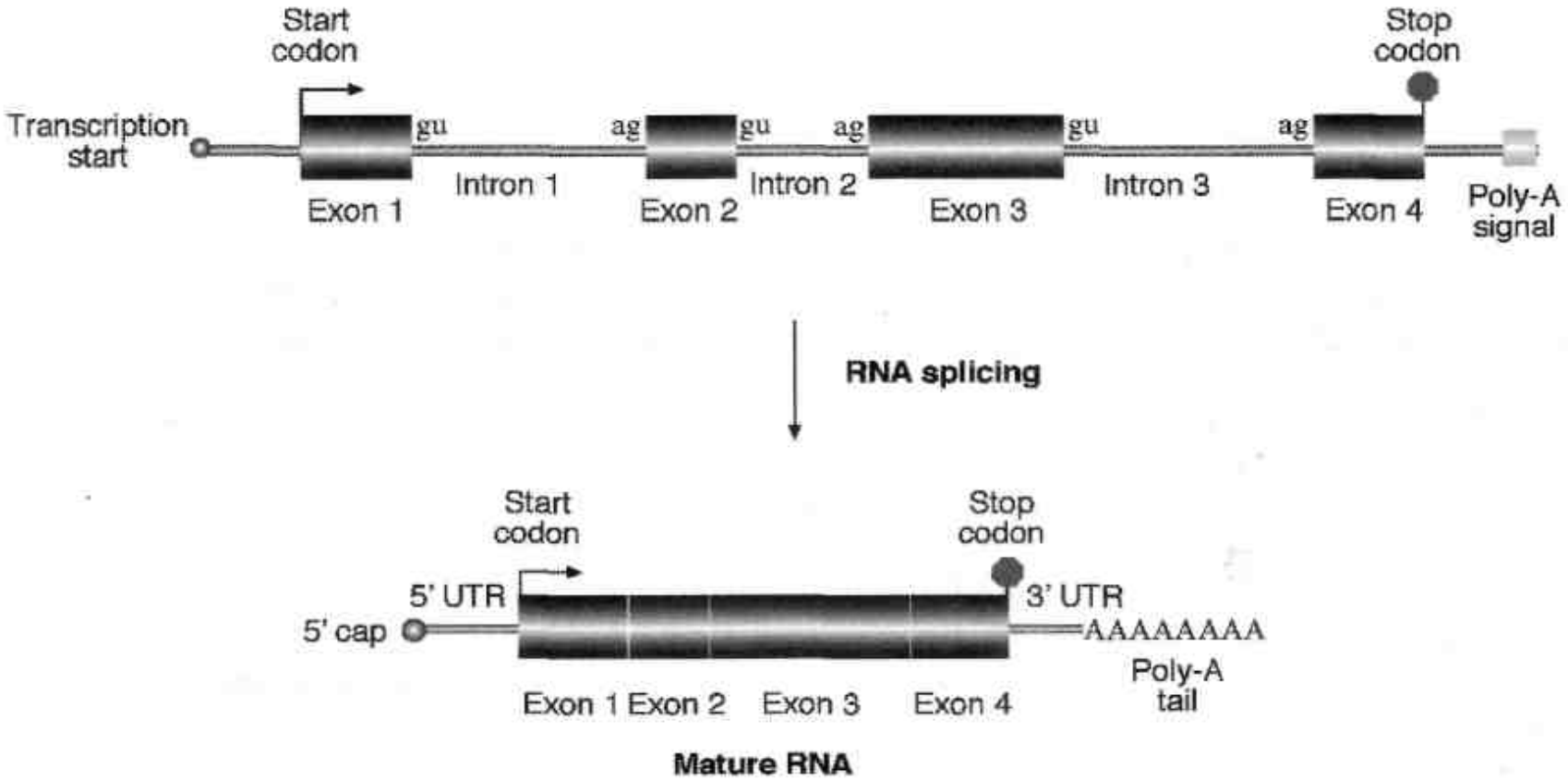


Figure 8.5: *Structure of a typical eukaryotic RNA as primary transcript from genomic DNA and as* mature RNA after posttranscriptional processing. *Abbreviations:* UTR, untranslated region; poly-A, polyadenylation.

