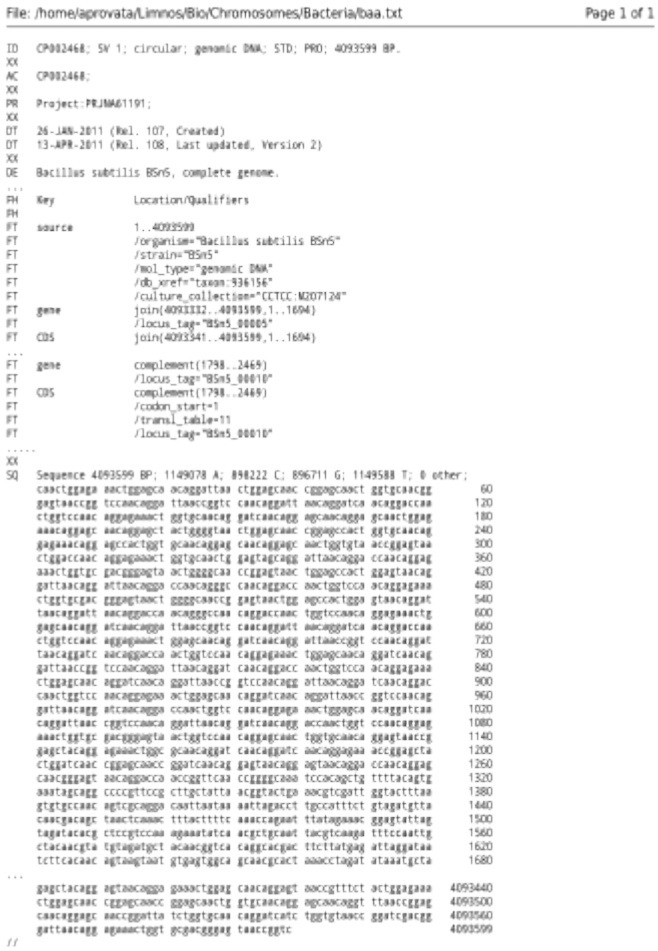
# Master's Thesis Proposal

**Thesis Title:** Complexity analyses in Bioinformatics: Comparison of DNA sequence complexity across organisms.
[Collaboration between the University of Athens & the National Center for Scientific Research "Demokritos" ]

**Aim:** The recent expansion in the field of Molecular Biology has produced complete **DNA sequences** of many organisms, ranging from higher Eukaryotes (humans, animals) to Bacteria and Viruses. These sequences carry most of the information about the organism's main characteristics. **Bioinformatics** has already designed specific computational tools to study and compare these sequences. Inspired from recent advances in the Theory of **Complex Systems** our aim is to use quantitative indices (**Correlation dimensions, Information dimensions, Fractal dimensions, Entropy production)** in order to quantify and compare the relative genomic complexity of test organisms belonging to different Kingdom: Animalia, Plantae, Fungi, Protista, Bacteria.

Typical image of a DNA sequence consisting of the 4 nucleotides (A,C,G,T) is depicted in the figure. This data originates from the bacterium ***Baccilus Subtilis*** which is commonly found in the human body and is widely used as a test organism for bacteria.


