A reading frame is one of the 6 ways to translate DNA sequence into codons (three reading frames on each of the two strands).

An open reading frame (ORF) is a sequence that contains no stop codon in at least one of the six reading frames.

Coding sequences are a subset of open reading frames.

ORF’s beginning with ATG comprise a special class. In organisms with uninterrupted genes (i.e. no introns) coding sequences are a subset of ORF’s beginning with ATG. In organisms that have introns, the first exon is an ORF beginning with ATG.

The exact distribution of ORF’s in random sequence is difficult to obtain. However it is easy to get a good approximation. In doing so, we make the simplifying assumption that an ORF begins and ends with a stop codon in some reading frame, and has no interior stop codons in that reading frame.

Define $N(c; L_c)$ to be the expected number of ORF’s of length at least $c$ codons (at least $3c$ nucleotides) in a given reading frame of random sequence of length $L_c$ codons.

More formally, if $Z_i$ is the length of the ORF that begins at position $i$ (in codons), then

$$N(c; L_c) = E \sum_{i=1}^{L_c} \mathbb{I}(Z_i \geq c) = L_c E \mathbb{I}(Z_i \geq c).$$

Since

$$P(Z_i \geq c + 1) = P(Z_i \geq c + 1 | Z_i \geq c) P(Z_i \geq c),$$

1
if we assume that all codons are equally likely then

\[ E(Z_i \geq c + 1) = \frac{61}{64} E(Z_i \geq c), \]

and hence

\[ N(c + 1; L_c) = \frac{61}{64} N(c; L_c), \]

which is equivalent to

\[ N(c + 1) - N(c) = -\frac{3}{64} N(c). \]

If \( c \) were continuous, then the left side approximates the derivative \( dN(c)/dc \), so we have an approximation to the ordinary differential equation

\[ dN(c)/dc = -\frac{3}{64} N(c), \]

which has solutions of the form

\[ N(c) = N(0) \exp(-3c/64). \]

Since \( N(0) \) is just the frequency of a stop codon, \( N(0) = 3L_c/64 \). Now if we ignore dependence between reading frames and convert from codons to nucleotides, we get the following approximation to the expected number of ORF’s of length at least \( \ell \) nucleotides in one of the three reading frames on a single strand:

\[ N(\ell) = \frac{3L}{64} \exp(-\ell/64), \]

where \( L = 3L_c \) and \( \ell = 3c \).
• It follows that in sequence of length $L = 64 \exp(\ell/64)/3$ one ORF of length at least $\ell$ is expected.

• An alternative approach is to directly randomize an observed sequence (i.e. permute it with a random permutation). Using the randomized sequence, for each $\ell = 1, 2, \ldots$ compute the number of ORF’s $N_\ell$ of length exceeding $\ell$ codons. Then fit an empirical model of the form $N_\ell = \alpha \exp(-\beta \ell)$, or $\log N_\ell = a - \beta \ell$ (where $a = \log \alpha$). This model can be fit using ordinary least squares:

$$
\begin{align*}
\beta &= -\frac{\sum_\ell (\ell - \bar{\ell})(\log N_\ell - \log N_\ell)}{\sum_\ell (\ell - \bar{\ell})^2} \\
 a &= \log N_\ell + \beta \bar{\ell}.
\end{align*}
$$

An advantage of this empirical approach over the theoretical approach derived above is that we can incorporate additional constraints such that the ORF must begin with ATG, or that the sequence has a specific GC content.

• **Sensitivity:** The fraction of true genes (exons) that are detected by a procedure.

• **Specificity:** The fraction of the candidate genes (exons) produced by a procedure that are true.

• Sensitivity and specificity are always traded off against each other.

• Any procedure for finding genes has a tuning constant that increases sensitivity and decreases specificity (or vice versa). A natural balancing point is to choose the tuning constant so that sensitivity and specificity are equal.
• If genes are to be identified based on ORF length, the tuning constant is the minimal length for an ORF to be considered candidate coding sequence.

• In organisms that lack introns, coding sequences begin with ATG, and there tend to be far more ORF’s beginning with ATG than expected by chance. For example, in yeast out of 6000 predicted genes (based on ORF information alone) only around 500 are spurious.

