Long Range Correlations Between Nucleotide Triplets in Human Chromosomes

Daniele Vitale[*], Ercan E. Kuruoglu[†], Osman Abul[‡]

May 28, 2010

Abstract

The study of statistical fluctuations in DNA sequences can reveal important characteristics of their organization. Particularly, in the last two decades, several studies have focused on the detection of Long Range Correlations (LRC) in DNA sequences, and on the study of particular long range dependence properties of nucleotides. Since protein coding is carried out by codons (nucleotide triplets), we conduct LRC analysis to see whether there is a long range correlation among nucleotide triplets. Our interest is not limited to auto correlations of single codons, but also extends to cross correlations over all possible pairs of nucleotide triplets. LRCs in DNA sequences are studied and quantified using two measures: the mutual information function and the correlation function. The analysis on nucleotide triplets reveal LRCs for the human Chromosomes 20 and 21. Moreover, some triplets that contain only nucleotides Adenine and Thymine are seen to exhibit correlation significantly higher than others.

Keywords: Bioinformatics, Long Range Correlations, Mutual Information

1 Introduction

It was in April 2003 that International Human Genome Sequencing Consortium concluded the sequencing of whole human genome consisting of approxi-

[*]Daniele Vitale is now with University of Pisa, Department of Computer Science. He was with ISTI-CNR while this work was performed.

[†]Ercan E. Kuruoglu is with ISTI-CNR, Pisa, Italy.

[‡]Osman Abul is with TOBB University of Economics and Technology, Ankara, Turkey.

mately three billion base pairs. One of the surprising results was the difference between the expected number of genes in the sequence, ≈ 120000, and the empirical one, ≈ 30000(≈ 25000 according to most recent results). Moreover, this protein coding part constitutes only 1.5% of the whole human genome. It later became clear that solely studying the coding zone is not sufficient, although necessary, to explain the complexity of human being.

Statistical methods provide an effective tool for understanding the features and organization of the DNA better. In fact, one of the characteristics of DNA sequences is that they are not statistically homogeneous and show inherent variability (Karmeshu & Krisnamachari, 2004). To this end, the study of statistical fluctuations in DNA sequences can reveal some important information about the organization and functions of genomes.

An interesting feature detected in the statistical analysis of DNA sequences is the presence of Long Range Correlations (LRC). Typically, genomes are treated as signals and LRC models are characterized by their decay behavior that is similar to power law functions. Scale-free property of power law functions allows one to detect *self-similar* features, an important class of correlations, in the organization of genomes.

Long range correlations were detected in DNA by several researchers, *e.g.*, (Peng *et al.*, 1992), (Li & Kaneko, 1992), (Voss, 1992), (Buldyrev *et al.*, 1995) and (Audit *et al.*, 2001). These studies revealed that significant long range dependence is observed between nucleotides. In this study, however, we are interested in whether such dependencies exist in other representations (triplet representation, in particular) of genomes too. Hence, our objective is to investigate long range dependencies between nucleotide triplet pairs and to compare them with that of nucleotide pairs. We aim not only to uncover autocorrelations between codons but also to study the cross-correlations between any nucleotide triplet pair.

The paper is organized as follows. Statistical dependence measures that are used to measure LRCs are presented in Section 2. The section also details our method for experimentation. Section 3 explores LRCs in nucleotide representation and presents our experimental results on three selected human chromosomes, *Chromosome 20, Chromosome 21*, and *Chromosome 22*. Similar analysis on codon representations of selected chromosomes are detailed in Section 4, and the results are provided and contrasted. Finally, Section 5 provides concluding remarks.

2 Statistical Dependence Measures

Statistical dependencies in symbolic sequences can be studied and quantified with different methods. In this work, two measures are used: (1) Mutual Information Function $I(\cdot)$ (Shannon, 1948), and (2) the Correlation Function $C(\cdot)$ (Herzel & Grosse, 1995).

