Case Study 1

Dengue virus (DENV) is the cause of dengue fever. It is a mosquito-borne, single positive-stranded RNA virus of the family Flaviviridae; genus Flavivirus. Five serotypes of the virus have been found, all of which can cause the full spectrum of disease.

I worked here with gene sequence of complete genome of virus to study it.

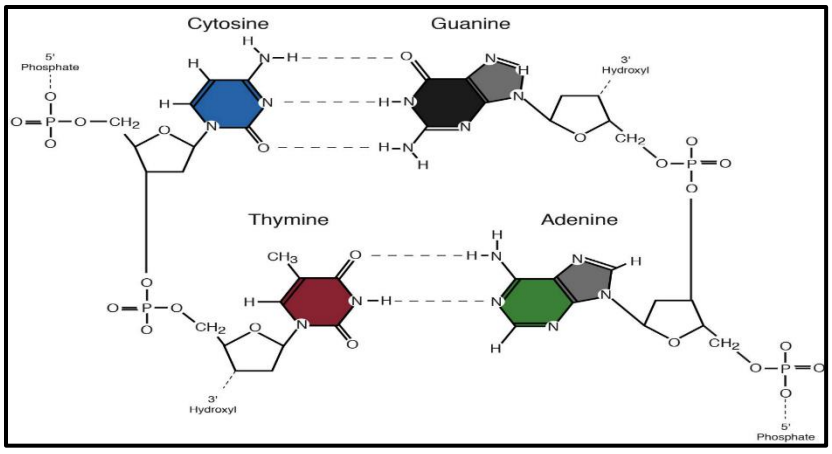
I use Seqinr package which has wonderful methods to analyze bioinformatics the biotechnology in sequence data.

Background

There are 4 nucleotides in DNA molecule: Adenine, Guanine, Cytosine, Thymine

These nucleotides are nitrogenous bases

GC content(guanine-cytosine content) is important in understanding genomic information it's % of bases on a DNA molecule that are either guanine or cytosine.



- GC-content (or guanine-cytosine content) is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine.

- This may refer to a specific fragment of DNA or RNA, or that of the whole genome.

- The GC pair is bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content; however. the hydrogen bonds do not stabilize the DNA significantly, and stabilization is due mainly to stacking interactions

- GC content helps us to study the stability of DNA.

- Bacterial genomes possess varying GC content (total guanines (Gs) and cytosines (Cs) per total of the four bases within the gen..) but within a given genome, GC content can vary locally along the chromosome, with some regions significantly more on less GC rich, than on average._ A low GCVAR indicates intra-genomic GC homogeneity and high GCVAR heterogeneity. The regression analyses indicated that GCVAR was significantly associated with domain (i.e. archaea or bacteria). phylum, and oxygen requirement. GCVAR was significantly higher among anaerobes(which means the species that doesn't require oxygen) than both aerobic and facultative microbes(which means highly organized or evolved species.

- It's very clear that GC content can help us to study genome of species, nature of species.

- Highly expressed genes were relatively G+C rich ,developmentally regulated genes were more A+T rich.

- Gene sequence analysis helps to understand the gene product or biology. . In bioinformatics, sequence analysis is the process of subjecting a DNA, RNA or peptide sequence to any of a wide range of analytical methods to understand its features, function. structure, or evolution. A collection of sequences does not, by itself. increase the scientist's understanding of the biology of organisms. However, comparing these new sequences to those with known functions is a key way of understanding the biology of an organism from which the new sequence comes. sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences.

- Difference between DNA and RNA in brief:
  DNA has two threads and RNA has one thread most viruses has RNA.

Starting Point of data analysis for geo-informatics

We have GC and AT content:

- GC bond have 3 hydrogen bonds: more GC content in any given genome simply means that more organism evolved and it has become mature
- AT bond have 2 hydrogen bonds
- simply GC content =(G+C)/(G+C+A+T)
- After calculating this ratio we can take decision and redraw inference whether that particular DNA microbe has matured or not.
- More GC content ➡ More stable of that particular genome.

What you do from scratch to make this study and that code?

1) import data you will analyze it from repos:
   - Genome information good repositories
   - US-National Center for Biotechnology information http://www.ncbi.nlm.nih.gov/genbank/
   - UK-European Molecular Biology Laboratory- http://www.ebi.ac.uk/
   - Japan-DNA Data Bank of Japan- http://www.ddbj.nig.ac.jp/

   Or simply if you want the same case study use den.FASTA file at my GitHub repo.

2) Install Seqinr by using install.packages("seqinr").

3) Use method which is called read.fasta(path_of_file) to read fasta format files. or you can use read.fasta(file.choose()) it will pop-up a window to choose file you want store data in R-object: den <- read.fasta(path)

4) Use edit(den) to show what is in your object and to verify that this object is list use class(den) and to know how elements in list use names(den)

5) To calculate GC content in first element of list use GC(den[[1]])

(R) Code

```
# load the library
library(seqinr)
# import the fasta file
den = read.fasta(file.choose())
#verify the file attributes
class(den)
#Extract sequence from "fasta file"
dseq <- den[[1]]
#verify sequence
edit(dseq)
#compute GC content
GC(dseq)
```

Conclusion

Scientists can know from value of GC content whether the virus is mature or not.