Nonlinear Analysis and Visualization of the Genome of Evolutionary Younger Organisms

Panayotis G. Katsaloulis*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
panayotis@panayotis.com

Abstract. Computational analysis of DNA sequences can distinguish areas within the genome according to their characteristics and recognize or forecast the functionality of such areas. Following this approach, the present Thesis focuses on the statistical analysis of DNA sequences at different scales, by developing biocomputational algorithms. At the DNA base level, oligonucleotides were used to calculate statistical properties in order to create a homogeneity map, with the power of prediction for coding and non-coding areas. At the chromosomal level, various statistical parameters were calculated, based on the distances of homologous oligonucleotides. Calculation of the oligonucleotide distributions inside a chromosome and characterization of their functional behavior throughout the chromosome were performed. At the population level, the focus was on calculating statistical attributes of DNA areas, and creating a Small Tandem Repeats (STR) database. The computational results have been cross checked with experimental biological data. The algorithms are open source under the GNU GPL.

1 Introduction

In recent years new revolutionary experimental methods in molecular biology have been developed. It is now possible to sequence DNA macromolecules with increased speed and accuracy. This has resulted in an explosive growth of the amount of biological data being stored in biological databases (such as NCBI, EMBL and DDBJ). We now have complete genomic sequences for many organisms, even for organisms with extensive genomes, such as human (*Homo sapiens*) and mouse (*Mus musculus*).

At today's rates, the amount of data inserted into biological databases will double every 18 months. It is clear that this tremendous amount of data is of no value, unless there exist tools for effectively searching and manipulating it. For this reason various biological packages have been developed, such as BLAST [1], FASTA [2], CLUSTAL [3], while other numerical approaches and algorithms are presented in references [4–9].

Despite the increasing number of available tools, the problem of categorizing oligonucleotides based on their statistical properties is still open. Following this

* Dissertation Advisor: Theoharis Theoharis, Assoc. Professor

path, we have developed biocomputational algorithms which calculate statistical properties of sequenced DNA chromosomes in different levels of abstraction; at the base level, at the chromosomal level and at the population level.

In order to understand the basic characteristics of the proposed algorithms, it is important to describe the fundamental functional structure of the DNA. Although the main operation of the DNA is to code for proteins, it is known that, in higher eukaryotic organisms, only a small percentage of the DNA is translated in order to produce proteins. These areas are called *coding areas*. The rest of the DNA has more structural than functional role and such areas are called *non-coding areas*. In order for the coding areas to be distinguishable by the enzyme which promotes the transcription (like RNA polymerase II), usually in the beginning of the coding regions there is a special DNA sequence called the *promoter*. Promoters have a length of hundreds of bases. Between two successive appearances of a promoter there are at least one coding and one non-coding sequence. Inside each promoter there are small oligonucleotides of length $m = 2 - 10$, which are steadily present, called *consensus sequences*. Known consensus sequences in eukaryotes are, among others, the `CG` and the `TATA` sequences.

In the next section we describe the SHMap algorithm, which calculates a homogeneity map of a chromosome and which has power of prediction of functionality of DNA. In section 3 we discuss analysis at the chromosomal level, by calculating statistical properties of oligonucleotide combinations. In section 3.3 we approach the analysis of oligonucleotides by a two-parameter description of size distribution. In section 4 we perform analysis at the population level, by constructing and validating a STR database and in section 5 we present the conclusions and discuss some open problems.

2 Analysis at the base level

The aim of this approach was to search for specific areas in the chromosomes, to find areas with different characteristics and to annotate unknown DNA areas. For this reason we have developed a distributed system which is able to execute algorithms in order to statistically manipulate a chromosome. We have also constructed an algorithm able to "map" areas inside the chromosome according to their statistical behavior. We thus consider another biological observation; the *lack of homogeneity* within eukaryotic chromosomes. Each eukaryotic chromosome consists of areas with different composition. Some areas can be described as "random" from the statistical point of view, whereas other areas have more "stable" consistency [9].

The proposed algorithm (Statistical Homogeneity Map - SHMap) [10] marks areas of the chromosome according to their "randomness". As a base measure we employ all possible oligonucleotides of length m. We distinguish the areas which are rich in different oligonucleotides, and those which consist of only a few oligonucleotides. The biological meaning of these areas and their numeric values is as follows:

– a lower value implies that the distances between oligonucleotides in this area are generally smaller than the given threshold. Having short distances means that the possibility of finding any given combination, starting from any position inside this area is high, or in other words, that most combinations are present and mixed in this area. Since this kind of behavior resembles "random" distribution, it is also expected that these areas include mostly coding DNA sequences.
– a higher value implies long distances between oligonucleotides (above the given threshold). Having long distances means that it is less probable to find the next occurrence of a certain oligonucleotide inside this area. Since we do not consider extensive DNA gaps in this implementation of the algorithm, but there is a contiguous coverage of bases, the reason for the long distances is the over-representation of few specific oligonucleotide combinations in this area, forcing the remaining majority of the oligonucleotides to be under-represented. This behavior is common in non-coding DNA sequences, where the presence of structures like poly-A (long sequences consisting only of adenine) are common.