A discrete sequence (*e.g.*, DNA sequence) can be abstracted as a symbolic string S of length L, where each position contains a single symbol from an alphabet $\mathcal{A} = \{X_1, X_2, \ldots, X_\lambda\}$ of cardinality λ. For DNA sequences, the alphabet is $\{A, C, G, T\}$. We use S_i $(i = 1, 2, \ldots, L)$ to denote the symbol in position i of string S, and $f_i(S)$ to denote the relative frequency of symbol $X_i \in \mathcal{A}$ within sequence S, *i.e.* $f_i(S) = |\{S_j : S_j = X_i, \forall j = 1, 2, \ldots, L\}|/L$. Similarly, the relative frequency of a pair of symbols $X_i \in \mathcal{A}$ and $X_j \in \mathcal{A}$ at directional distance k $(k \geq 1)$ within S is denoted by $f_{ij}(S, k)$, *i.e.* $f_{ij}(S, k) = |\{S_l : S_l = X_i, S_{l+k} = X_j, \forall l = 1, 2, \ldots, L - k\}|/(L - k)$. For clarity, when S is fixed or understood from the context, the notations $f_i(S)$ and $f_{ij}(S, k)$ are simplified to f_i and $f_{ij}(k)$, respectively.

Under the assumption that S can be treated as a realization of an ergodic and stationary stochastic process, we can use these frequencies as maximum-likehood estimators of the probability p_i, which is the probability to find the symbol X_i in any given position of the sequence, and of $p_{ij}(k)$, which is the joint probability of finding symbol X_i in an arbitrary position of the sequence and symbol X_j at k letters downstream. However, when S is a real DNA sequence, the assumption of stationarity and ergodicity are very strong and must be handled carefully. This is simply because of the fact that DNA sequences show evidence of non-stationarity. Fortunately, as discussed in (Bernaola-Galvan *et al.*, 2002) the measures we use can be calculated for non-stationary sequences as well whilst the correlation measure cannot be related to the power spectrum.

2.1 Mutual Information

The entropy of a random variable X with a probability distribution function p is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{1}$$

The entropy of a random variable X measures the degree of uncertainty of the valu of the variable. It is possible to define a conditional entropy $H(X|Y)$ that is the entropy of a random variable X conditioned on the knowledge of another random variable Y. This reduction of uncertainty caused by the

knowledge of another variable is called *mutual information*. Mutual information between two random variables X and Y is defined as:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

where $p(x,y)$ is the joint probability of $X = x$, and $Y = y$.

Mutual information is symmetric in X and Y, always non-negative, and equal to 0 if and only if X and Y are independent (Cover & Thomas, 2006). Extending the previous definitions, for a fixed S, we can write the mutual information between two symbols at distance k as follows.

$$I(k) = \sum_{i,j=1}^{\lambda} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j} \tag{3}$$

$I(k)$ can be interpreted as the divergence of the real joint distribution of two symbols at distance k from the null hypothesis of statistical independence of the event.

2.2 Auto-Correlation Functions

Correlation functions measure statistical dependence between two random variables up to second order statistics, and are therefore can not measure the full statistical dependence. However, in case of random variables with Gaussian Distribution, the correlation can measure the full dependence. The classical correlation function of two random variables X and Y is given as:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{4}$$

where $Cov(X,Y)$ is the co-variance, and σ_x and σ_y are the respective standard deviations of X and Y.

The above given definition is valid for any time series with discrete or continuous outputs. In the case of finite-alphabet sources, that is when the random variable can take values from only a limited number of choices, such as in the case of DNA, a special form can be obtained for the correlation function.

In particular, given two symbols X_i and X_j; their linear dependence at distance k can be viewed as the difference between their joint conditional probability and the product of their marginal distributions as shown in the following (Herzel & Grosse, 1995):

$$D_{ij}(k) = p_{ij}(k) - p_i p_j \tag{5}$$

Recall that in mutual information (Eq. 3), the difference was measured in logarithmic units. Eq. 5, however, can be seen as the mutual information function with the logarithm removed, or the classical correlation with the expectation operations replaced directly by the probability densities.

Defining a Boolean row vector $\vec{a} =< a_1, a_2, \ldots, a_\lambda >$ containing ones for the variables $X_i \in \mathcal{A}$; for which we would like to calculate the correlations, we can write a specific correlation function $C_{\vec{a}}(k)$ in the bilinear form $\lambda \times \lambda$ of the correlation matrix $\underline{D}(k)$ with the elements $D_{ij}(k)$ (Herzel & Grosse, 1995) as:

$$C_{\vec{a}}(k) = \vec{a} \cdot \underline{D}(k) \cdot \vec{a^T} = \sum_{i,j=1}^{\lambda} a_i \cdot D_{ij}(k) \cdot a_j \tag{6}$$

where $\vec{a^T}$ denotes the transpose of row vector $\vec{a}$.

