Spectral-Based DNA Sequence Analysis

A DNA sequence of length \( N \) can be written as \( S = S_0 S_1 \ldots S_{N-1} \) where \( S_i \in \{A,T,G,C\} \). Typically, the DNA sequence is rewritten as

\[
\]

where \( u_A[n], u_T[n], u_C[n] \) and \( u_G[n] \) are binary indicator sequences and \( \{a,t,c,g\} \) are weightings associated with the corresponding binary sequences. The binary indicator sequences take the value of 1 or 0 at location \( n \), depending upon whether the corresponding character exists at \( n \).

The goal of performing spectral analysis on DNA sequences is to highlight sequence structure and frequency components that may be present. Discrete Fourier transform (DFT) can be applied to the numerical sequence to analyze its spectral features. In particular, the power spectrum can be formed as

\[
\tilde{X}[k] = \sum_{j=A,G,T,C} \tilde{U}_j[k]^2
\]

where \( \tilde{U}_j[k] = \sum_{n=0}^{N-1} u_j[n] e^{-j2\pi nk/N} \). The spectral approach for DNA sequence analysis relies on the assumption that the spectrum for coding region is different from that for non-coding region due to codon usage bias. In particular, a coding region is identified if a peak at frequency \( 2\pi/3 \) is observed. However, the magnitude of the peak varies greatly. To increase the discriminating power, one can adjust the four weights, \( \{a,g,t,c\} \), in (1). For example, Anastassiou [5] obtained the weightings through an optimization process which maximizes the differences between the spectra formed from exons and introns in a set of “training” sequences.

Z-Curve Approach

The Z-curve approach [3] extracts features directly from the character-based DNA sequence. In particular, statistical information about the cumulative frequencies of the occurrence of individual nucleotide is used. Let the frequencies of bases A, C, G and T at positions 0, 3, 6, ...; 1, 4, 7, ... and 2, 5, 8, ... respectively be \( A_0, C_0, G_0, T_0; A_1, C_1, G_1, T_1; A_2, C_2, G_2, T_2 \), the nine features in the Z-curve approach are then defined as

\[
\begin{align*}
f_{3i} &= (C_i + G_i) - (A_i + T_i) \\
f_{3i+1} &= (A_i + C_i) - (G_i + T_i) \\
f_{3i+2} &= (A_i + T_i) - (C_i + G_i) \quad i = 0, 1, 2
\end{align*}
\]

The biological interpretation of the above three measures are as follows [3]: component \( f_{3i} \) displays the distribution of bases of the purine (A or G) and pyrimidine (C or T) types along the sequence. Component \( f_{3i+1} \) displays the distribution of the bases of amino (A or C) and keto (G or T) types. Component \( f_{3i+2} \) displays the distribution of the bases of the weak H-bond (A or T) and strong H-bond (G or C) types. These nine values form a feature vector which helps to distinguish the coding from non-coding regions. For example, a neural network can be employed or a Fisher discriminate analysis can be used to perform the classification [3].

Relationship between Z-Curve & Spectral Approaches

Both the spectral approach and the Z-curve exploit the three-periodicity in the coding region. To elucidate the relationship between the two, we studied the Z-curve features from a signal processing perspective. Using the binary indicator sequence \( u_b[n] \), the cumulative frequencies \( A_i, C_i, G_i, T_i \) can be written as,

\[
h_b = \frac{1}{N} \sum_{n=0}^{N-1} l_i[n] u_b[n] \quad b \in \{A,G,T,C\}, i = 0,1,2
\]

where \( l_i \) captures information relating to the nucleotide position and is defined as,

\[
l_i = \sum_{m=0}^{N-1} \delta(n-3m-i) \quad i = 0,1,2
\]

ISSN 0973-2063

www.bioinformation.net
Using (4) and (5), (3) can be written as,

\[
f_{3i+i'} = \frac{1}{N} \sum_{n=0}^{N-1} s_i[n] l_i[n]
\]

where \(s_0[n], s_1[n]\) and \(s_2[n]\) are the modified sequences and are formed from the DNA sequence \(x[n]\) with \(\{a,g,t,c\}\) equals to \{1,1,-1,-1\}, \{-1,-1,1,1\} and \{1,-1,1,-1\}, respectively. Using the Parvesal’s theorem, (6) can be written in the frequency domain as

\[
f_{3i+i'} = \frac{1}{N^2} \sum_{k=0}^{N-1} \tilde{S}_i[k] \tilde{l}_i[k]^*
\]

where \(\tilde{S}_i[k]\) is the N-point DFT of the modified sequences \(S_i[n]\), \(\tilde{l}_i[k]\) is the N-point DFT of sequence \(l_i\) defined in (5) and * denotes complex conjugate. Note that the DFT of sequence \(l_i\) can be written as a sum of delta functions.

