IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 4 Ver. II (Jul -Aug. 2014), PP 45-60
www.iosrjournals.org

What can we learn with Signal Analysis about Genomic information?

Amitabh Chaturvedi* and Archana Tiwari
School of Biotechnology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Airport Road, Bhopal- 462033 (M.P.), INDIA. Email: amitabh.2003@gmail.com, Tel: +91 755 2678873; Fax: +91 755 2678873

Abstract: *The novel biological information from genomic signal analysis for the conversion of genomic sequences into digital signals opens the chance to use powerful signal process strategies for handling genomic info. The study of advanced genomic signals reveals large-scale options, maintained over the size of whole chromosomes that will be tough to seek out by victimisation solely the symbolisation. supported genomic signal strategies and on applied math techniques, Herein, we tend to discuss the parameters of deoxyribonucleic acid sequences, that square measure invariant to transformations evoked by SNPs, junction or crossover by victimisation customary 'Wet' strategies of deoxyribonucleic sequence analysis, substantiated with IT techniques is illustrated, urging for analysis endeavours during in this direction. Therefore, the recently established sign cascades from freelance part Analysis is employed to characterize the variability shaping of the various strains isolated in several places square measure according from a proof analysis views.*

Keywords: *Amino Acid Sequence, Digital Signal Processing (DSP), Nucleotide Sequence, Independent Component Analysis (ICA), Nucleotide genomic signal (NGS), Single Nucleotide polymorphisms (SNPs)*

I. Introduction

The genomic signal conversion use during this work is simply a matched numeric illustration of the symbolic genomic sequences, chosen to be as very little biased as potential. For convenience, we tend to review within the next section the fundamentals of complicated genomic signal illustration. Genomic Signal Analysis virtually complete sequencing of the genomes of many eukaryotes, together with man (Homo sapiens) (J. Craig Venter 2001) and a number of other "model organisms" (C. Elegans 1998) has created the chance for comparative genomic analyses at scales starting from a personal individual gene or reference sequence to whole chromosomes. The general public access to most of this large quantity information of data of info offers associate degree new probability to information mine and explores thorough this extraordinary information installation to convert data into knowledge. The quality symbolical illustration of genomic info symbolic sequences of ester or amino acids has definite benefits in what issues storage, search and retrieval of genomic info, however limits the methodology of handling and process genomic info to pattern matching and applied mathematics analysis. Changing the polymer sequences into digital signals (Cristea 2002) opens the likelihood to use signal-processing strategies to the analysis of genomic information (Cristea 2001; Cristea 2002; Cristea 2003; Cristea 2004) . There are many tries to connect numerical values to the symbolic genomic sequences. The values correspond to bound properties of the element bases or of the encoded amino acids, within the case of the exons. The ensuing genomic signals develops interest for the study of the purposeful or structural options of polymer or super molecule, molecules which may be determined or influenced by the properties measured by the chosen signal values. As an example, the electro-ion-interaction potential has been used for the study of folding (Lalovic 1990; Cosic 1994) , whereas some values expressing the chance of committal to writing are accustomed acknowledge protein-coding regions in polymer sequences (Fickett 1982; Anastassiou 2000) . The analysis within the frequency domain, i.e., the Fourier remodel, has been for the most part utilized in of these approaches. Alternative approaches to the whole-genome analysis of prokaryotes, supported skew and integrated skew plots (Lobry 1996; Freeman 1998; Grigoriev 1998) are conferred in literature. The integrated skew diagrams look quite like the cumulated section curves within the case of being genomes, however the cumulated section given by (C. Elegans 1998) for the complicated illustration (Anastassiou 2000) could be a totally different perform than the skew functions (nG-nC)/(nG+nC), (nA-nT)/(nA+nT) or the similar ones. The cumulated section conserves all the data concerning the corresponding polymer strand, so a symbolic sequence is re-constructed from its cumulated section diagram. This can be not the case for the skew diagrams that extract every partial information, action bound specific properties of polymer strands. A similar holds true for the purine-excess or the keto-excess functions. On the opposite hand, the skew plots are created employing a window of an exact (arbitrary) breadth that limits the resolution of the analysis, whereas the cumulated section incorporates one-base resolution. The extremes of the integrated skew diagram are place in relation with the origins and termini of body replication. The genomic signal conversion within the gift work is simply associate degree objective (one-to-one) numeric illustration of the symbolic genomic sequences, chosen to be as very little biased as potential. This straightforward methodology has already tried its potential in revealing massive scale options of polymer sequences maintained over distances of 106 – 108 base pairs, together with each committal to writing and non-coding regions, at the size of whole genomes or chromosomes (Cristea 2002; Cristea 2003) . One among the foremost conspicuous results is that the unwrapped section of polymer complicated genomic signals varies virtually linearly on all investigated chromosomes, for each prokaryotes and eukaryotes. The dimensionality of the unwrapped section of polymer complicated genomic signals reveals large-scale regularities within the distribution of pairs of sequential nucleotides, like Chargaff's 1st order rules for the frequencies of incidence of nucleotides (Chargaff 1951) alongside the genes, the human and therefore the muscles genomes share doubly as long alternative extra–gene polymer sequences.