Consider that the alphabet $\mathcal{A}$ is fixed to DNA alphabet of the four letters, and $\vec{a} =< 1, 0, 0, 1 >$; then $C_{\vec{a}}(k)$ measures the auto-correlation of symbol set $\{A, T\}$ at distance k. Similarly, if $\vec{a} =< 1, 1, 1, 0 >$, then $C_{\vec{a}}(k)$ measures the auto-correlation of symbol set $\{A, C, G\}$ at distance k. For clarity, when $\vec{a}$ is fixed or understood from the context, the notation $C_{\vec{a}}(k)$ is simplified to $C(k)$.

Generalizing the above measure introduced in (Herzel & Grosse, 1995) to:

$$C_{\vec{a},\vec{b}}(k) = \vec{a} \cdot \underline{D}(k) \cdot \vec{b^T} = \sum_{i,j=1}^{\lambda} a_i \cdot D_{ij}(k) \cdot b_j \tag{7}$$

one can define a cross-correlation measure as well. For example, if $\vec{a} =< 1, 0, 0, 0 >$ and $\vec{b} =< 0, 1, 0, 0 >$, then $C_{\vec{a},\vec{b}}(k)$ measures the cross correlation at distance k between A and C.

The Taylor expansion of mutual information function (Eq. 3) results in the following equation, where $o(D_{ij}^3)$ represents third and higher order terms (Herzel & Grosse, 1995):

$$I(k) = \frac{1}{2ln2} \sum_{i,j=1}^{\lambda} \frac{D_{ij}^2(k)}{p_i p_j} + o(D_{ij}^3) \tag{8}$$

The equation shows that mutual information is approximately proportional to squared correlation functions (Herzel & Grosse, 1995) as the contribution of $o(D_{ij}^3)$ is negligible. So, in the sequel $C^2(k)$ will be plotted in figures for an easier comparison with mutual information function.

2.3 Long-Range Dependence

The concept of long-range dependence (LRD) is strictly related to slow decay of correlations in stochastic processes. Models for this kind of dependence were first introduced by Mandelbrot and his colleagues (Mandelbrot & Ness, 1968, Mandelbrot & Wallis, 1968) as an attempt to empirically explain phenomenon observed by Hurst (Hurst, 1951, 1956). Hurst discovered that water level evolution of river Nile grows with an unusual rate, $R/S - statistic$. Mandelbrot explained the phenomenon (known as Hurst Phenomenon) as an unusual behavior in time with a long memory. Long range dependences (a.k.a. long memory process) are characterized by a power law decay of correlations. This is in contrast with processes for which the correlations decay exponentially.

Several studies in the first half of nineties have found examples of long range correlations in DNA. Long-range correlations in intron-containing genes and in untranscribed regulatory DNA sequences were found in 1992 by Peng et al. (Peng *et al.*, 1992). Same year, Li and Kaneko (Li & Kaneko, 1992) found examples of LRCs in non-coding DNA, and detected a $1/f^{\alpha}$ behaviour in the power spectrum. Similar observations were repeated by Voss in (Voss, 1992). Later, Buldyrev et al. provide a study long-range correlation properties of coding and non-coding DNA sequences in (Buldyrev *et al.*, 1995) LRD analysis were also carried out to investigate properties of complete genomes of many species including human genome (Bernaola-Galvan *et al.*, 2002, Yu *et al.*, 2001). A multifractal analysis of genome sequences was provided in (Li & Holste, 2005).

The presence of LRCs in DNA sequences therefore has been long known, though its origin remains unclear. A possible interpretation is that they can be generated by processes such as duplication, single-site mutations, and deletion of existing segments of sequence (Li *et al.*, 1994). Further biological explanation is provided in (Mackiewicz *et al.*, 2002). The characteristics seem to be closely related to the evolutionary process (Isohata & Hayashi, 2003, Messer *et al.*, 2005a). Messer (Messer *et al.*, 2005b) and others confirmed this interpretation by further studies that proved the emergence of long range correlations in sequences generated using models based on evolutionary dynamics. Following it, (Messer & Arndt, 2006) implemented a tool capable of measuring amplitude and decay exponent of LRCs of a given sequence and to generate new random sequences with user specified LRCs.

