STUDY OF DNA SEQUENCE IN HUMAN GENES WITH STATISTICAL MECHANICS

Shwe Sin Oo*

Abstract

All living things have DNA. Human DNA contains about 38,000 genes. Characteristics of these living things depend on DNA sequence. Each gene sequence is determined by the way the base amino acid pairs (AT and CG) are arranged. In this work, human DNA base pairs were studied by statistical mechanics to investigate if the existing arrangement of these base pairs in human genes are in statistical equilibrium according to Maxwell-Boltzmann statistics.

Key words: DNA, base amino acid pairs (AT and CG), Maxwell-Boltzmann statistics

Introduction

Statistical Mechanics and the Structure of DNA

In a system of N particles where each particle can access to different energy levels E_1, E_2, E_3, etc., one can talk about the most probably partition where n_1 number of particles are in E_1, n_2 number of particles in E_2 etc. When the system is in this most probable partition, we say that the system is in statistical equilibrium. In classical statistical mechanics, Maxwell-Boltzmann distribution determines the number of particles in each energy levels for the statistical equilibrium. For the system to be in statistical equilibrium, Maxwell-Boltzmann distribution says that the number of particles in energy level E_i is given by

$$n_i = g_i \exp(- a - b E_i)$$

Where a and b are constants and g_i is the intrinsic probability of a particle to be in the state with energy E_i If all the energy levels are equally accessible, we will take $g_1 = g_2 = g_3 = \ldots = g$ etc.. The constant b is related to temperature by

$$b = 1/kT$$

*. Assistant Lecturer, Department of Physics, Loikaw University

Where k is the Boltzmann constant and T is the temperature. In order to understand how to determine if a system of N particles is in statistical equilibrium, consider a system of N=5000 particles with available energies E_1=0, E_2=E, and E_3=2E where E is some known energy. Now let us suppose that we know how many particles are in each energy levels. For example, let us take n_1=1000, n_2=3500 and n_3=500. Since the total number of particles is N=5000 and, to be in the statistical equilibrium we must use the expression for each n_i given by the Maxwell-Blotzmann distribution. i.e.

$$5000 = n_1 + n_2 + n_3 \qquad (A)$$

At statistical equilibrium, we get (for simplicity, in this example we take g =1 without loss of generality)

$$5000=\exp(-a -b E_1) + \exp(-a-b E_2) +\exp(-a -b E_3)$$

Recall that, in this example, $E_1 = 0$, $E_2 = E$ and $E_3 = 2$ E. Now by naming $y = \exp(-bE)$, $n_1= \exp(-a)$, $n_2 = \exp(-a-bE) = n_1 y$, and $n_3= \exp(-a-2bE)= n_1 y^2$

Equation (A) becomes

$$5000 = n_1 + n_1 y + n_1 y^2 \qquad (B)$$

The total energy is given by

$$E(total) = n_1 E_1 + n_2 E_2 + n_3 E_3$$

From the given numbers

$$E(total) = (1000 \times 0) + (3500 \times E) + (500 \times 2E) =4500 E$$

Now, to be in statistical equilibrium (SE), the total energy is given by

$$E (total\text{-}SE) = \exp(-a -b E_1)\ E_1 + \exp(-a -b E_2)\ E_2 + \exp(-a -b E_3)\ E_3$$

$4500 E= E (n_1 y) + 2E (n_1 y^2)$ which leads to

$$4500= n_1 y + 2 n_1 y^2 \qquad (C)$$

We can eliminate n_1 between (B) and (C) and get

$$11y^2+y-9=0$$

Which gives y = +0.86 where we have taken only the positive root since $y = \exp(-bE)$.

We can then find n_1 from either (B) or (C) and obtain n_1=1923. We also obtain $n_2 = n_1 y = 1653$ and $n_3 = n_1 y^2 = 1422$. Therefore we have found the populations of each energy level at statistical equilibrium. We also note that, the given configuration n_1=1000, n_2=3500 and n_3=500 is not the most probable partition. That means it was not in statistical equilibrium. If we found that the original n_is are very close to the calculated ones using Maxwell-Boltzmann distribution, we can conclude that the system is in statistical equilibrium.