The study of pathway disruption is vital to understand cancer biology. Advances in high turnout technologies have led to the speedy accumulation of genomic information. The explosion in obtainable information has generated opportunities for investigation of combined changes that disrupt biological functions; this in turn created a desire for procedure tools for pathway analysis (Tsui 2007) .

Digital Signal process (DSP) applications in Bioinformatics have received nice attention in recent years, wherever new effective strategies for genomic sequence analysis, like the detection of writing regions, are developed. The use of DSP principles to analyze genomic sequences requires defining an adequate representation of the nucleotide bases by numerical values, changing the ester sequences into statistic. Once this has been done, all the mathematical tools typically utilized in DSP square measure employed in determination tasks like identification of super molecule coding DNA regions, identification of reading frames (Juan V. Lorenzo-Ginori 2009) .

Based on the mathematical analysis of complicated genomic signals, the review presents an outline of the foremost relevant applications of DSP algorithms within the analysis of genomic sequences, showing the most results obtained by mistreatment of these techniques, analyzing their relative blessings and downsides, and providing relevant examples. We have a tendency to finally analyze some views of DSP in Bioinformatics, considering recent analysis results on algebraically structures of the ordination, that recommend alternative new DSP applications during this field, still because the new field of Genomic Signal process. This additionally gift a model of the "patchy" longitudinal structure of body's associate in nursing shows that it derives from an supposed ancestral extremely ordered chromosome structure as a results of processes connected to species separation and protection at molecular level. The complicated genomic signal part may be connected to molecular potentials like the unbalanced element bonds of nucleotides. Such potentials will describe the interaction of a deoxyribonucleic acid section with proteins and with alternative deoxyribonucleic acid segments in processes like replication, transcription or crossover. Specially, this model will justify the functioning of deoxyribonucleic acid enzyme as a "Brownian machine" throughout replication, by the conversion of random molecular movements into an ordered gradual advance of the accelerator on the deoxyribonucleic acid strand. The speed of movement may be expressed as a function of the temperature and also the slope of the part.

II. Signal Analysis Learning

The conversion of ester sequences into digital genomic signals permits victimization signal-processing strategies for the analysis of genomic knowledge. Nucleotide genomic signal (NGS) analysis reveals stunning regularities within the distribution of nucleotides and pairs of nucleotides, in each prokaryotes and eukaryotes. These structural and applied mathematics restrictions of genomic sequences would be tough to spot by victimization of applied mathematics and pattern matching strategies, as in normal symbolic sequence analysis. NGS analysis is additionally economical within the analysis of native structural options, specifically for the study of infective agent variability. This can be vital for the molecular level detection of mutations that induce drug resistance, providing the practitioner with data required for a quick and correct call, and avoiding the length and big-ticket phonotypical clinical studies requesting infective agent culture (Cristea 2008) .