3 LRCs in nucleotide sequences

In this section, we calculate LRCs in single nucleotide representations of human genome. Particularly, we measure $I(k)$ and $C(k)$ for varying k on three human chromosomes (Chromosome 20, 21 and 22) separately. A similar study has been carried out by Beirer et al. (Beirer, 2003, Holste *et al.*, 2003) on the same set of chromosomes. We first validate our results comparing with the results obtained by (Beirer, 2003), and hence verify the efficiency of the quasi-logarithmic sampling scheme we used.

First of all, we briefly describe the software tool, called LRCUtils, we developed to detect long range correlations in DNA sequences. The software is implemented in C++ and computes the functions described in Section 2.1 and 2.2. The functions on a given sequence is computed up to $k = 10^6$. For faster computation, the functions are calculated on samples of k on quasi-logarithmic scale, *i.e.*, for $k = 10, 20, \ldots, 100, 200, \ldots, 10^3, 2 * 10^3, \ldots$.

For this part of the experimentation, we are particularly interested in auto-correlations of weakly-binding nucleotides, *e.g.*, A and T. So, we set $\vec{a}$ to $< 1, 0, 0, 1 >$ and compute $C(k)$ accordingly, and denote $C(k)$ values as $C_{WW}(k)$. We are also interested in the auto-correlation of purine nucleotides (A and G) to compute $C_{RR}(k)$. $\vec{a} =< 1, 0, 1, 0 >$ is used to do so. As shown in Figure 1, $I(k)$ and $C(k)^2_{WW}$ shows LRCs up to 10^4 bp, for Chromosome 20 and 21, and up to 10^5 bp for Chromosome 22. The power law exponents γ of LRCs were estimated by least-square regression. The estimated γ values are shown in Table 1.

As can be seen by comparison with Table 2, which reports power law exponents calculated by Beirer, the values are perfectly compatible despite the nonuniform sampling we used.

	Chr20	Chr21	Chr22
$2\gamma_I$	0.457	0.415	0.52
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{WW}$	0.403	0.367	0.474
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{RR}$	1.32	1.28	1.45
range of fit(bp)	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$

Table 1: *Power law coefficients computed by LRCUtils.*

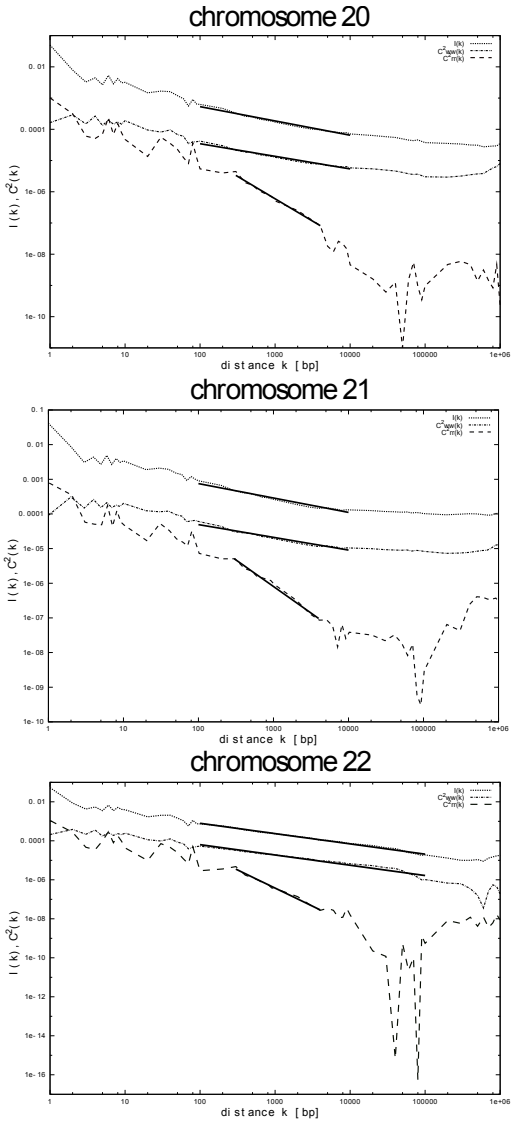

Figure 1: *Long range dependencies in nucleotide representation*

	Chr20	Chr21	Chr22
$2\gamma_I$	0.46	0.41	0.51
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{WW}$	0.40	0.37	0.48
range of fit(bp)	$100 .. 10^4$	$100 .. 10^4$	$100 .. 10^5$
$2\gamma_{RR}$	1.49	1.67	1.63
range of fit(bp)	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$	$100 .. 5 \times 10^3$

Table 2: *Power law coefficients computed by (Beirer, 2003).*

4 LRCs in Codon Sequences

As shown in the previous section, previous work in the literature considered genomes as nucleotide sequences to analyze LRCs (Beirer, 2003, Isohata & Hayashi, 2003, Peng *et al.*, 1992). In this section, we consider genomes as sequences of nucleotide triplets or more specifically as codon sequences. It is possible to consider genomes as sequences of pairs, quadruples of nucleotides; the main reason for the choice of codons (triples of nucleotides) is due to their biological interest.