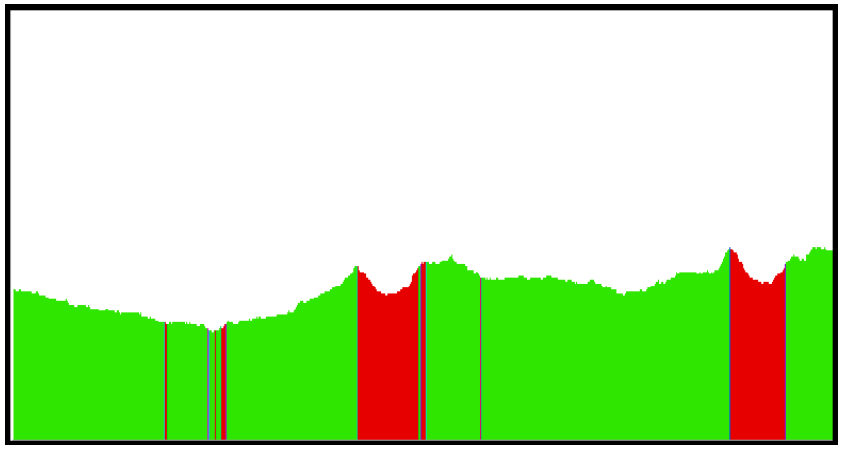

Fig. 1. Statistical Homogeneity Map (SHMap) for the annotated chromosome 17 of *Mus musculus*. The visible region is between the bases $12040000 - 12070000$. The size of oligonucleotides is $m = 6$. The lower the value on the vertical axis, the higher the homogeneity (and thus high possibility for coding sequences) and vice versa. Red color marks "CDS" areas, as found in the GeneBank genome file, green marks non-coding areas and blue marks areas with both coding and non-coding parts.

Using this algorithm it is possible to distinguish areas which are rich in oligonucleotide combinations (lower value) from those which are poorer (higher values). Since the coding regions appear to be richer in oligonucleotide combinations, we expect them to be inside the areas with the lower values. Since this

approach is statistical and not biochemical, there is less accuracy in the positioning of the exons and introns. It can be used as a tool to point to DNA areas which need to be further investigated by traditional biochemical methods.

Since the execution time of this algorithm was rather long, we parallelized it under the GRID environment using MPICH. This achieves significant speed-up compared to the sequential version of the algorithm. The tests have been performed on machines with 1, 2, 4 and 8 processors. We have found that the speed-up increases monotonically, as expected, while the parallel cost appears almost constant, independent of the number of processors.

In order to test the validity of the produced results of *Mus musculus* chromosome 17, we compare them to the actual biological information gathered from a biological database. The annotated chromosome 17 of *Mus musculus* from NCBI database has been used. Figure 1 shows such a comparison for the area $12040000 - 12070000$. We show the produced SHMap of this chromosome together with the coding sequence areas (CDS) on the same graph. The higher the values on the vertical axis the lower the homogeneity of the corresponding area of the chromosome. Red color denotes CDS and green color the rest of the chromosome, as obtained from the GeneBank file. We note that this approach does not produce a detailed distinction between coding and non-coding regions, but it can be used to focus the researcher on areas which have a high probability of being CDS. It is thus able to highlight areas of potential importance, since the structure and statistical characteristics of the DNA sequence are related to its functionality.

3 Analysis at the chromosomal level

3.1 Methodology

Recently, a common pattern was investigated in the size distribution of coding and non-coding regions of DNA [11, 9]. For higher organisms (eukaryotes) it was shown that the size distribution of coding sequences follows short range distributions, such as Gaussian or exponential, while the size distribution $P(S_{non-cod})$ of non-coding sequences follows long range laws:

$$P(S_{non-cod}) \sim S_{non-cod}^{-1-\mu}, \qquad where \quad 0 \leq \mu \leq 2 \tag{1}$$

where $S_{non-cod}$ is the size of the non-coding sequence in a large DNA strand and the value of μ is called the power law (or critical) exponent.

