LONG RANGE CROSS CORRELATIONS BETWEEN NUCLEOTIDE TRIPLETS IN HUMAN CHROMOSOMES

Davide Vitale, Ercan E. Kuruoglu and Osman Abul

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

ISTI-CNR, Pisa, Italy / TOBB, Ankara, Turkey
ercan.kuruoglu@isti.cnr.it

TOBB ETÜ
University of Economy and Technology

PROBLEM STATEMENT

We would like to answer:

➢Does the long range dependence seen in nucleotides generalise to nucleotide triplets and in particular to codons?

➢Are there long range cross correlations between different nucleotide triplets?

LONG RANGE DEPENDENCE

Long-range dependency (LRD) relates to the rate of decay of statistical dependence; decays more slowly than an exponential decay, typically a power-like decay.

Long range dependence was observed in teletraffic, hydrology and linguistics.

➢ Long range **auto**correlation has been observed in human chromosomes by several researchers:

- In intron containing genes and in untranscribed regulatory DNA sequences [Peng et al. 92]
- In non-coding DNA [Li and Kaneko. 92]
- LRCs analysis were also carried out to investigate properties of complete genomes

➢ The presence of LRCs is a fact, the origin is partially unknown

- Duplication of single mutations and deletion

➢Possibly related to the evolutionary process

- Messser et al (2005) showed that using models based on evolutionary process leads tosequences which reveal LRCs

➢ We calculate the long-range dependence between different pairs of nucleotide triplets

QUANTIFYING DEPENDENCE

DNA as sequence *S* of length *L*

$$\mathbb{A}=\{A,C,G,T\} \qquad \lambda=|\mathbb{A}|=4$$

Linear Dependence: Considering *k* as the distance in base pair between two nucleotides (at position i and j)

$$f_i \qquad\qquad p_i$$

$$f_{ij}(k) \quad\Longrightarrow\quad p_{ij}(k)$$

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

$$\vec{a}=<a_1,a_2,\dots,a_\lambda>$$

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

$$\vec{a}=<a_1,a_2,\dots,a_\lambda> \qquad \vec{b}=<b_1,b_2,\dots,b_\lambda>$$

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

➢Ideally we are interested in the mutual information

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

➢The Taylor expansion of mutual information function results in

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

Herzel and Grosse, 1995

➢The equation shows that mutual information is approximately proportional to squared correlation functions

- The contribution of $o(D_{ij}^3)$ can be considered negligible.

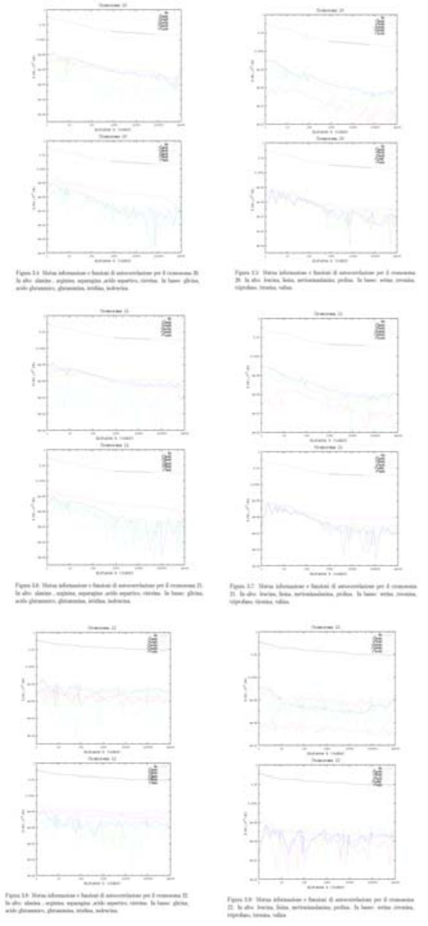
$$\mathbb{A}=\{AAA,AAT,\dots,TTT\} \qquad \lambda=|\mathbb{A}|=64$$

➢Considering genomes as sequences of nucleotide triples (codon)

- Important! We refer to generic nucleotide triplets, not necessarily corresponding to nucleotide triples that encode amino acids

➢Distance k is now expressed in triplets of base pair

➢We experimented for finding LRCs in chromosome 20, 21 and 22 of the human genome



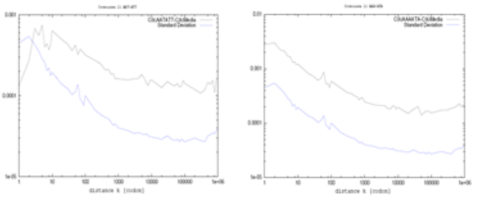
ANALYSIS RESULTS

➢We look for triplets that has significant LRC

➢An average Ck was computed also for the 2080 possible couples(AAA-TTT = TTT-AAA)

➢23 couples are found that showed significantly higher than average (more than 3 strandard deviations away from the mean)

➢All combinations of AAA, AAT, ATA, ATT, TAA, TAT, TTA, TTT



Comparison of average cross-correlation and a nucleotide triplet pair that is showing significant long range dependence

AAA,AAT	Lysine,Asparagine
AAA,ATA	Lysine,Isoleucine
AAA,ATT	Lysine,Isoleucine
AAA,TAA	Lysine,Stop
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

CONCLUSIONS

➢We observed significant long range correlations between 23 pairs of nucleotide triplet pairs

➢We developed a software "BioUtils" for the long range dependence analysis.

FUTURE WORK

➢ What is the biological significance of the results reported in this work?

➢Verifying whether other chromosomes present similar LRC Charateristics

➢Investigating on possible relationship between triplets in non-coding DNA and real codons

➢LRD analysis for different species