III. Genomic Signal Representation

The symbolic sequences are regenerated to advanced digital signals employing a mapping, reviewed here for convenience. The four nucleotides will be organized in categories consistent with the three main dichotomies in their organic chemistry properties**:**

3.1 Molecular structure: A and G are measure purines (R), whereas C and T are measure pyrimidines (Y);

3.2 Strength of link: bases A and T are measure connected by two hydrogen element bonds (W - weak bond), whereas C and G are measure connected by three hydrogen element bonds (S - sturdy bond);

3.3 Radical content: A and C contain the amino (NH_3) cluster (M class) within the major grove, whereas T and G the keto (C=O) cluster (K class). A vector tetrahedral illustration of the nucleotides (3) results on this basis.

By let alone the reduced amino-keto separation, the illustration will be brought during a plane, permitting victimisation the advanced illustration of nucleotides shown in Fig. 1 and expressed in Equations (i):

$$\begin{array}{ll}
a = 1+j, & \\
c = -1-j, & d = 1/3(1+j) \\
g = -1+j, & h = 1/3(1-j) \\
t = 1-j, & b = 1/3(-1-j) \\
 & v = 1/3(-1+j) \qquad \text{(i)} \\
w = 1 & \\
y = -j & \\
s = -1 & \\
r = j & \\
k = m = n = 0 &
\end{array}$$

Apart of the vectors like the four nucleotides (A, C, G, T), equations (i) and Fig. 1 conjointly comprise the IUPAC symbols for the mentioned nucleotide pair categories (S, W, R, Y, M, K), also as for the categories containing three nucleotides (B = = ¬A, D = = ¬C, H = = ¬G, V = = ¬T), or four nucleotides (N - case of such-and-such nucleotide). These symbols occur within the deoxyribonucleic acid sequences generated by the genotyping system delineate within the previous paragraph, owing to the multiplicities determined either by the variability among the virus population or by noise.

The genomic signals ensuing from the conversion of symbolic ester sequences into digital sequences victimisation the mapping given in Equations (i) and Fig. 1 are investigated by victimisation section analysis, nucleotide path analysis, independent component analysis, cluster analysis, phylogenetic analysis by neighbor-joining and most probability ways (Cristea 2005).

IV. Phase Analysis

The section of a fancy range could be an ambiguous magnitude with the amount 2π. The quality mathematical convention restricts the section to the domain (-π, π). For the of mapping in Equation (i), absolutely the values of all ester advanced representations square measure are identical, whereas the phases square measure given by:

$$\begin{array}{l}
a = 1 + j = \sqrt{2}\angle\frac{\pi}{4};\ g = -1 + j = \sqrt{2}\angle\frac{3\pi}{4}; \\
c = -1 - j = \sqrt{2}\angle\frac{-3\pi}{4};\ t = 1 - j = \sqrt{2}\angle - \frac{\pi}{4}
\end{array} \qquad \text{(ii)}$$

The cumulated section is that the add of the phases of the advanced numbers during a sequence from the primary entry up to this one. Within the case of Equations (i), the cumulated section is said to the amount of chemical element bases by:

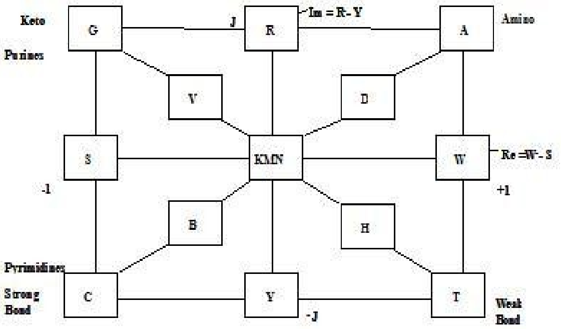

Fig. 1. Advanced illustration of nucleotide and nucleotide categories (IUPAC Symbols)

$$\theta_c = \frac{\pi}{4}\left[3\left(n_G - n_T\right) + \left(n_G - n_T\right)\right] \qquad \text{(iii)}$$

Where nA, nC, nG and nT metric weight unit and are measure of the numbers of A, cytosine, G and T nucleotides within the sequence, up to this location.