At this point, we introduce the most basic ideas about DNA (Deoxyribo Nucleic Acid) and its structure. The discovery and the study of the structure of DNA has been one of the most exciting and challenging experiences that scientists have faced. As early as 1940s scientists started using bacteria and viruses in the study of genetics. The rapid reproduction rate of these simple life forms allowed the scientists to make a detail study of the structure of the genes. Even at that time, there were evidence that DNA played the role of a storage for genetic information. In 1953, James Watson and Francis Crick discovered the double helix structure of DNA. Now it is well known that DNA has a double helix structure with two sugar-phosphate backbones with rungs consisting of amino-acid base pairs. [1] In DNA, there are four types of amino-acids. (i) A (Adenine) (ii) T (Thymine) (iii) C (Cytosine) (iv) G (Guanine). Thymine (T) always pairs up with Adenine (A) and Guanine (G) always pairs up with Cytosine (C), hence AT and CG pairs. This DNA molecule is very long and the human DNA if stretched out would be about 2 meters long. It is like a long ladder with the two side rails made out of sugar and phosphate molecules and the rungs are entirely made out of AT (Adenine and Thymine) and CG (Cytosine and Guanine) pairs. Human DNA contains billions of these amino- acid pairs. There are approximately about 30,000 genes in human DNA. Each gene has a distinct sequence of AT and CG base pairs. Depending on the sequence of these base pairs, different types and amounts of proteins are produced in the organism. Each gene can contain from a couple of hundred base pairs to millions of base pairs. In between the genes in the DNA there are sequences which are the reminiscence of our evolution.

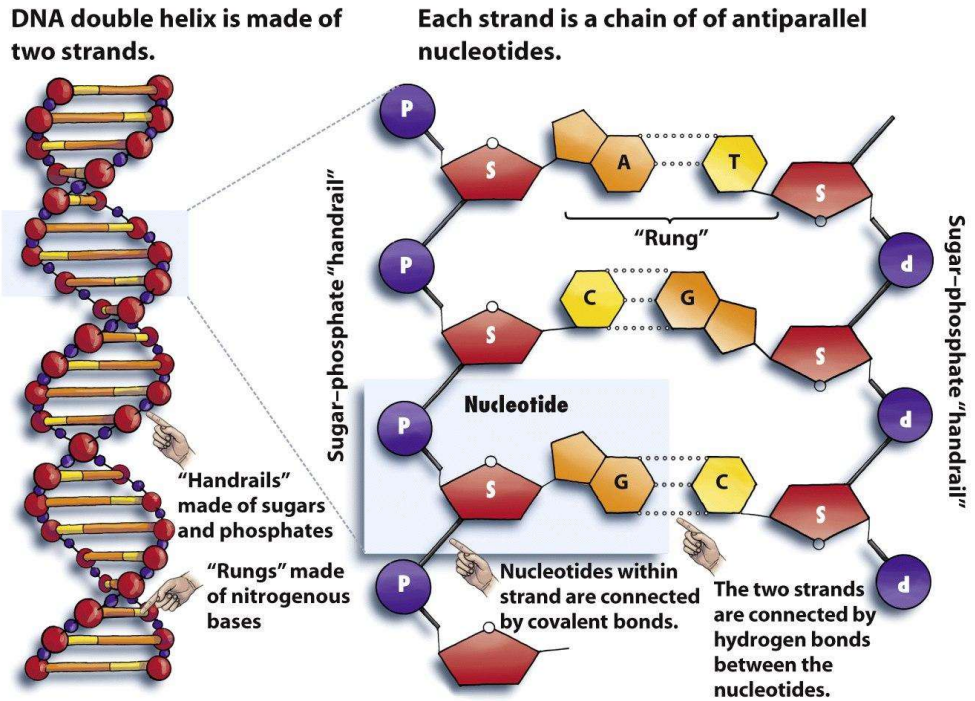

Figure 1. Structure and replication of DNA.