Using the conjugate property of a real sequence, (9) can be written as

\[
f_{3i+i'} = \frac{1}{3N} \sum_{m=0}^{2} \tilde{S}_i[m] \left( Nm \right) \frac{2\pi m}{3} e^{-j \frac{2\pi m}{3}}
\]

Substituting (8) into (7) gives,

\[
f_{3i+i'} = \left( \frac{1}{3N} \sum_{m=0}^{2} \tilde{S}_i[m] \right) \left( \frac{2\pi m}{3} \right) e^{-j \frac{2\pi m}{3}}
\]

Using the conjugate property of a real sequence, (9) can be written as

\[
f_{3i+i'} = \left( \frac{1}{3N} \sum_{m=0}^{2} \tilde{S}_i[m] \right) \left( \frac{2\pi m}{3} \right) e^{-j \frac{2\pi m}{3}}
\]

Eq (10) shows the relationship between the Z-curve features \(f_{3i+i'}\) and the spectra of the three modified sequences \(S_i[n]\) at frequency \(N/3\). It clearly shows that the Z-curve features measure different compositions of the nucleotides along the sequence as well as any 3-periodicity present in the coding region. For example, the DC value \(\tilde{S}_0[0]\) measures the difference between the distribution of the bases of purine and pyrimidine types, \(\tilde{S}_1[0]\) measures the difference between the distribution of the bases of amino and keto types, and \(\tilde{S}_2[0]\) measures the difference between the distribution of the bases of the weak H-bond and strong H-bond. The term \(\tilde{S}_i[N/3]\) implies a sampling at every third position and detects the 3-periodicity characteristic. If the magnitude of \(\tilde{S}_i[N/3]\) is large, a peak is observed in which a coding region is identified.

**Comparative Analysis**

Both the FT approach in (2) and the Z-curve approach ((9) or (10)) attempt to measure the 3-periodicity in the DNA sequence. Nevertheless, there are significant differences between them.

Although \(S_0 = U_A + U_G - U_T - U_C\), \(S_0^2 \neq |U_A|^2 + |U_G|^2 + |U_T|^2 + |U_C|^2\). Thus, the weighting in the FT approach is different from the weighting used in the Z-curve approach. The former considers each spectrum independently as \(\sum_{j \in \{A,G,T,C\}} w_j |U_j|^2\) while the Z-curve approach considers the spectra of the modified sequences.

The periodicity assumption is also different between the FT approach and the Z-curve approach. In the Z-curve approach, the periodicity assumption applies with regards to the biological properties and the nucleotide positions induced by the different base combination. In contrast, the periodicity assumption in the FT approach is made regardless of the biological properties. It simply sums up the spectra of different nucleotide indicator sequences independently. To demonstrate, let’s consider an artificial sequence \(\{T, A, G, C, G, A\}\). In the FT approach, this gives rise to four binary indicator sequences, \(\{0, 1, 0, 0, 1\}\) (A), \(\{0, 0, 1, 0, 1\}\) (G), \(\{1, 0, 0, 0, 0\}\) (T) and \(\{0, 0, 0, 1, 0, 0\}\) (C). Periodicity cannot be observed in any sequence. In the Z-curve approach, the modified sequence \(s_0[n]\) is \{-1, 1, 1, -1, 1, 1\}, which shows strong 3-periodicity. Finally, three modified sequences which characterize different biological properties are considered in the Z-curve approach whereas only the original sequence is considered in the FT approach.
FT approach. Hence, the FT approach considers only one spike at $2\pi/3$ for classification whereas the Z-curve approach considers both the DC value and the value at $2\pi/3$ of three modified sequences.

In view of the above analysis, we proposed to apply the FT to the modified sequences $s_0[n], s_1[n]$ and $s_2[n]$. The power spectra of the modified sequences are first formed. The three DC values and the three values at $2\pi/3$ can then be used for sequence classification. These DC and $2\pi/3$ features are in fact closely related to the nine Z-curve features as seen in (10), and they carry similar biological interpretation as the Z-curve features.