The unwrapped section is that the corrected section of the samples during a advanced sequence, such absolutely the worth of the distinction between the sections of any two neighbouring parts is formed smaller than π. This demand will forever be happy by adding or subtracting the acceptable multiple of 2π to or from the phase of this component. For the chosen advanced illustration of nucleotides (J. Craig Venter 2001), the positive transitions A→G, G→C, C→T, T→A confirm a rise of the unwrapped section with π/2 , correspondingly, the negative transitions A→T, T→C, C→G, G→A confirm a decrease with –π/2 , whereas all alternative transitions square measure neutral, so that

$$\theta_u = \frac{\pi}{2}\left(n_+ - n_-\right), \qquad \text{(iv)}$$

Where n_+ and n_- are measure the numbers of the positive and negative transitions, severally.

The cumulated section could be a signature of the signal reflective its 1st order statistics, whereas the unwrapped section could be a second order data point feature, determined by the distribution of ester pairs (Cristea 2004; Cristea 2004; Cristea 2004). Owing to the applied mathematics significance of the advanced illustration phases, it's generally convenient to precise the cumulated section not in radians, however in nucleotides (i.e., dividing (iii) with π/4), and therefore the unwrapped innovate pairs of nucleotides (i.e., dividing (iv) with π/2).

V. Nucleotide Path Analysis

Nucleotide path analysis is predicated on the development of the cumulated add of the vectors representing the nucleotides during a sequence, equally to the well-known Fresnel diagram utilized in optics. On victimisation the 3D illustration outlined in (Cristea 2002), the axes correspond to the numbers of weak minus sturdy bonds, amino minus keto, and purines minus pyrimidines, respectively:

$$x = n_w - n_s, y = n_M - n_K, z = n_R - n_Y \qquad \text{(v)}$$

The 2D (complex) illustration outlined in (i) results by the projection of the higher than 3D illustration on the y axis, so the x coordinate becomes the original component, whereas the z coordinate becomes the unreal element. These kinds of diagrams provide an international image of the nucleotide distribution during a deoxyribonucleic acid strand. The A, G, C and T axes are measure orienting towards the four corners of the square in Fig. 1. The nucleotide path diagrams are measure like the nucleotide walks, the excellence consisting within the selection of opposite ester species that is like the selection of the projection direction of the 3D vector path.

A third dimension, like the advancement on the deoxyribonucleic acid strand, will be value-added to the previous 2D nucleotide path. This 3D diagram shows the gradual accumulation of the variations of nucleotide species on the deoxyribonucleic acid sequence and permits a straightforward comparison of varied isolates of identical virus.

VI. Independent Element Analysis

A convenient thanks to reveal single nucleotide polymorphisms (SNPs) and alternative variability induced changes during a set of connected nucleotide sequences, and to ascertain potential links between such events is that the Independent Component Analysis (ICA). The set of nucleotide sequences might correspond to isolates of identical virus in numerous patients or in numerous phases of amendment beneath the combined choice pressure of the response and of treatment. ICA permits separating statistically freelance variations, therefore showing the links between at the same time occurring changes, generally at rather distant locations on the deoxyribonucleic acid strands. ICA assumes that a group of noticeable signals x1, x2, …, xn are measure linear mixtures of some indirectly accessible, however statistically freelance, supply signals s1, s2, …, sn . Equations of the form:

$$x_j = a_{j1}s_1 + a_{j2}s_2 + \Lambda + a_{jn}s_n = \sum_{k=1}^{n} a_{jk}s_k, \qquad \text{(vi)}$$

are assumed for all j = 1… n, with constant or virtually constant entries of the blending matrix entries ajk; j,k = 1… n.

The problem is that the "blind supply separation", i.e., finding (all or some of) the independent components sk, once knowing solely the mixed signals x1, x2… xn, however not the blending matrix entries. The answer is predicated on a hypothesis that infers a reciprocal of the Central Limit Theorem, i.e., the hypothesis that independent (non-mixed) signals ought to be non-gaussian (Hyvarinen 2001). Numerous measures of non-gaussianity or “contrast functions” will be used, like kurtosis, negentropy, mutual info, Kullback-Leibler divergence, or maximum likelihood estimation. The leads to this paper are obtained by victimisation the FastICA algorithmic program, a free (GPL) MATLAB package in the market available at http://www.cis.hut.fi/projects/ica/fastica/, which will implement a quick fixed-point algorithmic program for ICA and projection pursuit.