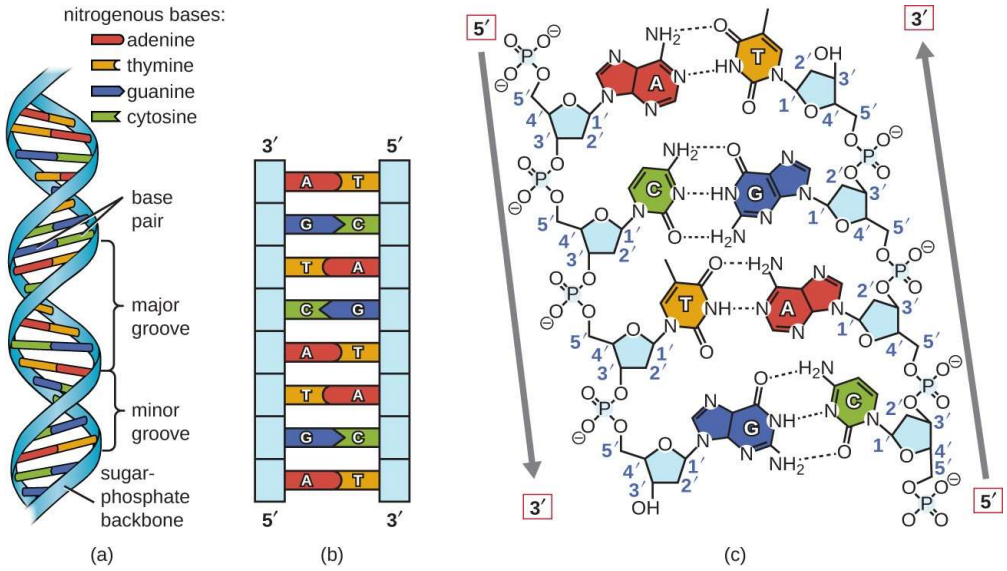

Figure 2. Structure and function of DNA.

All living things have DNA which stores the genetic information. The material that make DNA is the same regardless of whether it comes from a tree or an insect or a human. What is important is the sequence of these base pairs (amino acid pairs). The difference in the sequence makes each organism different.

Materials and Methods

DNA and Statistical Equilibrium

Now, we look at DNA from the point of view of statistical mechanics which is a very powerful tool in studying many particle systems. Let suppose a given gene has a total of 1000 base pairs with 375 pairs of ATs and 625 pairs of CG. Obviously there are millions of possible ways to arrange these base pairs in this gene. The following is one possible segment of a gene.

```
A T T T G C C G T G A T G C T T G G A T G
| | | | | | | | | | | | | | | | | | | |
T A A A C G G C A C T A C G A A C C T A C
```

Since these amino acids A, T, C and G interact with each other, an AT pair between a CG and an AT will have different energy than an AT pair between CG and GC. Since an A is always attached to a T and a C is always attached to a G, we can consider AT as a single particle and a GC as a single particle. Since their energies will depending on what their nearest neighbors, each AT and CG have different accessible energy levels. There are ten possible ways to sandwich a base pair. For example, an AT base pair can be in between

```
AAA     AAT     A A C     A A G
| | |   | | |   | | |     | | |
T T T   T T A   T T G     T T C

T A T     T A C     T A G
| | |     | | |     | | |
A T A     A T G     A T C
```

```
C A C        C A G
| | |        | | |
G T G        G T C

             G A G
             | | |
             C T C
```

Similarly we can sandwich a CG base pair the same way and we will get ten different possibilities again.

Since the interactions between different amino acids vary, the net energy for each AT base pair for the above sandwiching possibilities are all different. This gives us ten different energy levels for AT pair and ten levels for a CG pair.

Now, for a given gene sequence, we have a known number of AT base pairs and CG base pairs and they are arranged in a specific sequence. The natural question to ask is, if the sequence put this many particle system in a statistical equilibrium. As far as we are aware, no one has studied the question of statistical equilibrium regarding a gene sequence.

In the following, we outline our method for finding out if a DNA gene sequence is in statistical equilibrium.

First, we take each base pair AT and CG as single particles. This is justified because, A and T are always together and C and G are always together. We take AT and CG to be two different types of particles. The number of AT s and CG s can simply be obtained by counting them from the gene sequence under study. Next we need to consider the energies of each AT and CG. The change in free energies (Gibb free energy) of the base pairs which takes the nearest neighbor in to account is given by John Santa Lucia [3]. In his work, Lucia obtained the energies by making a comparison among the experimental results from seven different laboratories. [3]. There exists many other studies of base pair stacking and parameterization of the energies.

First consider the case of the base pair AT. Now for the total number NT of ATs we can write at equilibrium

$NT(AT) = g_1 \exp(-a - b E_1) + g_2 \exp(-a - b E_2) + g_3 \exp(-a - b E_3) + \ldots..$

By assuming that intrinsic probabilities for all levels to be the same, i.e. $g_1=g_2=\ldots= g = 1$ we can write

$$NT(AT)= \exp(-a-bE_1) (1 + \exp(-b (E_2-E_1)) + \exp(-b (E_3-E_1) + \ldots.)$$

$$NT(AT)= n_1 (1 + \exp(-b (E_2-E_1)) + \exp(-b(E_3-E_1)) + \ldots) \quad (D)$$

Here the value of NT(AT) is obtained from counting the numbers of ATs in the gene sequence.