GENE PREDICTION IN EUKARYOTES

- The nascent transcript from a eukaryotic gene is **modified in three different ways** before becoming a mature mRNA for protein translation.
- **The first event** is capping at the 5' end of the transcript, which involves methylation at the initial residue of the RNA.
- **The second event** is splicing, which is the process of removing introns and joining exons. The molecular basis of splicing is still not completely understood. What is known currently is that the splicing process involves a large RNA-protein complex called spliceosome.
- **The third** modification is polyadenylation, which is the addition of a stretch of As (~250) at the 3' end of the RNA.

GENE PREDICTION IN EUKARYOTES

- The **main issue** in prediction of eukaryotic genes is the identification of exons, introns, and splicing sites.

- From a computational point of view, it is a very complex and challenging problem.

- Because of the presence of split gene structures, alternative splicing, and very low gene densities, the difficulty of finding genes in such an environment is likened to finding a needle in a haystack. The needle to be found actually is broken into pieces and scattered in many different places. The job is to gather the pieces in the haystack and reproduce the needle in the correct order.

GENE PREDICTION IN EUKARYOTES

- The good news is that there are still some conserved sequence features in eukaryotic genes that allow computational prediction.

- For example, the splice junctions of introns and exons follow the GT-AG rule in which an intron at the 5' splice junction has a consensus motif of GTAAGT; and at the 3' splice junction is a consensus motif of (Py)12NCAG

- Distribution observed for donor sites in human (GT):

	-3	-2	-1	1	2	3	4	5	6
a	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
c	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
g	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
t	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

- Distribution observed for acceptor sites in human (AG):

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
a	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
c	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
g	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
t	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9

- Information content:

$$I_j = |\sum_i -f(i,j) * log(f(i,j)/q(i))|$$

where $i = \{a, c, g, t\}$, j is the position (column), $f(i,j)$ is the observed frequency for symbol i at position j, and $q(i)$ is the distribution of symbol i (in our case $q(i) = 0.25$).

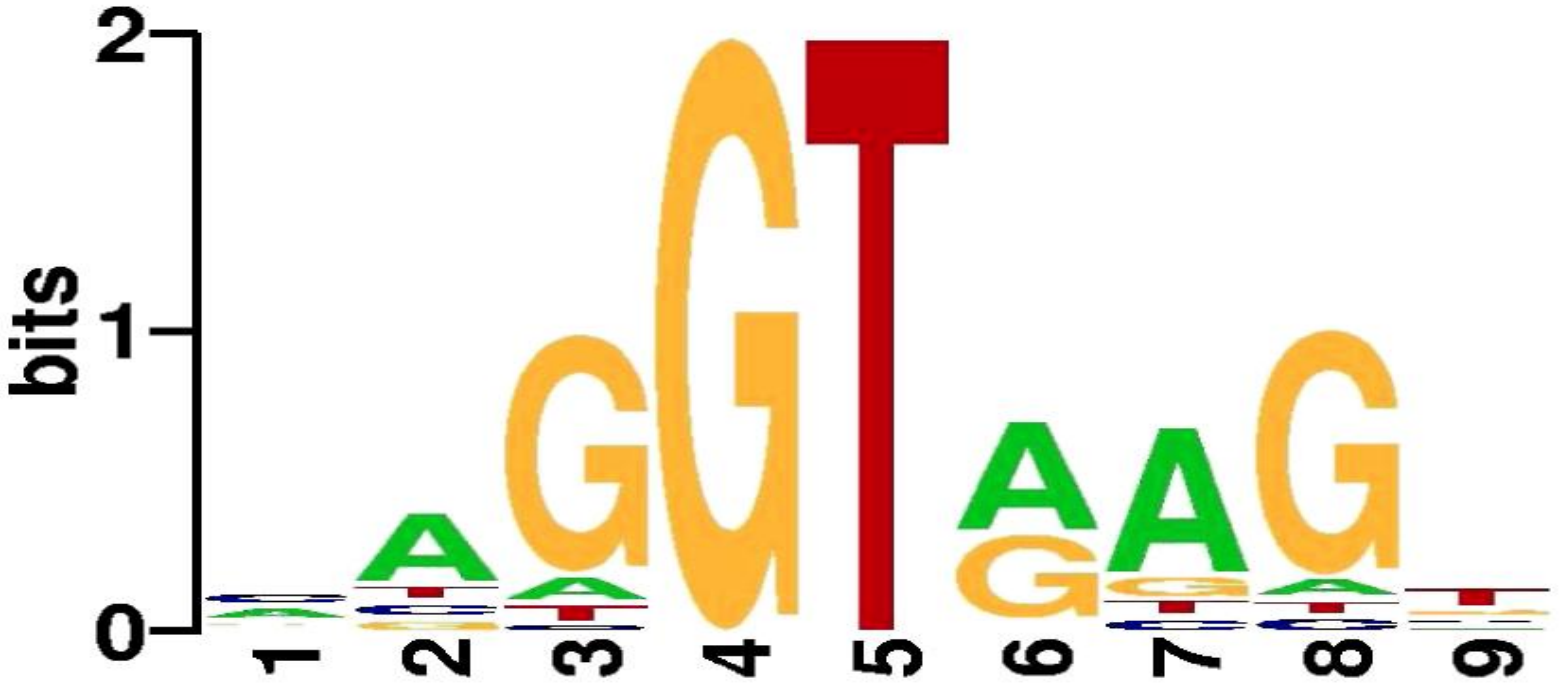
- A column with uniform distributed nucleotides has a low information content
- A column with unevenly distributed nucleotides has a higher information content