VII. Cluster Analysis

Cluster analysis will realize novel profile groupings however aren't designed to dependably reproduce groupings that are unit renowned from independent sources of knowledge. Therefore, there's generally an analytical challenge in understanding however the novel groupings relate to renowned grouping.

For work the dependableness of the ICA estimates, we have a tendency to performed a bunch analysis by mistreatment icasso, an interactive visual image technique and computer code package obtainable at (Himberg 2004). Its fundamental principle is to repeatedly run ICA formula, and visually estimate the bunch of the calculable independent elements within the signal house. Basically, the clusters ought to be compact, containing elements that area unit near one another and well separated from the remainder.

Classification approaches take renowned groupings and make rules for dependably assignment genes or conditions into these teams. These approaches usually need a collection of examples that the expression knowledge is related to some constitution or categorization.

VIII. Phylogenetic Analysis

The sequencing of deoxyribonucleic acid and protein has become simple and quick with the employment of machine-controlled tools. As sequence databases grow, they not solely offer information for man of science, however additionally challenge the computer engineers, mathematicians and statisticians to interpret this information. Equally searches and multiple alignments of sequences cause the question of finding the link between the sequences or maybe the organisms.

To analyze deoxyribonucleic acid sequences are visually inspected, aligned, gap stripped and valid mistreatment the multiple sequence editors BioEdit (Hall 1999). A pair wise distance matrix was generated with the DNADist program of the Phylip package (Felsenstein 1999). The tree topology was inferred by the neighbor-joining technique with the Fitch-Margoliash and least-square methods.

Analysis during this space will offer solely a sight into the divergence of molecules, however additionally into the progress of evolution.

IX. Genome Signatures

Analyses of deoxyribonucleic acid sequences from cultivated microorganisms have discovered genome-wide, taxa-specific nucleotide integrative characteristics, named as genome signatures. These signatures have sweeping implications for understanding genome evolution and potential application in classification of Meta genomic sequence fragments. However, very little is understood relating to the distribution of genome signatures in natural microorganism communities or the extent to that environmental factors form them (Gregory J Dick 2009).

As a matter of truth, the discrimination which might be achieved between eukaryotes and prokaryotes mistreatment, the genomic signature provides a genomic basis to the organic phylogenetic root of classification of species. Moreover, distances between sequences of phylogenetically shut species area unit of identical order because of the similar distances between subsequences of an order of those species. It has potential to look at that the distribution of word frequency broadly. Uncommon frequencies can be detected initially (Gregory J Dick 2009).

The genomic signature seems as a strong tool for monitoring the mechanisms of deoxyribonucleic acid maintenance from which the deoxyribonucleic acid structure results.

X. Genome Sequence Comparison with Practical Genomic

An important process technique for extracting the wealth of information hidden in human genomic sequence data is to match the sequence therewith from the corresponding region of the mouse genome, probing for segments that area unit preserved over evolutionary time. Moreover, the approach generalizes to comparison of sequences from any two connected species. The underlying principle (which is extravagantly confirmed by observation) is that a random mutation during a practical region is sometimes hurtful to the organism, and thus unlikely to become mounted within the population, whereas mutations during a non-functional region area unit absolve to accumulate over time. The potential worth of this approach is thus enticing that the general public and personal comes to sequence the human genome are unit currently turning to sequencing the mouse, and that we can shortly be ready to compare the human and mouse sequences of our favorite genomic region. We have a tendency to area unit presently witnessing an explosion of laptop tools for comparative analysis of two genomic sequences. Here the capabilities of two new network servers for comparison genomic sequences from any combine of closely connected species area unit sketched. The Synthetic Gene Prediction Program SGP-1 utilizes sequence comparisons to boost the flexibility to find protein-coding segments in genomic knowledge. PipMaker tries to see all preserved genomic regions, notwithstanding their perform (Thomas Wiehe 2000).

Mainly three digital signal processing methods are used: linear predictive coding, wavelet decomposition, and fractal dimension are studied to provide a comparative evaluation of the clustering performance of these methods on several microarray datasets. The results of this study show that the fractal approach provides the best clustering accuracy compared to other digital signal processing and well known statistical methods (Istepanian RS 2011).

