DNA in chromatin: how to extract structural, dynamical and functional information from the analysis of genomic sequences using space-scale wavelet techniques

Seminar on 24 Oct 2005 host by:
Alain Arneodo
Laboratoire Joliot-Curie / Laboratoire de Physique
Ecole Normale Supérieure de Lyon
46 allée d’Italie, 69364 Lyon Cedex 07, FRANCE
Resume composed by JIA Hongbo and MEMARI NAMIN Peyman

The main topic of this seminar by Dr. Alain Arneodo is an interdisciplinary study on the structural characteristics of DNA, in a theory frame of random walking and by applying the techniques of wavelet analysis to the DNA code sequences. The results have shown the relations between the long-range correlations and the structure of nucleosome for different types of DNA.

1 Introduction: Hierarchical structure of eukaryotic DNA

The complete hierarchical structure of eukaryotic DNA is shown in the figure on the right. Each DNA molecule has been packaged into a mitotic chromosome that is 50,000x shorter than its extended length. But how does it come out in this way? The statistical study on DNA code sequences will indicate the mechanism.
2 Model and method of analysis

2.1 The DNA walker representation

Each nucleotide is associated with a numerical value, +1 for purines (A/G) and -1 for pyrimidines (C/T). The DNA code sequence, which is expressed by the four bases A/G/C/T, is now represented by a walker on the line. The value associated with the $i^{th}$ nucleotide defines the $i^{th}$ step $S(i)$ of the walker. As shown in the figure below. One gets a signal from a specific DNA sequence. Therefore one could further apply some signal processing methods to analyze it.

$$f_n = \sum_{i=1}^{n} S(i)$$

To characterize such a signal, one uses the root-mean square of the height fluctuations, which is exponentially increased by the length of signal:

$$W(L) = \sqrt{\langle f^2(x) \rangle} - \langle f(x) \rangle^2 \sim L^H$$

where $H$ is the roughness exponent, indicating the properties of correlations in the signal.

- $0.5 < H < 1$ Long-range correlations (LRC)
- $H = 0.5$ uncorrelated
- $0 < H < 0.5$ anti-correlations

One defines the correlation function for a certain distance $l$

$$C_f(l) = \langle f(x)f(x+l) \rangle - \langle f(x) \rangle^2 \sim l^{2H}$$

which has a exponential behavior with $2H$ index.
2.2 Wavelet analysis of fractal signals

The wavelet analysis is a very common method for signal processing, which can be expressed by the following transforming equation:

\[ T_g(a,b) = \int_{-\infty}^{+\infty} g\left(\frac{x-b}{a}\right)f(x)dx \]

The wavelet transform works like a 'mathematic microscope' which allows us to locate \( b \) as positions of the singularities of the signal \( f \) and to estimate \( a \) as their strength \( h(x) \). The figure on the right shows a wavelet analysis of the DNA sequence of the bacteriophage \( \lambda \).

3 Results of analysis on DNA sequences

3.1 The roughness exponent \( H \)

By linear fitting the behavior of singularities versus code length, one gets the roughness exponent and then the properties of long-range correlations, as shown on the right.

- \( H = 0.5 \) no LRC
- \( 0.5 < H < 1 \) LRC

The DNA walker reflects some characteristics of the DNA structure, like the local curvature (nucleosome position) and local flexibility (DNAse-I sensitivity).
According the comparison between analysis results of S.cerevisiae and E.coli, the author proposed a hypothesis that, LRC in small-scale regime is the signature of the nucleosomal structure. More analysis on the sequences of Pox virus and Archaebacteria show that, LRC in small-scale regime exists in the former one but not in the latter one.

3.2 Influence of the DNA sequence on the formation and dynamics of nucleosomes

The figure on the right shows the influence of LRC in DNA sequence on the formation and dynamics of nucleosomes, that LRC favor DNA loop formation and DNA loop mobility, which is further confirmed by a 2-D elastic model and AFM studies.

4 Large-scale analysis of the Human genome

Now the author demonstrates how it goes from DNA sequence analysis to the modeling of replication in higher eukaryotes.

4.1 Compositional skew

In general, the symmetrical properties of the strands should follow the “Parity Rule type 2”, where \([A] = [T] \& [G] = [C]\) in each strand. But there exist deviations from this property, which is estimated by the compositional skews:

\[
S_{co} = \frac{[C] - [G]}{[C] + [G]}
\]

\[
S_{nt} = \frac{[A] - [T]}{[A] + [T]}
\]
Compositional skew is due to local biases in a strand in the course of biological mechanisms. The figure below shows some skew profiles around human replication origins.

The upward jumps rate \( \Delta S \) is about 24%. Intergenic regions show upward transitions of the skew.

4.2 Multi-scale detection of jumps in human skew profile

As shown in the figure above, for a sequence of Human chromosome 12, the wavelet based analysis detected of 2415 upward jumps and of a similar number of downward jumps.

5 Conclusion & commentary

The works of Alain Arneodo group demonstrates an active approach for statistics and analysis on the DNA sequence to find out the underlying mechanisms for the hierarchical structure. By applying the method of wavelet analysis, one could get a lot of substantial information on the structural characteristics of DNA.