Since in evolutionary recent eukaryotes the majority of DNA is non-coding [12], coding DNA may be considered as 1-dimensional small "islands" (segments) floating in an 1-dimensional "ocean" (medium) of non-coding DNA. Promoters indicate the beginning of coding regions and it is evident that between two promoters exists at least one coding and one non-coding region (see Figure 1).

$$S_P = S_{cod}^{(1)} + S_{cod}^{(2)} + ... + S_{cod}^{(n)} + S_{non-cod}^{(1)} + S_{non-cod}^{(2)} + ... + S_{non-cod}^{(m)} \tag{2}$$

where S_P is the distance between two consecutive promoters, $S_{cod}^{(i)}$ is the size of the ith exon and $S_{non-cod}^{(j)}$ is the size of the jth non-coding region between the two adjacent promoters. n and m are the numbers of exons and non-coding regions (including introns), respectively. Since in general in eukaryotes

$$S_{non-cod} \gg S_{cod} \tag{3}$$

it is expected that

$$S_P \simeq S_{non-cod}^{(max)} \tag{4}$$

i.e. the distance between two consequent promoters is proportional to the size of the maximum internal non-coding region. Consequently, the distribution of separations between occurrences of the same promoter, $P(S_P)$, will have similar statistical characteristics with the size distribution of non-coding regions $P(S_{non-cod})$. We thus expect that:

$$P(S_P) \sim S_P^{-1-\mu}, \qquad where \quad 0 \leq \mu \leq 2 \tag{5}$$

Under this aspect, we have explored the statistical properties of the distances S between oligonucleotides in different organisms [13, 14]. Since some of these oligomers are consensus sequences, by observing their distance distributions, we expect that their statistics may reflect the biological meaning of these sequences and specifically they should present long range characteristics. Such an observation will bring together a theoretical approach (calculation of statistical properties between oligonucleotides), with experimental results (existence of consensus sequences). For this analysis we have chosen to study the following representative eukaryotes with long sequenced genome or even entire chromosomes: *Homo sapiens*, *Mus musculus*, *Saccharomyces cereviciae*, *Oryza sativa* and *Arabidopsis thaliana*.

3.2 Statistical parameters

An algorithmic tool was developed, which was able to calculate distances between oligonucleotides and various statistical parameters, such as average distance, distance deviation and distance size distribution. In the produced histograms, short scales represent the appearance of the oligonucleotide within genes, while long scales represent the separation of intergenic regions. For this reason we concentrate on the power law exponent $-\mu$ extracted from the tails of the size distribution, which is usually between the bounds $1 \leq ln[P(S)] \leq 6$ (in double logarithmic scale). Sometimes these bounds vary, accordingly to the size, content and origin of a given chromosome. At the end, the oligonucleotides are sorted according to their value of μ.

The general conclusion is that sequences which have the lowest values of $|\mu|$, are those which are consensus sequences of the RNA polymerase II promoter, and this holds for all organisms studied. This is consistent with the hypothesis that the inter-distance distribution between promoters is a power law distribution

and has long range correlations. This is also in accordance with the hypothesis that the size distribution of the non-coding areas follows a power law, since in higher organisms the separation between promoters consists mainly of intergenic non-coding regions.

Although the results are common for the majority of the organisms tested with oligonucleotides having the `CG` combination, it is interesting to focus on *A. thaliana*, which stands out of line. The sequences of *A. thaliana* that have the smallest value of $|\mu|$ are different from the corresponding sequences found in all other organisms tested in this work. The quadruplet with the smallest value of $|\mu|$ is the `TATA` sequence, followed mostly by point mutations of the `TATA` pattern. This result fits well with our original hypothesis, since this sequence is one of the consensus sequences (boxes) of RNA polymerase II. In this situation the most dominant consensus sequence, in statistical terms, seems to be `TATA` and not the `CG` pattern, which usually appears in all other cases.

Sequences which appear to have the largest values of $|\mu|$, are those which are rich in adenine or thymine. This is obvious mostly in organisms like *S. cereviciae* and *O. sativa*, and not in mammals, where these sequences have a more random composition. It is rather a consequence of the *poly-A* sequences (long sequences consisting only of adenine) which might affect the order.

3.3 Two parameter description of size distribution

Although special care was taken, so that only the linear regions in the double logarithmic plots are taken into account, in some cases when computing the power law exponent μ this approach is not optimal [15]. For this reason it is interesting to perform parametric curve fitting over the size distribution. The equation which was chosen was:

$$y(x) = Ax^{-1-m}e^{-kx} \tag{6}$$

Three fitting parameters are used, A, m and k. Parameter m corresponds to the intermediate scales, parameter k to the large scales, while A is a normalisation parameter. Parameters m and k appear in the exponent and thus show the statistical tendencies of the distribution. The value of m corresponds also to the critical exponent, similar to μ in Equation 5. Using a combined Levenberg-Marquardt with Gauss-Newton method, the three parameters $(A, -1-m, -k)$ are calculated.