This counsel the goal of genomics is to grasp the link between an organism's order and its constitution. The term genomics is usually used loosely to seek advice from the various potential approaches to understanding the properties and performance of the whole thing of an organism's sequences and gene product.

XI. DNA Walk

Graphical representations of deoxyribonucleic acid sequences area unit helpful as a result of the permit visual observations of nucleotide composition, nucleotide pair patterns, and sequence evolution. A deoxyribonucleic acid sequence consists of four bases: adenine (A), thymine (T), cytosine (C), and guanine (G). One necessary task within the study of genome sequences is to see densities of specific nucleotides and to grasp the implications for exon, or coding regions. Many ways for addressing this downside diagrammatically area unit is given in literature (Arneodo A. 1998).

The first step is to convert the four-letter genome alphabet into some numerical format, the way already projected in (Peng C. 1992; Arneodo A. 1998) which relatively plots the purine (A, G) and pyrimidine (C, T) content among deoxyribonucleic acid sequences by employing a two-dimensional deoxyribonucleic acid walk. Let us denote a deoxyribonucleic acid sequence of length $N\,as\,X = \{x[i]; i = 1,2,.....,N\}$. for a position k among this sequence. The value of the pyrimidine present is defined as $x[k] = +1$ whereas the purine present is defined as if a pyrimidine $x[k] = -1$. The deoxyribonucleic acid walk sequence of length $N\,as\,S = \{s[i]; i = 1,2,.....,N\}$, wherever for any position k we've accumulative total of the $x[i]\,for\,1 \le i \le k$ delineated by

$$s[k] = x\underset{i=1}{\overset{k}{i}}.$$

From the sequence S, observations are created relating to solely their relative content of purines and pyrimidines. Though this observation is enlightening, the deoxyribonucleic acid walk construct could also be extended to an additional helpful type by increasing the spatiality of the numerical deoxyribonucleic acid sequence (Peng C. 1992).

Suggesting that this technique applies to a very important category of sequence interactions, that is, transcriptional regulation via transcription factors (TFs) that bind to specific attention or silencer sites on deoxyribonucleic acid. The tactic addresses the question: “which of the genes during a genome are unit doubtless to be regulated by one or additional TFs with renowned deoxyribonucleic acid binding specificity?” It takes advantage of the actual fact that several TFs show cooperatively in transcriptional activation that manifests itself in closely spaced TF binding sites. Such "clusters" of binding sites area unit most unlikely to occur unintentionally alone, as hostile individual sites, that area unit typically superabundant each within the order and in promoter regions. applied math info regarding binding website clusters within the order, is complemented by info regarding (i) renowned biochemical functions of the TF, (ii) the structure of its binding website, and (iii) perform of the genes close to the cluster, to spot genes doubtless to be regulated by a given transcription issue.

XII. Signal Representability

Data scattering ratio

One of the issues that arise once operational with giant sets of knowledge, particularly once describing advanced systems or processes, or once generated by such systems or processes with a probably chaotic dynamics, is the way to represent information adequately. The ultimate understanding is just about all sets of knowledge or signals with human operators that the Graphical illustration, permits grasping quickly options hidden in piles of numerical information, is desirable. As shown within the previous sections, symbolic nucleotide sequences will be regenerated into digital genomic signals by mistreatment of the advanced (2D) quadrantal illustration derived from the tetrahedral (3D) illustration of nucleotides. The study of advanced genomic signals mistreatment signal process ways reveals giant scale options of chromosomes that will be troublesome to note by applying solely the applied mathematics or pattern matching ways presently utilized in the analysis of symbolic genomic information. On the opposite hand, within the context of operational with an outsized volume of knowledge at varied resolutions and visualizing the results to create them out there to humans, the matter of knowledge represent ability becomes crucial. Here, in the study we have shown an analysis of data representation based on the concept of data catering ratio for a pixel. Represent ability diagrams are derived for many typical cases of normal signals and for genomic signals. It is shown that the variation of genomic information on nucleotide sequences, specifically the cumulated and unwrapped section, will be unreal adequately as easy graphic lines for low and enormous scales, whereas for medium scales (thousands to tens of thousands of base pairs) the applied mathematics descriptions need to be used (Cristea 2002).