It is important to note that, we refer to generic nucleotide triplets as codons to not necessarily correspond to 20 nucleotide triplets that encode amino acids. Starting from this choice, we fix the alphabet as all possible combinations of three letter nucleotides, *i.e.*, $\mathcal{A} = \{AAA, AAC, \ldots, TTT\}$ of cardinality 64. The max distance k is always 10^6, but now is not expressed in base-pair but in triplets of base-pair. As a pre-processing step, input sequences are divided into triplets, starting from the first position to the last position. If the sequence length is not a multiple of three, then it is truncated to nearest multiple by removing one or two symbols at the end.

Mutual information, $I(k)$, are computed at varying k values as explained in the previous section. The results are plotted in Figure 2. $I(k)$ shows LRCs in Chromosome 20 up to 10^5 and in Chromosome 21 up to 10^4. Average correlations, $C(k)^2$, are plotted in 3 which again demonstrate long-range dependence. However, looking at auto-correlations of some specific nucleotide triplets we do not observe long-range dependence. Yet, there are some couples of triplets (Table 4) which exhibit decay behavior on $C(k)^2$ in parallel to $I(k)$. In other words, not all triplet pairs show LRCs but some do.

As mentioned before, we are working on generic nucleotide triplets but it could be interesting to note that among the triplets reported in Table 4 ATT, ATC and ATA encodes the same amino acid Isoleucine, while GGT, GGC and GGG encodes the same amino acid Glycine. This encouraged us to look for pairs of triplets that show significant $C(k)$ values. To do so, we calculated an average auto-correlation value $C(k)$ for every possible couple of nucleotide triplets at varying k. The order is considered irrelevant, for example the couples AAA-TTT and TTT-AAA are treated the same. This procedure resulted in 2080 possible couples. We selected from them such couples that have significantly higher $C(k)$ values, with respect to the overall average, for every value of k. 23 couples have been found to satisfy the criteria. They are all combinations of AAA, AAT, ATA, ATT, TAA, TAT, TTA, and TTT. To verify whether these values are also statistically significant, we compare the difference between the value of $C(k)$ for these couples and the average

auto-correlation values and the standard deviation. Given any k, we have 2080 random variables $C_{ij}(k)$. Their aggregate distribution, according to the central limit theorem, tends to be a normal distribution $N(\mu(k); \sigma^2(k))$ where $\mu(k)$ is the average auto-correlation and $\sigma^2(k)$ is the squared standard deviation. Defining $\delta_{ij}(k)$ as the difference between $C_{ij}(k)$ and the average $C(k)$, the observation can be considered statistically significant if $\frac{\delta_{ij}(k)}{\sigma(k)} > 3$. For the pairs we considered, this value is almost 10, far beyond 3. In Figure 4, for instance, the comparison between $\delta_{ij}(k)$ for the pair AAA-ATA and the standard deviation $\sigma(k)$ is shown.

	Chr20	Chr21
$2\gamma_I$	0.21	0.11
range of fit(tbp)	10^3 .. 10^5	10^3 .. 10^4

Table 3: *Power-law coefficients of the mutual information $I(k)$, and squared auto-correlation function $C^2(k)$ calculated on triplets for Chromosome 20 and 21.*

ATC	ATA
ATT	ATA
ATT	ATC
GGT	GGC
GGT	GGG
GGC	GGG

Table 4: *Nucleotide triplet couples for which $C(k)^2$ has a similar decay behavior to $I(k)$.*