In a graph with axes $-k$ and $-1-m$ we denote by X the positions of oligonucleotides containing at least once the promoter signature `CG`, while all other oligonucleotides are denoted by circles.

This anlysis has been performed on mammalian DNA (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*), on a bird (*Gallus gallus*), on an insect (*Drosophila melanogaster*), on a nematode (*Caenorhabditis elegans*) and on a dicot plant DNA (*Arabidopsis thaliana*).

In evolutionary recent organisms (mammals and birds) we can distinguish two areas clearly separated (Figure 2). The lower/right part consists of oligonucleotides in which the `CG` sequence appears at least once (X symbol - group α),

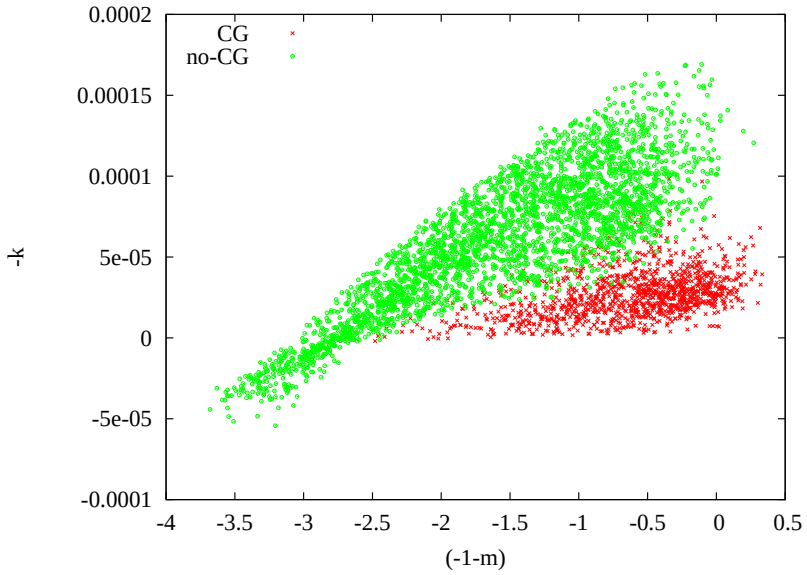

Fig. 2. Two dimensional plot of $(-1-m)$ against $-k$ in human chromosome 21 for quintuplets and hexaplets. Every point corresponds to one oligonucleotide combination. The combinations which consist of at least one occurrence of the `CG` sequence appear in the plot with the X symbol (red color), whereas the oligonucleotides which do not have this pattern appear with the O symbol (green color). The points (oligonucleotides) are divided into two groups/clusters in a claviform shape: the first consists *only* of oligonucleotides with at least one occurrence of `CG` sequence (lower right), while the second contains *exclusively* all other oligonucleotides.

whereas in the upper/left part there are oligonucleotides which do not have the CG sequence (O symbol - group β). It is interesting that there is a clear separation of these two groups; if an oligonucleotide has the CG sequence it belongs to the α group, whereas if it does not have the CG sequence it belongs to the β group.

This is in accordance with the observation that CG plays an important role as a consensus sequence in a promoter. If we focus on the position of the two areas within the graph, we note that the α areas have the smallest values of $-1-m$ and the value of $-k$ is around zero. This means that for these sequences the influence of the exponential part is very small, i.e. the distributions of sequences having the CG signature have almost pure long range nature. The other sequences have values of $-k$ which are away from zero, meaning that the influence of the polynomial part is much stronger. The difference between these two types of distributions might reflect the special biological meaning of CG, as an important part of a promoter.

The other organisms, namely *D. melanogaster*, *C. elegans* and *A. thaliana*, do not present the same behaviour. The distinction between the α and β areas is not clear. Still, the values of the parameters have the same meaning, as in the previous organisms. Again, the sequences that include the CG sub-sequence have the smallest values of $-1-m$ and the value of $-k$ is around zero.

4 Analysis at the population level

The analysis described above can not be used to calculate statistical properties of populations. The biochemical methods used to perform the DNA sequencing of a chromosome, consist of very expensive and time consuming procedures. There are also other reasons, such as mutations, which make statistical manipulation of the whole genome pool of a population an impossible task with present-day tools.