• In vertebrates, the observed ORF distribution is consistent with the randomized distribution. Most long ORF’s do not correspond to coding sequence, so additional signals must be sought for identifying coding sequence.

• In vertebrates, exons must be flanked by splice donor (3’) and acceptor (5’) sites. These provide additional information that can be used to narrow down the pool of long ORF’s, increasing specificity without harming sensitivity.

• The pattern of the splice sites is not fixed. Its variation is described by a position weight matrix (PWM). This is a formal way of expressing a consensus sequence. For the acceptor position it is

$$
\begin{array}{cccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
A & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 & 0 \\
C & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 2 & 0 & 0 \\
G & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\
T & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\
\end{array}
$$

A candidate splice acceptor receives a score based on the PWM.
For example, CTGTTCACACAG would receive 28. Higher scores correspond to better matches to the consensus sequence.

- One must set a threshold PWM score such that an ORF is only considered to be a candidate exon if it is flanked by a sequence with score exceeding the threshold. One way to set the threshold is using a training set consisting of some known coding sequences and some sequences that are known to be non-coding. Based on the training set, the threshold can be set so that sensitivity = specificity (or some other value).

- An alternative is to set the threshold based on the theoretical distribution of PWM scores in random data. This distribution can be obtained using dynamic programming. Let $Z_k$ denote the PWM score for a random sequence of length $k$ (i.e., only using the first $k$ columns of the PWM), and let $X_k$ denote the PWM score for column $k$ alone. We obtain the following recursion:

$$P(Z_k = b) = \sum_{a \leq b} P(Z_k = b | Z_{k-1} = a)P(Z_{k-1} = a)$$

$$= \sum_{a \leq b} P(X_k = b - a)P(Z_{k-1} = a).$$

Thus the distribution of $Z_k$ can be built up from $k = 1, 2, \text{etc.}$

- For the splice acceptor site it is found that the sensitivity = specificity point occurs with a threshold of 29, and this gives only 50% of each.

- Similarly, one can derive PWM’s for regulatory elements such as TATA and CCAAT boxes. This provides some information about the location of the 5′ end of genes. However only around
1 in 84 good matches to the TATA PWM are actual TATA boxes (around 1 in 12 matches to CCAAT are actual regulatory elements). Thus a single regulatory element can never be used to find the 5′ end of a gene.

- Most promoters contain several regulatory elements. By searching for a combination of matches, sensitivity and specificity can be improved. However there are genes that have only a single element (or that contain none of the known elements).

- Other sources of information: (i) matches to EST libraries, (ii) matches to known protein sequence, (iii) near matches to known protein sequence in another organism, (iv) hexamer bias, (v) other (yet unknown) signals.

- Key messages: (i) the signals currently used to find coding sequences are weak, so it is not possible to identify most genes without also obtaining a large number of false positives, (ii) training-based methods “find more of the same”, so they are inherently conservative.

- Big picture question: the transcriptional machinery succeeds at recognizing genes, so the information is there. Two alternatives are (i) epigenetic information (information beyond what is contained in the sequence, i.e. methylation or chromatin structure) plays a major role in the innate mechanism by which genes are identified, (ii) there are as-yet undiscovered sequence signals that can be used to identify genes.

If (i) is true, then the goals that motivated large-scale sequencing may never be met.

**PWM’s and Multinomial Distributions**
There is a close connection between a PWM and a multinomial distribution. Suppose we observe \( N = 200 \) instances of the acceptor splice site. That is, we observe 200 sequences of length 12. We compute the nucleotide frequencies in each of the 12 positions, giving the following array:

\[
\begin{array}{ccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
A & p_1^A & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & p_{12}^A \\
C & p_1^C & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & p_{12}^C \\
G & p_1^G & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & p_{12}^G \\
T & p_1^T & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & p_{12}^T \\
\end{array}
\]

Similarly, we have a null distribution \( q^A, q^C, q^G, q^T \) that will usually be set according to the marginal nucleotide frequencies.