	Chr21
$2\gamma_{C(k)_{ATC-ATA}}$	0.86
range of fit(tbp)	10^1 .. 4 $\times$ 10^2
$2\gamma_{C(k)_{ATT-ATA}}$	0.77
range of fit(tbp)	10^1 .. 3 $\times$ 10^2
$2\gamma_{C(k)_{ATT-ATC}}$	0.56
range of fit(tbp)	10^1 .. 7 $\times$ 10^2
$2\gamma_{C(k)_{GGT-GGC}}$	0.74
range of fit(tbp)	3 $\times$ 10^1 .. 5 $\times$ 10^2
$2\gamma_{C(k)_{GGT-GGC}}$	0.69
range of fit(tbp)	10^1 .. 8 $\times$ 10^2
$2\gamma_{C(k)_{GGT-GGG}}$	0.47
range of fit(tbp)	10^1 .. 2 $\times$ 10^3

Table 5: *Power-law coefficients of auto-correlation function $C(k)$ calculated on triplets pairs listed in Table 4 for Chromosome 21.*

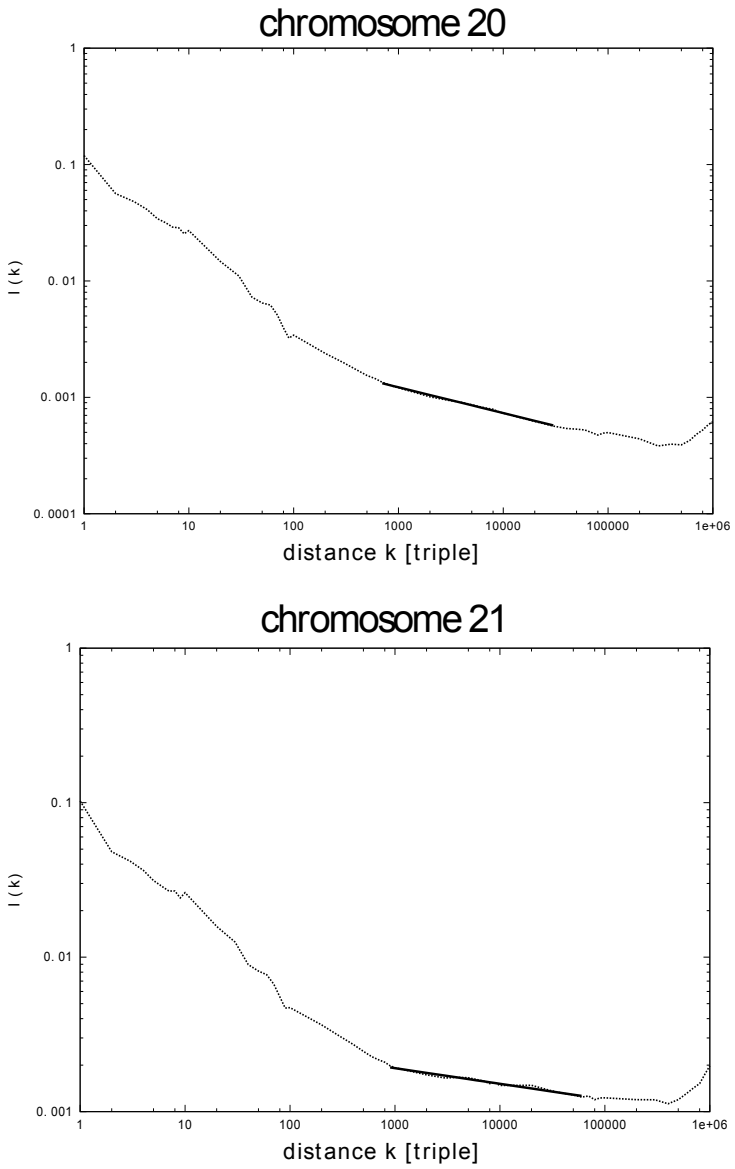

Figure 2: *LRC analysis on triplet representations of Chromosome 20 and 21.*

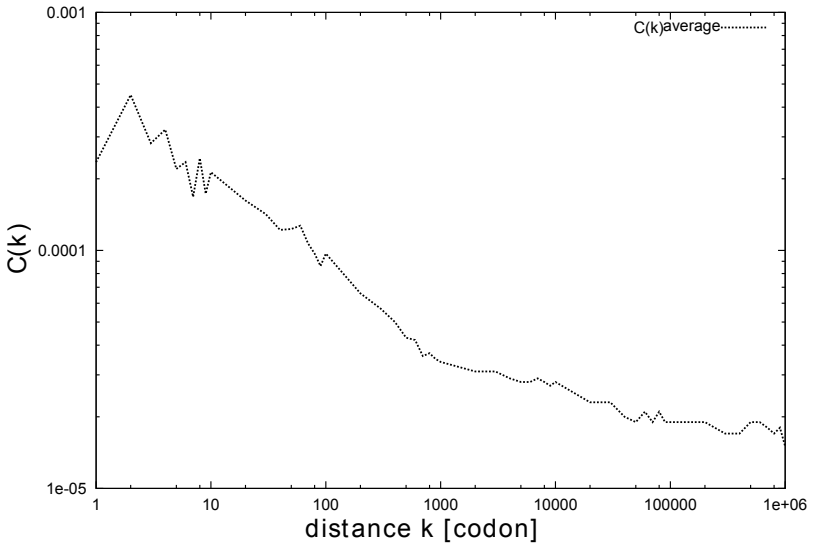

Figure 3: *Average auto-correlation function for chromosome 21*

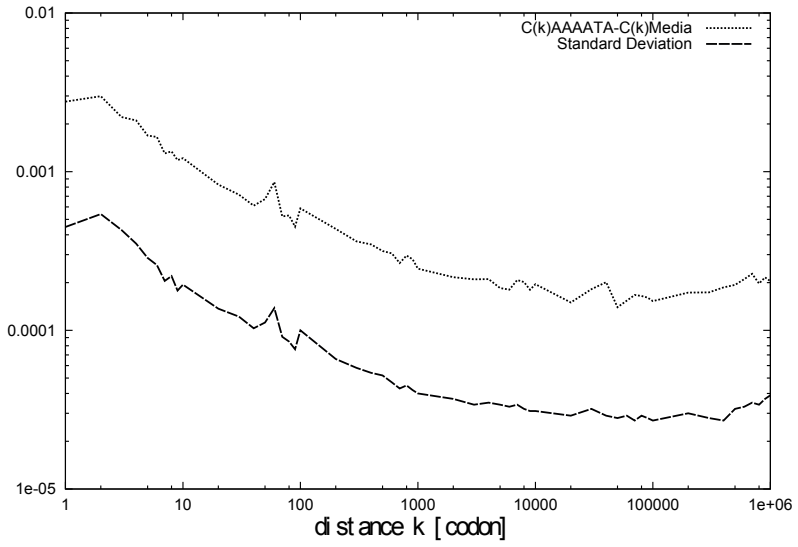

Figure 4: $\delta_{ij}(k)$ *for the pair AAA-ATA and the standard deviation* $\sigma(k)$

AAA, AAT	Lysine, Asparagine
AAA, ATA	Lysine, Isoleucine
AAA, ATT	Lysine, Isoleucine
AAA, TAA	Lysine, StopCodon
AAA, TAT	Lysine, Tyrosine
AAA, TTT	Lysine, Phenylalanine
AAT, ATA	Asparagine, Isoleucine
AAT, ATT	Asparagine, Isoleucine
AAT, TAT	Asparagine, Tyrosine