	-3	-2	-1	1	2	3	4	5	6
a	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
c	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
g	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
t	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

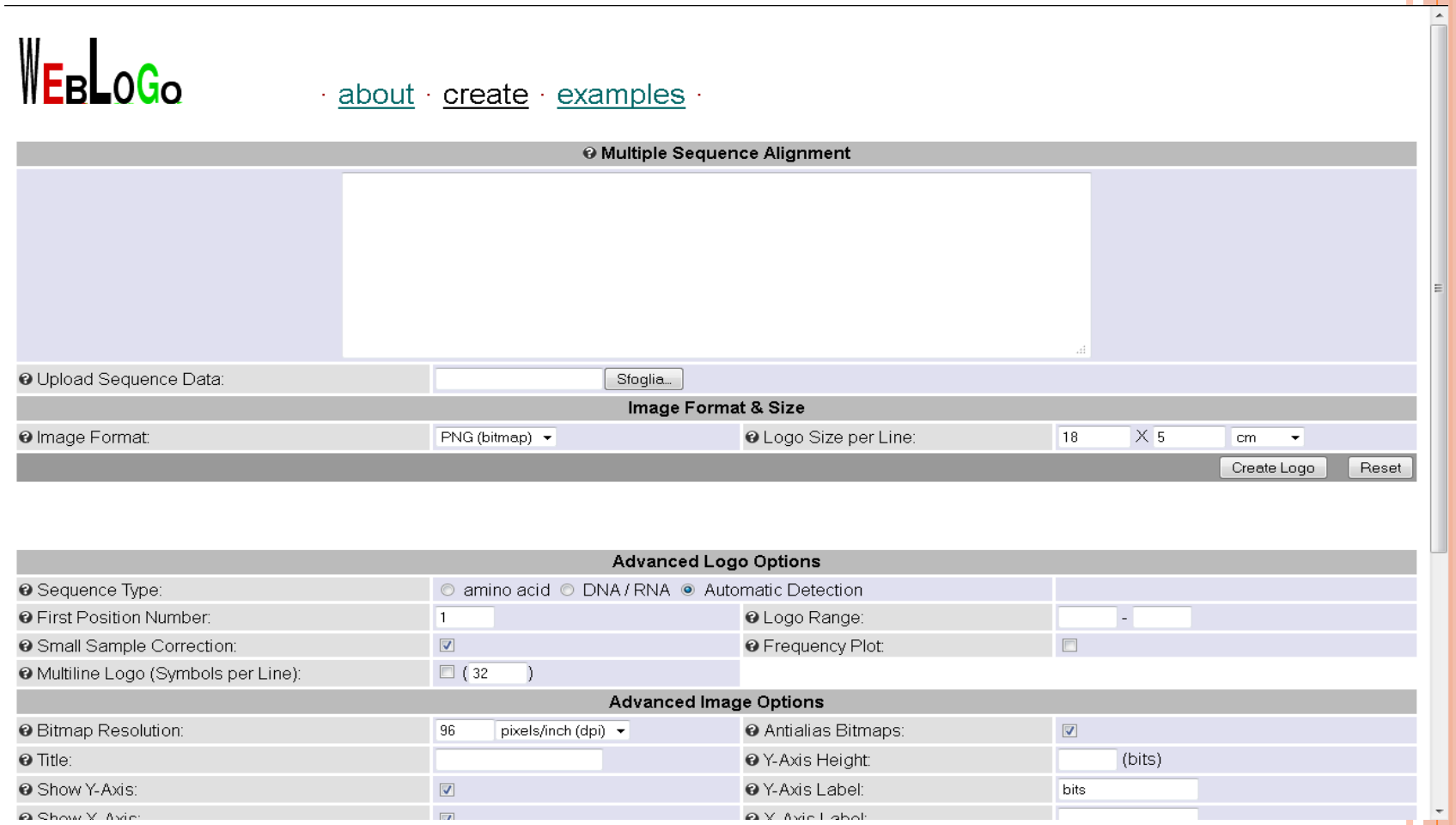
$$
\begin{aligned}
I_{-3} &= |\sum_i -f(i,-3) * log_2(f(i,-3)/0.25)| \\
&= |-0.34 * log_2(0.34/0.25) \\
&\quad -0.363 * log_2(0.363/0.25) \\
&\quad -0.183 * log_2(0.183/0.25) \\
&\quad -0.114 * log_2(0.114/0.25)| \\
&= 0.13
\end{aligned}
$$