For convenience the essential ideas and notations to permit the understanding of the new theoretical and sensible results are measure delineated in figure. 2. Figure. 2 shows the plot as a line of the digital signal $s[i], i \in I^s = \{1,...,L\}$, where L is that the length of the sequence or sub-sequence to be pictured. An element is extracted and enlarged to permit examination absolutely for comparing the absolute value of the variation Vy of the signal for the set of samples pictured by the element with the element height Py measured in signal units. For the case in this figure, Vy < Py so the graphical illustration of the information by a line with the breadth of an element is adequate at that point and actually for the entire sequence. The scale of the screen in pixels is fixed e.g., the same old size Nx = 1024 by Ny = 768 pixels. To optimally use the screen to represent the information, the out there screen area should be fitted to the data: the horizontal screen size Sx in variety of samples must be chosen capable the length L of the sequence (or sub-sequence) to be pictured, whereas the screen vertical size Sy in information units should be chosen equal to the absolute value of the variation of the information within the pictured sequence:

$$S_y = \max_{i \in I^s}(s[i]) - \min_{i \in I^s}\left(s\left[i\right]\right). \qquad \text{(vii)}$$

Correspondingly, the horizontal and vertical element size is given in variety of samples and information units by:

$$P_x = \frac{S_x}{N_x}, P_y = \frac{S_y}{N_y}, \qquad \text{(viii)}$$

The variation of the information for the set of samples corresponding to an element is:

$$V_y(h) = \max_{i \in I_h^P}(s[i]) - \min_{i \in I_h^P}(s[i]), \qquad \text{(ix)}$$

Where $I_h^P = \{(h-1)P_x + 1, K, hP_x\}; h = 1, K, N_x$. As mentioned above, the adequacy of the illustration by an element of the set of Px data samples, it comprises can be characterised by ratio:

$$Q(h) = \frac{V_y(h)}{Py}, \qquad \text{(x)}$$

called the data scattering ratio for an element h. If $Q \le 1$, an element represents properly all the information samples it covers. Once all the elements within the line satisfy this condition, a line having the breadth of an element can represent the information adequately. If Q is below to two or three units for every element of the sequence fitted in the screen, a line may pictured the data properly; however the width of the line must correspond to the maximum value of Q. When Q is larger, the information is densely and quite uniformly distributed, so Q is almost same for all the elements and there aren't any outliers, the data will be represented adequately by one or two of lines showing the maximum and minimum values of the data for each element. Finally, if the information is scattered and there are outliers, this approach is not any longer sensible but the statistical description of data is more appropriate for its representation. The element will be thought-about a window of size Px (Cristea 2002). If the information distribution is shut enough to a standard distribution, the information will be delineating for every such window by the norm and also the variance. A line giving the norm and two others for delimiting some confidence interval expressed in terms of the quality deviation will be accustomed to represent the information. It is customary to represent solely the norm as a line, and to point out the boldness intervals by vertical segments for a set of the elements chosen with a definite periodicity on the road. Within the following we have a tendency to analyze the represent ability of many sorts of information and signals, as well as genomic signals, in terms of their represent ability characteristic

$$\tilde{Q} = \frac{\tilde{V}_y}{P_y} = f(P_x), \qquad \text{(xi)}$$

Where $\tilde{Q} = \tilde{V}_y / P_y$ is that the average information scattering magnitude relation for all the elements within the pictured line, with

$$\tilde{V}_y = \underset{h=1,k,N_x}{mean}\left(V_y(h)\right),$$

While Px is that the element horizontal size. While drawing the representability diagram showing the representability characteristic [5], logarithms within the basis a pair of are used for both abscissa and ordinate. Attributable to that, the element size $P_x^{(k)}$ are enlarged in an exceedingly geometrical scale with magnitude relation to:

$$P_x^{(1)} = 1, \ldots\ldots, P_x^{(k)} = 2^{k-1}, \ldots., P_x^{(k_{\max})} = 2^{k_{\max}-1}, \qquad \text{(xii)}$$