AAT, TTT	Asparagine, Phenylalanine
ATA, ATT	Isoleucine, Isoleucine
ATA, TAA	Isoleucine, StopCodon
ATA, TAT	Isoleucine, Tyrosine
ATA, TTA	Isoleucine, Leucine
ATA, TTT	Isoleucine, Phenylalanine
ATT, TAA	Isoleucine, StopCodon
ATT, TAT	Isoleucine, Tyrosine
ATT, TTT	Isoleucine, Phenylalanine
TAA, TAT	StopCodon, Tyrosine
TAA, TTT	StopCodon, Phenylalanine
TAT, TTA	Tyrosine, Leucine
TAT, TTT	Tyrosine, Phenylalanine
TTA, TTT	Leucine, Phenylalanine

Table 6: *Couples that have significantly higher $C(k)$ values, with respect to the overall average. The second column reports the respective aminoacid names.*

5 Conclusion

In this work, we have investigated statistical long range correlation analysis on nucleotide triplets and found LRCs in human Chromosomes 20 and 21, but not a clear LRC characteristic in Chromosome 22.

We have demonstrated that the triplets of nucleotides composed of Adenine and Thymine showed a significantly higher correlation in Chromosomes 20 and 21. It is important to underline that these chromosomes have a content of Adenine and Thymine greater than the content of Guanine and Cytosine, so a future direction is to verify if other chromosomes present similar LRC characteristics.