File: /home/aprovata/Limnos/Bio/Chromosomes/Bacteria/baa.txt Page 1 of 1

```
ID   CP002468; SV 1; circular; genomic DNA; STD; PRO; 4093599 BP.
XX
AC   CP002468;
XX
PR   Project:PRJNA61191;
XX
DT   26-JAN-2011 (Rel. 107, Created)
DT   13-APR-2011 (Rel. 108, Last updated, Version 2)
XX
DE   Bacillus subtilis BSn5, complete genome.
...
FH   Key             Location/Qualifiers
FH
FT   source          1..4093599
FT                   /organism="Bacillus subtilis BSn5"
FT                   /strain="BSn5"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:936156"
FT                   /culture_collection="CCTCC:M2017124"
FT   gene            join(4093332..4093599,1..1694)
FT                   /locus_tag="BSn5_00005"
FT   CDS             join(4093334!..4093599,1..1694)
...
FT   gene            complement(1798..2469)
FT                   /locus_tag="BSn5_00010"
FT   CDS             complement(1798..2469)
FT                   /codon_start=1
FT                   /transl_table=11
FT                   /locus_tag="BSn5_00010"
.....
XX
SQ   Sequence 4093599 BP; 1149078 A; 898222 C; 896711 G; 1149588 T; 0 other;
     caactggaga aactggagca acaggattaa ctggagcaac cggagcaact ggtgcaacgg        60
     gagtaaccgg tccaacagga ttaaccggtc caacaggatt aacaggatca acaggaccaa       120
     ctggtccaac aggaggaaact ggtgcaacag gatcaacagg agcaacagga gcaactggag      180
     aaacaggagc aacaggagct actggggtaa ctggagcaac cggagccact ggtgcaacag       240
     gagaaacagg agccactggt gcaacaggag caacaggagc aactggtgta accggagtaa       300
     ctggaccaac aggaggaaact ggtgcaactg gagtagcagg attaacagga ccaacaggag      360
     aaactggtgc gacggggagta actggggcaa ccggagtaac tggagccact ggagtaacag      420
     gattaacagg attaacagga ccaacagggc caacaggacc aactggtcca acaggagaaa       480
     ctggtgcgac gggagtaact ggggcaaccg gagtaactgg agccactgga gtaacaggat       540
     taacaggatt aacaggacca acaggccaa caggaccaac tggtccaaca ggaggaactg        600
     gagcaacagg atcaacagga ttaaccggtc caacaggatt aacaggatca acaggaccaa       660
     ctggtccaac aggaggaaact ggaggcaacag gatcaacagg attaaccggt ccaacaggat     720
     taacaggatc aacaggacca actggtccaa caggagaaac tggagcaaca ggatcaacag       780
     gattaaccgg tccaacagga ttaacaggat caacaggacc aactggtcca acaggagaaa       840
     ctggagcaac aggatcaaca ggattaaccg gtccaacagg attaacagga tcaacaggac       900
     caactggtcc aacaggagaa actggagcaa caggatcaac aggattaacc ggtccaacag       960
     gattaacagg atcaacagga ccaactggtc caacaggaga aactggagca acaggatcaa      1020
     caggattaac cggtccaaca ggattaacag gatcaacagg accaactggt ccaacaggag      1080
     aaactggtgc gacggggagta actggtccaa caggagcaac tggtgcaaca ggagtaaccg     1140
     gagctacagg agaaactggc gcaacaggat caacaggatc aacaggagaa accggagcta      1200
     ctggatcaac cggagcaacc ggatcaacag gagtaacagg agtaacagga ccaacaggag      1260
     caacgggagt aacaggacca accggttcaa ccgggcaaa tccacagctg ttttacagtg       1320
     aaatagcagg ccccgttccg cttgctatta acggtactga aacgtcgatt ggtactttaa      1380
     gtgtgccaac agtcgcagga caatttaataa aattagacct tgccatttct gtagatgtta     1440
     caacgacagc taactcaaac tttacttttc aaaccagaat ttatagaaac ggagtattag      1500
     tagatacacg ctccgtccaa agaaatatca acgctgcaat tacgtcaaga tttccaattg      1560
     ctacaacgta tgtagatgct acaacggtca caggcacgac ttcttatgag attaggataa      1620
     tcttcacaac agtaagtaat gtgagtggca gcaacgcact aaaacctagat ataaatgcta     1680
...
     gagctacagg agtaacagga gaaactggag caacaggagt aaccgtttct actggagaaa     4093440
     ctggagcaac cggagccaacc ggagcaactg gtgcaacagg agcaacaggt ttaaccggag     4093500
     caacaggagc aaccggaatta tctggtgcaa caggatcatc tggtgtaacc ggatcgacgg     4093560
     gattaacagg agaaactggt gcgacgggag taaccggtc                            4093599
//
```

**Typical DNA sequence (for analysis)**

**Methodology:**

For each test organism the following steps will be followed:

- The most recent genomic sequences of the organism will be downloaded from the genomic databases (EMBL, GenBank)
- Statistical characteristics of the organism will be calculated (e.g. number of A,C,G,T basepairs, coding percentage, repetition percentage, etc).
- Quantitative indices, such as the information dimension, fractal dimension, correlation dimension and entropy production will be evaluated.
- Classification of the organisms will be undertaken, based on the above indices.

Comparison of the above dimensionality classifications with the classical phenomenological classifications will be undertaken.

**Literature:**


- C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, "Mosaic Organization of DNA Nucleotides", *Physical Review E* **49**, 1685-1689 (1994).
- C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, "Long-Range Correlations in Nucleotide Sequences", *Nature* **356**, 168-171 (1992).
- W Li, J Freudenberg, P Miramontes , "Diminishing return for increased mappability with longer sequencing reads: implications of the k-mer distributions in the human genome", *BMC Bioinformatics*, **15**, 2 (2014).
- A. Provata, C. Nicolis, G. Nicolis, "DNA viewed as an out-of-equilibrium structure", *Physical Review E* **89,** 052105 (2014) .
- A. Provata, C. Nicolis, G. Nicolis, "Complexity measures for the evolutionary categorization of organisms", *Computational Biology and Chemistry,* **53**, 5–14 (2014).


**Collaboration:**

This Master Thesis is performed in the framework of a collaboration between:
-the Department of Informatics of the University of Athens (Prof. T. Theoharis),
-the Institute of Nanoscience and Nanotechnology of the National Center for Scientific Research "Demokritos" (Dr. A. Provata)

**Requirements:**

We are looking for an MA student with excellent knowledge of:

a) image analysis techniques
b) programming language C/C++ (or other)
d) English

Additional Skills (not required) could be useful:
a) Knowledge of Linux operating system
b) Knowledge of Fractal and Multifractal Analysis

**Interested students, please contact:**


Dr. A. Provata
Statistical Mechanics and Complex Dynamical Systems Laboratory
Institute of Nanoscience and Nanotechnology
National Center for Scientific Research "Demokritos"
15310 Athens, Greece

tel: +30 210 6503964
fax: +30 210 6511766
E-mail: a.provata@inn.demokritos.gr