Homework #3-Physics 498Bio

by Soon Yong, Chang – Nov/16/01

Review of the paper: *Minimum* Entropy Approach to Word Segmentation Problems by Bin Wang, from LANL archive physics/0008232 v1 29/Aug/2000.

INTRODUCTION & BACKGROUND

DNA is a text written with 4 distinct "letters". However, different from an ordinary text the DNA strand is a single long sentence devoid of delimiters such as space, comma, period, etc. In order to be able to "read" the information of the DNA, the first step is to be able to identify the "words" in the DNA sequence. This identification of "words" is called segmentation.

In this paper, the non-coding portion of DNA is focused, where the regulatory elements of the genes are found. Different from the coding portion where the word size is limited to 3 letters (codon) this region of DNA allows for larger flexibility in terms of the possible size of the words. The proper understanding of the segmentation holds the promise of better understanding the non-coding region of DNA.

The problem of correct segmentation is aggravated by the obvious fact that the same letter can belong to two different words (adjacent). This is equivalent to taking a pair of consecutive letters and to consider two cases: the pair belongs to the same word or each letter of the pair belongs to two different words. It is easy to see that a sequence of N letters can contain $N-1$ adjacent pairs. If segmentation is formally defined as the choice of connectivity between $N-1$ adjacent pairs, we can easily see that there are 2^{N-1} possible segmentations. So our "configurational" space grows exponentially as a function of the sequence length.

As seen in English, we should be able to recognize that DNA obeys certain "grammar" and thus its sequence must show somehow the non-randomness. It is here where the idea of *minimizing* "entropy" is a plausible concept, in contrast to maximizing entropy to find the thermally stable state. The actual DNA is a highly organized entity midst the sea of randomness and noise. The so-called segmentation entropy is defined and it is tested in some simple cases.

METHODS

First of all, the so-called the segmentation entropy (SE) has to be introduced with some constraints. The obvious constraint is the total number of letters;

$$\sum_l n_l l = N$$

where l= size of the word and n_l = n# of words of size l. The possible number of segmentations is given by:

$$\frac{\left(\sum_l n_l\right)!}{\prod_l n_l!}$$

Actual calculations give a fantastically large number even for a relatively small size "paragraphs", which can hint the expensive computational cost.

There are some strong assumptions made in this paper:

1). As we can see in the plain English, the word consist of vowels and consonants and most of the vowels can be discarded with the possibility that the word is still intelligible: for example "airplane" "arpln", " student" "stdnt", "computer" "cmptr", etc. And in some cases, even some consonants can be taken off. The apparent "redundancy" of letters contributes to easier segmentation when the continuous non-delimited sentence.

2). The same situation is applied in the sequence of non-coding DNA of the eukaryotes. The segmentation must be aided somehow by the non-randomness of the letters.

3). From the statistical point of view, the "correct" segmentation entropy would be minimum if the correct segmentation were found.

The segmentation entropy proposed by this paper is:

$$S = -\sum_{i=1}^{M} \frac{m_i l_i}{N} \ln\left(\frac{m_i l_i}{N}\right)$$

where m_i is the number of word w_i with length l_i. The minimization of this entropy will give the segmented sequence with m_i times the word w_i.

A way of testing this entropy is to have what is known to be the correct segmentation (in English text) and make random variations such as to choose two adjacent words and exchange their lengths. In this way the original two words evolve into two different words. This allows local variations of the original segmentation (after a long series of exchanges we will be able to sweep through the entire configurational space). In each step the segmentation entropy is recalculated. It is shown (calculated) that the initial segmentation is at least a local minimum of the segmentation entropy. The evolution of the segmentation entropy grows steadily with the random variations until it reaches a plateau with small fluctuations after a determined number of steps. And it is very unlikely that it will ever begin to go down to the original value or below.

Not only the sequential variations were studied but also truly random sampling of a large number of configurations. In both cases, it is suggested that the original segmentation corresponds, in fact, to the global minimum of this entropy. This is a sort of "weak" empirical proof that the above defined entropy works. This is weak, because there are some extreme cases where the entropy is small or zero: as when the entire sequence is considered to be a single word or when each word consists of only one letter. However, these cases could be disregarded for reasons other than statistical.

DISCUSSION

A purely statistical study of the segmentation was attempted without the knowledge of the DNA "vocabulary" or "grammar". This is not deliberate but the present state of matters. The validity of the segmentation entropy was tested with relatively short sequence of English language. The actual application to the DNA is still to be seen.

This concept is much broader and general than DNA sequencing. In fact, any symbolic sequence could be studied by this method. However, the vastness of the configuational space hinders its immediate implementation and a more efficient algorithm for sampling should be suggested. In fact, a version of Monte Carlo simulation where instead of the energy, the use of entropy would produce a random walk with a bias thorough the most probable configurations.