$$I_{+1} = |\sum_i -f(i,+1) * log_2(f(i,+1)/0.25)| = 2$$

	-3	-2	-1	1	2	3	4	5	6
a	34.0	60.4	9.2	0.0	0.0	52.6	71.3	7.1	16.0
c	36.3	12.9	3.3	0.0	0.0	2.8	7.6	5.5	16.5
g	18.3	12.5	80.3	100	0.0	41.9	11.8	81.4	20.9
t	11.4	14.2	7.3	0.0	100	2.5	9.3	5.9	46.2

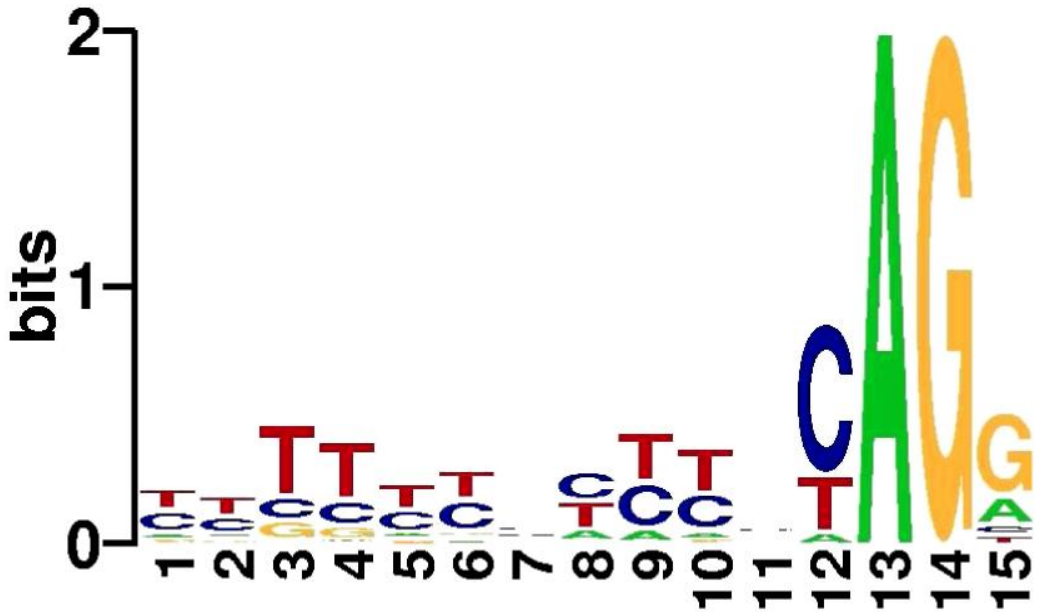


LOGO
HTTP://WEBLOGO.BERKELEY.EDU/



GENE PREDICTION IN EUKARYOTES

	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1
a	11.1	12.7	3.2	4.8	12.7	8.7	16.7	16.7	12.7	9.5	26.2	6.3	100	0.0	21.4
c	36.5	30.9	19.1	23.0	34.9	39.7	34.9	40.5	40.5	36.5	33.3	68.2	0.0	0.0	7.9
g	9.5	10.3	15.1	12.7	8.7	9.5	16.7	4.8	2.4	6.3	13.5	0.0	0.0	100	62.7
t	38.9	41.3	58.7	55.6	42.1	40.5	30.9	37.3	44.4	47.6	27.0	25.4	0.0	0.0	7.9



GENE PREDICTION IN EUKARYOTES

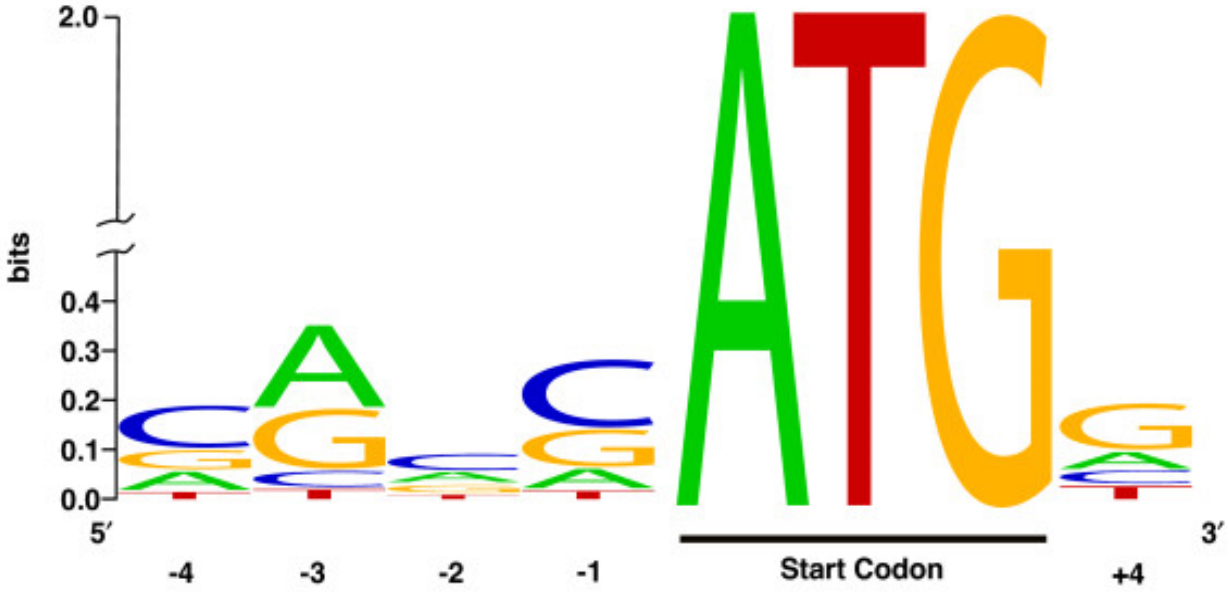
- Some statistical patterns useful for prokaryotic gene finding can be applied to eukaryotic systems as well.

- For example, nucleotide compositions and codon bias in coding regions of eukaryotes are different from those of the non-coding regions.

- Hexamer frequencies in coding regions are also higher than in the noncoding regions.

GENE PREDICTION IN EUKARYOTES

- Most vertebrate genes use ATG as the translation start codon and have a uniquely conserved flanking sequence call a *Kozak sequence* (CCGCCATGG).