For further research, the most important challenge is to give a biological significance to the 14 distinct pairs of aminoacids (other than pairs including stop codons) which showed high LRC. It would also be interesting to clarify a relationship between generic nucleotide triplets in non coding zone and triplets in the coding part of DNA (real codons). If a relationship exists, it would be necessary to study the possible biological implications. It could help, for instance, to better understand how non-coding DNA is involved in evolutionary process. It was difficult to realize this on Chromosomes 20, 21, and 22 since these chromosomes contain a very low percentage of coding DNA. Hence, analysis on other chromosomes worth to be investigated and this remained as a future work.

It could be interesting to investigate these properties also on genomes of other species, because it could offer useful information to establish whether there is a universality in the long range dependence observed between nucleotide triplets.

References

Audit, B., Thermes, C., Vaillant, C., D'Aubenton-Carafa, Y., Muzy, J.F., A., & Arneodo. 2001. Long-range correlations in genomic DNA: A signature of the nucleosomal structure. *Physical Review Letters*, **86**(11), 2471–2474.

Beirer, S. 2003. *Modelling long-range Correlations in Genomic DNA sequences.* M.Phil. thesis, Technische Universitat Berlin.

Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., & Oliver, J.L. 2002. Study of statistical correlations in DNA sequences. *Gene*, 105–115.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simmons, M., & Stanley, H.E. 1995. Long range correlation

properties of coding and noncoding DNA sequences: GenBank analysis. *Physical Review E*, **51**(5), 5084–5091.

Cover, T. M., & Thomas, J. A. 2006. *Elements of Information Theory*. 2nd edn.

Herzel, H., & Grosse, I. 1995. Measuring correlations in symbol sequences. *Physica A*, **216**(4), 518–542.

Holste, D., Grosse, I., Beirer, S., Schieg, P., & Herzel, H. 2003. Repeats and correlations in human DNA sequences. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, **67**(6), 061913/1–061913/7.

Hurst, H.E. 1951. Long term storage capacities of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 776–808.

Hurst, H.E. 1956. Methods of using long-term storage in reservoirs. *Proceedings of the Institution of Civil Engineers*, **5**, 519–543.

Isohata, Y., & Hayashi, M. 2003. Power Spectrum and Mutual Information Analyses of DNA Base (Nucleotide) Sequences. *Journal of the Physical Society of Japan*, **72**(3), 735–742.

Karmeshu, & Krisnamachari, A. 2004. Sequence Variability and Long Range Dependence of DNA: An information theoretic perspective. *LNCS*, 1354–1361.

Li, W., & Holste, D. 2005. Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. *Phys. Rev. E*, **71**(4), 041910.

Li, W., & Kaneko, K. 1992. Long Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Non-coding DNA Sequence. *EPL(Europhysics Letters)*, **7**(17), 655–660.

Li, W., Marr, T.G., & Kaneko, K. 1994. Understanding long-range correlations in DNA sequences. *Physica D*, **75**, 392–416.

Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., M.Dudkiewicz, Laszkiewics, A., Dudek, M.R., & Cebrat, S. 2002. Replication associated mutational pressure generating long-range correlation in DNA. *Physica A*, **314**(25), 646–654.

Mandelbrot, B.B., & Ness, J.W. Van. 1968. Fractional Brownian motions fractional noises and applications. *SIAM Review*, **10**(4), 422–437.

Mandelbrot, B.B., & Wallis, J.R. 1968. Noah, Joseph, and operational hydrology. *Water Resources Research*, **4**(5), 909–918.

Messer, P.W., & Arndt, P.F. 2006. CorGen-measuring and generating long-range correlations for DNA sequence analysis. *Nucleic Acids Research*, **34**, 692–695.

Messer, P.W., Arndt, P.F, & Lässig, M. 2005a. Solvable Sequence Evolution Models and Genomic Correlations. *Phys. Rev. Lett.*, **94**(13), 138103.

Messer, P.W., Arndt, P.F., & Lässig, M. 2005b. Universality of long-range correlations in expansion randomization systems. *Journal of Statistical Mechanics:Theory and Experiment*, **2005**(10), P10004.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., & Stanley, H.E. 1992. Long-range correlations in nucleotide sequences. *Nature*, 168–170.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Tech. J*, 168–170.

Voss, R.F. 1992. Evolution of long-range fractal correlations 1/f noise and in DNA base sequences. *Physical Review Letters*, **68**(25), 3806–3809.

Yu, Z., Anh, V.V., & Lau, K. 2001. Measure representation and multifractal analysis of complete genomes. *Phys. Rev. E*, **64**(3), 031903.