Now if we are given a sequence \( X_1, \ldots, X_{12} \), where \( X_i \in \{A, C, G, T\} \), then we can consider the likelihood ratio

\[
LR = \frac{p_1^{X_1} \cdots p_{12}^{X_{12}}}{q^{X_1} \cdots q^{X_{12}}},
\]

or the log-likelihood ratio

\[
LLR = \log p_1^{X_1} + \cdots + \log p_{12}^{X_{12}} - \log q^{X_1} - \cdots - \log q^{X_{12}}.
\]

Grouping the terms in the LLR a different way yields

\[
LLR = \log p_1^{X_1} - \log q^{X_1} + \cdots + \log p_{12}^{X_{12}} - \log q^{X_{12}}.
\]

If we let \( S_j(X) = \log p_j^X - \log q^X \) for \( X \in \{A, C, G, T\} \), then the \( S_j(X) \) are the entries of the PWM.
To make things look nice, let $M$ be the smallest $S_j(X)$ over all $j$ and $X$, then take $\tilde{S}_j(X) \equiv S_j(X) - M$. This gives an equivalent PWM with positive entries.

A major shortcoming of the PWM/multinomial approach to identifying matches to a target consensus sequence (e.g. 5' splice acceptor site) is that it ignores information in the interactions between different positions within the sequence.

The full multivariate distribution of a sequence of length $m$ covers $4^m$ points, which is large even when $m$ is moderate (e.g. $4^{12} \approx 17 \times 10^6$). Thus we will not be able to model the joint distribution of the target sequence unless it is quite short. We need a more parsimonious representation that can still take advantage of some multivariate structure. Here are some approaches:

1. **Inhomogeneous Markov Distributions.** Suppose that we model the distribution of $X_1, \ldots, X_m$ as a first order inhomogeneous Markov chain:

   \[ Q(X) = p_1(X_1)p_2(X_2|X_1) \cdots p_m(X_m|X_{m-1}). \]

   There are 3 parameters in $p_1$ and 12 parameters in each of $p_2, \ldots, p_m$, giving $12m - 9$ parameters overall. We can estimate the parameters from training data $X^{(1)}, \ldots, X^{(n)}$ using the method of moments:

   \[ \hat{p}_1(X) = \frac{\sum_i \mathcal{I}(X_1^{(i)} = X)}{n}. \]

   \[ \hat{p}_k(X|Y) = \frac{\sum_i \mathcal{I}(X_k^{(i)} = X \land X_{k-1}^{(i)} = Y)}{\sum_i \mathcal{I}(X_{k-1}^{(i)} = Y)}. \]

   Again we can use a likelihood ratio to identify the matches.
2. **Maximal Dependence Decomposition** (Burge 1997). For each pair of locations $i, j$, carry out a $\chi^2$ test of the independence of the nucleotides at positions $i$ and $j$, giving a test statistic $T_{ij}$. Let $S_i = \sum_j T_{ij}$, and let $k$ be such that $S_k$ is the largest of the $S_i$. Position $k$ has more dependence with the other positions than any other position, so we split on position $k$. That is, we model the sequence as $P(X) = P(X_{-k}|X_k)P(X_k)$, where $X_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$. To fit the model, construct independent multinomial distributions $\mu_A, \mu_T, \mu_G, \mu_C$ for $X_{-k}$, where $\mu_A$ is computed using only the training sequences with A at position $k$, and so on. Let $\pi_A, \pi_T, \pi_G, \pi_C$ denote the frequency of A, T, G, and C at position $k$, and let $f_A, f_T, f_G, f_C$ denote the marginal frequencies over all positions.

Now suppose we are given a sequence $X$ to classify. We use the log likelihood ratio $\rho = \log P(X)/R(X)$ to discriminate target from non-target sequences. Since $\rho = \log P(X_{-k}|X_k) + \log P(X_k) - \log R(X)$, or $\rho = \log P(X_{-k}|X_k) - \log R(X_{-k}) + \log \pi(X_k) - \log f(X_k)$, the score can be viewed as the application of one of four different PWM’s to $X_{-k}$, plus a likelihood ratio term for $X_k$. 