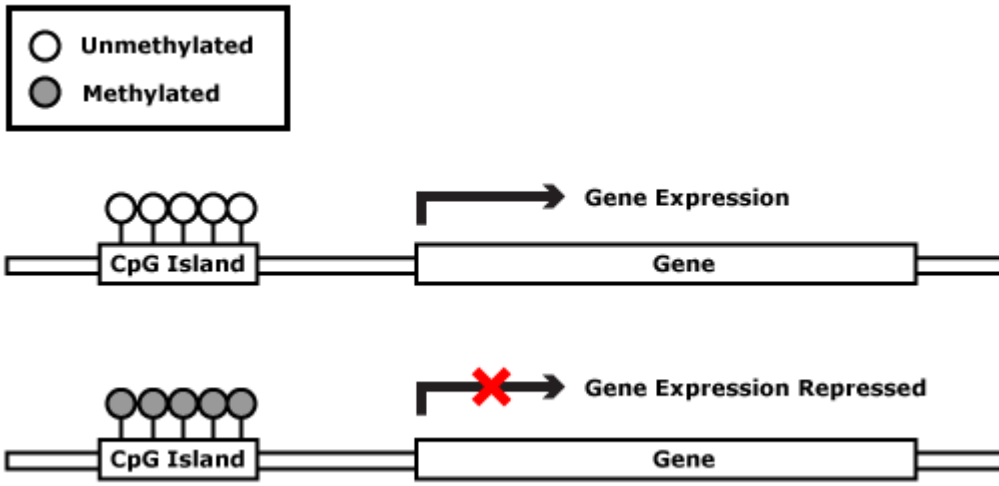


GENE PREDICTION IN EUKARYOTES

- In addition, most of these genes have a high density of CG dinucleotides near the transcription start site. This region is referred to as a CpG island, which helps to identify the transcription initiation site of a eukaryotic gene

Exon 1 CpG Island: 12634..12767

```
11941 ttataagatc ccctccctc taaatcctgt ccttctatca cttcatcctt CGctctcctt
12001 taaaatgaga cagttgtcag caggaatcct gCGcaagaac acaccaccct gtttcataga
12061 agatatctca ggtaatgtgc aaacaCGggt ttttaaaCGg agCGcatttt tctcatttgt
12121 taatatcacc acctaaatca tctcttgcct aaaacaagga gtagaaagtg aatgaaggaa
12181 ggaacaggtg atggtcagtg tcctttctac gcctcaaaat ttaagagttt atgtgaaaat
12241 tcataaatat taatctcaat ccaggttaag caaaattttt tgctctcctc tttagaaatt
12301 tctggttgcc aaagttccag aaattgcttc ctcattcctg agcctttcat tttctCGatt
12361 tctccattat gtaaCGggga gctggagctt tgggcCGaat ttccaattaa agatgatttt
12421 tacagtcaat gagccaCGtc agggagCGat ggcaccCGca ggCGgtatca actgatgcaa
12481 gtgttcaagc gaatctcaac tCGttttttc CGgtgactca ttccCGgccc tgcttggcag
12541 CGctgcaccc tttaacttaa acctCGgcCG gcCGccCGcc gggggcacag agtgtgCGcc
*12601 gggcCGCGCG gcaattggtc ccCGCGcCGa cctcCGccCG CGagCGcCGc CGcttccctt
*12661 cccCGcccCG CGtcctccc cctCGgcccc gCGCGtCGcc tgtcctcCGa gccagtCGct
*12721 gacagcCGCG gCGcCGCGag cttctcctct cctcaCGacc gaggcaggta aaCGccCGgg
12781 gtgggaggaa CGCGggCGgg ggcaggggag cCGCGggggc CGagtgagga cccCGggcct
12841 CGggtcccag gCGcaagggt gccCGgcCGg gCGgggtCGg gaccccagtg aggaggggcc
12901 gggggctgcc cCGCGggCGc gtgaCGgtct CGggcctgcc CGgctgCGct ggtctcCGct