For this reason we have followed a different approach, in order to statistically treat a large population. We have looked at specific sequences with known polymorphism, precise position in the chromosomes and hereditary character. In *Homo sapiens*, areas found to have these properties are Restriction Fragment Length Polymorphism (RFLM) and STR (Short Tandem Repeats). In this Thesis, we have looked at STR sequences.

We have taken into account a random population of 200 subjects and created a STR database for this population. In order to check the validity of the database we have used forensic statistical indices, such as *Heterozygocity*, *Homozygocity*, *Hex*, *Probability Match*, *Power of Discrimination* and *Power of Information Content*. We have developed a multi-platform environment, able to calculate the statistics of the STR database of this population and the genotype probability factor of an unknown subject. This database has been cross examined with published databases from the bibliography.

5 Conclusions and open problems

We present a set of algorithms for the statistical analysis of DNA data at different levels. Namely, at the base level we have created the SHMap algorith with predictive power on the funcitonality (coding versus non-coding) of non-annotated DNA sequences. At the chromosomal level we have proposed statistical measures and algorithms which bring forward the special properties of the `CG` doublet, as well as other sequences which promote RNA transcription. At the level of population we propose algorithms which compute the genotype probability factors.

All our algorithms extract functionality properties using mathematical and statistical characteristics. Although the approach was completely based on mathematical terms, the sequences which stand out are those which have specific biological meaning (e.g. consensus sequences of promoters). This is significant, not only because it is extra proof for the special function of these sequences, but also because they could be used on organisms for which we have the DNA sequence but do not know much about the functionality of their genome.

In future studies other statistical properties will be considered. Among them is the effect of repeat sequences, special treatment for complementary oligonucleotides, taking into account punctuation marks etc. It would even be possible to define a phylogenetic tree based on the values of the power law exponent μ or similar statistical characteristics. In this tree, one could determine the evolutionary distance between organisms in terms of changes in the RNA polymerase itself or in the corresponding consensus sequences.

Analysis at the base level could also be enhanced by taking into account biological information, such as promoters and consensus sequences, punctuation marks, reading frame and polymerase reading direction. Thus, the prediction power of the algorithm in determining coding and non-coding areas, or in general areas with characteristics similar to coding DNA, could be enhanced.

It will be interesting to examine at which point in the evolutionary scale the clustering properties discussed on section 3.3 disappear. For this task it is necessary to obtain more data on the DNA structure of organisms between birds and insects. With this approach it would be possible to check when long range laws in inter-promoter distances appeared and how they are preserved.

The algorithms developed in this Thesis can be downloaded from our website `http://www.bioinfo.gr/` and are under the GNU GPL licence.

References

1. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Mol Biol*, 215:403, 1990.
2. W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85:2444, 1988.
3. D.G. Higgins, A.J. Bleasby, and R. Fuchs. Clustal V: improved software for multiple sequence alignment. *CABIOS*, 8:189, 1992.
4. C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356:168, 1992.

5. W. Ebeling and G. Nicolis. Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons and Fractals*, 2:635, 1992.
6. S. Karlin and V. Brendel. Chance and statistical significance in protein and DNA sequence analysis. *Science*, 257:39, 1992.
7. S. Karlin and V. Brendel. Patchiness and correlations in DNA sequence. *Science*, 259:677, 1993.
8. Y. Almirantis and A. Provata. The 'clustered structure' of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and non-coding sequences. *Bull. Math. Biol.*, 59:975, 1997.
9. A. Provata and Y. Almirantis. Scaling properties of coding and non-coding DNA sequences. *Physica A*, 247:482, 1997.
10. P. Katsaloulis, T. Theoharis, and A. Provata. Statistical algorithms for long DNA sequences: oligonucleotide distributions and homogeneity maps. *Scient. Prog.*, 13:177, 2005.
11. Y. Almirantis and A. Provata. Long- and short-range correlations in genome organization. *Stat. Phys.*, 97:233, 1999.
12. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.Da. Watson. *Molecular Biology of the Cell*. Garland Publishing Inc., New York, 2002.
13. P. Katsaloulis, T. Theoharis, and A. Provata. Statistical distributions of oligonucleotide combinations: applications in human chromosomes 21 and 22. *Physica A*, 316:380, 2002.
14. P. Katsaloulis, T. Theoharis, W.M. Zheng, B.L. Hao, A. Bountis, Y. Almirantis, and A. Provata. Long-range correlations of RNA polymerase II promoter sequences across organisms. *Physica A*, 366:308, 2006.
15. P. Bernaola-Galvan, J. L. Oliver, P. Carpena, O. Clay, and G. Bernardi. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. *Gene*, 333:121, 2004.