Now for the total energy of ATs, we write

$ET(AT)= g_1 \ E_1 \exp(-a - b E_1) + g_2 E_2 \exp(-a-b E_2) + g3 E_3 \exp(-a - b E_3) + \ldots.$

Again, we take $g_1 = g_2 = g_3 =\ldots = g =1$, we write

$ET(AT) = \exp (-a- b E_1) (E_1 + E_2 \exp(-b(E_2-E_1))+ E_3 \exp(-b(E_3-E_1))+\ldots.)$

$$ET(AT) = n_1 (E_1+E_2 \exp (-b(E_2-E_1))+ E_3 \exp(-b(E_3-E_1))+\ldots.) \quad (E)$$

The value of ET(AT) on the left hand side of equation (E) is obtained from

$$ET(AT) = E_1n_1 + E_2n_2+ E_3n_3+\ldots.$$

Where all the n_i are obtained from counting the gene sequence.

Now, if our gene sequence is in statistical equilibrium, equations (D) and (E) must be consistent for a given temperature. Note that NT(AT) is obtained from counting and $(E_i - E_1)$ are obtained experimentally. Since $b = 1/kT$, temperature comes in through the value of b. Therefore for a given temperature, we can obtain n_1 from equation (D). (The most appropriate temperature to use here is the normal human body temperature which is about 37 degree C.) This n_1 value must also satisfy equation (E) if the system is in statistical equilibrium. A similar set of equations for CG can also be written and the same set of criteria must be applied to test statistical equilibrium.

Results and Discussion

In order to test our method of determining if a DNA sequence is in statistical equilibrium, we obtain a gene sequence from data base provided by The National Center for Biotechnology Information under the National Institute of Health, United States of America. [2]

Species: Homo sapiens (human)
Chromosome:1
Gene : S100A10
Number of rungs: 294

ATGCCATCTCAAATGGAACACGCCATGGAAACCATGATGTTTACAT
TTCACAAATTCGCTGGGGATAAAGGCTACTTAACAAAGGAGGACC
TGAGAGTACTCATGGAAAAGGAGTTCCCTGGATTTTTGGAAAATCA
AAAAGACCCTCTGGCTGTGGACAAAATAATGAAGGACCTGGACCA
GTGTAGAGATGGCAAAGTGGGCTTCCAGAGCTTCTTTTCCCTAATT
GCGGGCCTCACCATTGCATGCAATGACTATTTTGTAGTACACATGA
AGCAGAAGGGAAAGAAGTAG

Species: Homo sapiens (human)
Chromosome: 1
Gene: SPRR4
Number of rungs: 240

ATGTCTTCCCAGCAGCAGCAGCGGCAGCAGCAGCAGTGCCCACCC
CAGAGGGCCCAGCAGCAGCAAGTGAAGCAGCCTTGTCAGCCACCC
CCTGTTAAATGTCAAGAGACATGTGCACCCAAAACCAAGGATCCA
TGTGCTCCCCAGGTCAAGAAGCAATGCCCACCGAAAGGCACCATC
ATTCCAGCCCAGCAGAAGTGTCCCTCAGCCCAGCAAGCCTCCAAG
AGCAAACAGAAGTAA

We tested our method on these two genes using the energy values provided by Lucia [3] and we found that the condition for statistical equilibrium is not satisfied as expected. Calculation was done only using a pocket calculator. Here we note that these gene sequences are short, with 294 base pairs and 240 base pairs respectively.

Conclusion

In order to make a more concrete conclusion, we need to test on more genes with more number of base pairs. For long gene sequence where the number of base pairs involved can easily exceeds millions, calculations cannot be performed without the help of computer. We are in the process of developing a computer code which will read very long sequence (in millions) of the base pairs and test the statistical equilibrium. The other logical steps in this research would be to improve upon the energy values of the base pairs. One way would be to make a quantum mechanical model to calculate the energy levels of the base pairs using some well-known interaction potentials such as the harmonic oscillator model or Morse potential.

Acknowledgement

I would like to thank Dr Maung Maung and Dr Soe Myint Thein, Pro- Rectors, Loikaw University and Professor Dr. Than Than Myint, Head of department of Physics, Loikaw University for their kind permission to carry out this work. I am deeply indebted to my Advisor Professor Dr. Khin Maung Maung (Department of Physics and Astronomy, University of Southern Mississppi, USA), for his valuable advices for this work.

References

1. Benjamin A. Pierce , W. H. Freeman and company, NY. (2002) “Genetics”
2. Data Base, National Center for Biotechnology Information under the National Institute of Health, United States of America.
3. John Santa Lucia, Proc. Natl. Acad. Sci. “A unified view of polymer, dumbbell, and oligonucleotide DNA nearest neighbor thermodynamics”, USA, Vol 95, pp 1460, 1998
4. NY. McGraw Hill (1997) “McGraw Hill Encyclopedia of Science and Technology”