12961 CGggtgaggc ggcttggctt CGcttttcag gttaggaaag ctcctttac tgCGCGttgg
13021 ggggctgggg gagctggCGg agccaCGtta gggaggtCGg tggCGcCGgg gtgtctcagc
13081 gccccctgca cccCGCGCGg gtcCGgccca gCGggCGatc gctggCGccc agggaactcc
13141 gggagggcCG ccagCGggct cCGcaggCGc ggggCGggga ggggCGcctg ggggcCGCGg
13201 ggctCGCGct cccCGccCGt tggcCGcccc tCGgaggcCG agatCGgggc ccagaaCGcc
13261 ccttggcaaa gcctggCGct tcCGCGatgc ccagagggtg cttgggggga tggagagagg
13321 ggCGccCGcc ggggtagttc CGggagcctc ggtgcctccc gcCGcagctg cagCGttcct
13381 ccCGggaggc ggcccagccc ttcatcctCG cCGcctgagc ttctcCGagg ggggctgcag
13441 ccttgCGgcc gttgccacCG cctggagaag CGgcccaCGc ggactgaCGg gCGgggCGg
13501 ggcctCGggc ctCGgCGggg gCGgggtcCG gggaggcccc accctctgtt ctccaggggc
13561 ggggagagag gagctgcagg tctgCGgcct ggccccaggt gCGatggCGg accccagctt
13621 ggccagtcac attcctccca gtccccctgg agggagaaCG ctggccatgg ggggctccaa
13681 ggaacaacca gcctCGgatg aCGacccttg ggtcacCGgt ctccccacct gtgCGgcagg
13741 CGccttcaCG tttcattatt aaacaatggg gagaaatcca tgtttactgt ccttttagg
13801 aatttttgc tcttctcttt gaggtggctg taggaaatag attttttttt taacctCGca
13861 attccaccac ggtcacatcc atcctCGcca tCGcagagcc acagctctcc gtttttgttt
13921 cctagcctcc agattctcac acaacacagt gcagtttcac tgctgtaatg atgaggatct
13981 tcatggcCGc gttatttttct tgttctgaga gcatcaCGgt ttaattagca gttccccata
14041 tgatttgaag tgtttccCGt ttccttaggg aaaactcctg gtagaatagg attaaggatt
14101 tttacaaata taattatcaa aaacatagga acagggaatt ggataaatat gttaaacttc
14161 tggaaaaatc aacaaCGctc ttagatttgt agaagaaagg aaaaaatcac cagtggaaag
14221 gagcaatttt acttacacaa acacagagaa ggtcttacag tgaaaaaaag ctaaccagta
```



GENE PREDICTION IN EUKARYOTES

- ***Homology-Based Programs***

- Homology-based programs are based on the fact that exon structures and exon sequences of related species are highly conserved.

- When potential coding frames in a query sequence are translated and used to align with closest protein homologs found in databases, near perfectly matched regions can be used to reveal the exon boundaries in the query. This approach assumes that the database sequences are correct.

GENE PREDICTION IN EUKARYOTES

- It is a reasonable assumption in light of the fact that many homologous or paralogous sequences to be compared with are derived from cDNA massive sequencing of the same species.

- **With the support of experimental evidence, this method becomes rather efficient in finding genes in an unknown genomic DNA.**

- The drawback of this approach is its reliance on the presence of homologous in databases. **If the homologous are not available in the database, the method cannot be used.**

GENE PREDICTION IN EUKARYOTES

TABLE 8.2. Accuracy Comparisons for a Number of Ab Initio Gene Prediction Programs at Nucleotide and Exon Levels

	Nucleotide level			Exon level				
	Sn	Sp	CC	Sn	Sp	(Sn + Sp)/2	ME	WE
FGENES	0.86	0.88	0.83	0.67	0.67	0.67	0.12	0.09
GeneMark	0.87	0.89	0.83	0.53	0.54	0.54	0.13	0.11
Genie	0.91	0.90	0.88	0.71	0.70	0.71	0.19	0.11
GenScan	0.95	0.90	0.91	0.70	0.70	0.70	0.08	0.09
HMMgene	0.93	0.93	0.91	0.76	0.77	0.76	0.12	0.07
Morgan	0.75	0.74	0.74	0.46	0.41	0.43	0.20	0.28
MZEF	0.70	0.73	0.66	0.58	0.59	0.59	0.32	0.23

Note: The data sets used were single mammalian gene sequences (performed by Sanja Rogic, from www.cs.ubc.ca/~rogic/evaluation/tablesgen.html.

Abbreviations: Sn, sensitivity; Sp, specificity; CC, correlation coefficient; ME, missed exons; WE, wrongly predicted exons.