So, that the screen horizontal size $S_x^{(k)} = N_x P_x^{(k)}, k = 1, K, k_{\max}$, also will get double at every step for a hard and fast Nx. The amount of steps necessary to hide the full sequence of length L is $k_{\max} = \left\lceil \log_2 \frac{L}{N_x} \right\rceil + 1,$ where [x] denotes the smaller whole number larger than or equal to x. In this case, the most important screen equals or exceeds the length of the sequence. The amount of screens necessary to represent the entire sequence at step k is:

$$N_s^{(k)} = \left\lceil \frac{L}{S_x^{(k)}} \right\rceil = \left\lceil \frac{L}{N_x 2^{k-1}} \right\rceil \qquad \text{(xiii)}$$

If the length L of the sequence is not an influence of two, then the last screen at every step k, as well as the most important screen for the last step, may not be fitted to the information and can be excluded from the diagram. When L = 2^m and N_x = 2^s, all screens are horizontally fitted to the information and their variety $N_s^{(k)} = 2^{m-s+1-k} = 2^{k_{\max}-k}$ can type a geometrically decreasing sequence with magnitude relation 1/2, from $2^{\text{kmax-1}}$ to 1. On the opposite hand, every screen (window) is vertically fitted to the information, by selecting its vertical size equal to the absolute value of the variation of the information therein screen:

$$S_y^{(k)}(j) = \max_{i \in I_j^s}(s[i]) - \min_{i \in I_j^s}(s[i]), j = 1, K, j_{\max}^{(k)} \qquad \text{(xiv)}$$

Where $I_j^s = \left\{(j-1)S_x^{(k)} + 1, K, jS_x^{(k)}\right\}$ are the indices of the samples pictured within the screen j and $j_{\max}^{(k)} = N_s^{(k)}$ the amount of screens at step k.

A 3D diagram are accustomed showing the variation of the typical information scattering magnitude relation for the elements in every of the screens accustomed covers all the length of the sequence L at varied pixel sizes (Cristea 2002).

12. a. Monotonous signals

In the case of monotonously increasing signals, the relation (xiv) for the vertical size of screen j becomes:

$$S_y^{(k)}(j) = s\left[jS_x^{(k)}\right] - s\left[(j-1)S_x^{(k)} + 1\right], \qquad \text{(xv)}$$

So that the typical screen height results:

$$\tilde{S}_y^{(k)} = \underset{j=1K\,N_s^{(k)}}{mean}\left(S_y^{(k)}(j)\right) = \frac{1}{N_S^{(k)}} \sum_{j=1}^{N_S^{(k)}} \left(s\left[jS_x^{(k)}\right] - s\left[(j-1)S_x^{(k)} + 1\right]\right) \qquad \text{(xvi)}$$

Using this expression will be re-written as:

$$\tilde{S}_y^{(k)} = \frac{2^{k-1}N_x}{L}\left[s[L] - s[1] - \sum_{j=1}^{N_S^{(k)}-1}\left(s\left[jS_x^{(k)} + 1\right] - s\left[jS_x^{(k)}\right]\right)\right], \qquad \text{(xvii)}$$

Where the total contains signal variations between samples at distance one, sub-sampled with the step $S_x^{(k)}$. An identical expression holds for monotonously decreasing signals, so the typical screen height for monotonous signals results:

$$\tilde{S}_y^{(k)} = \frac{2^{k-1}N_x}{L}\left(s[L]-s[1]-\left(j_{\max}^{(k)}-1\right)mean\left(|d|\right)_{\downarrow S_x^{(k)}}\right) \qquad \text{(xviii)}$$

Where $mean\left(|d|\right)_{\downarrow S_x^{(k)}} = \underset{j=1,K,j_{\max}^{(k)}-1}{mean}\left(\left|d\left[j\,S_x^{(k)}\right]\right|\right) = \frac{1}{j_{\max}^{(k)}-1}\sum_{j=1}^{j_{\max}^{(k)}-1}\left|d\left[j\,S_x^{(k)}\right]\right|$ (xix)

is the average absolute variation of the signal between samples at distance one, down-sampled with the step $S_x^{(k)}$. Similarly, from (Cristea 2002) results the typical variation of the information for sets of samples reminiscent of pixels:

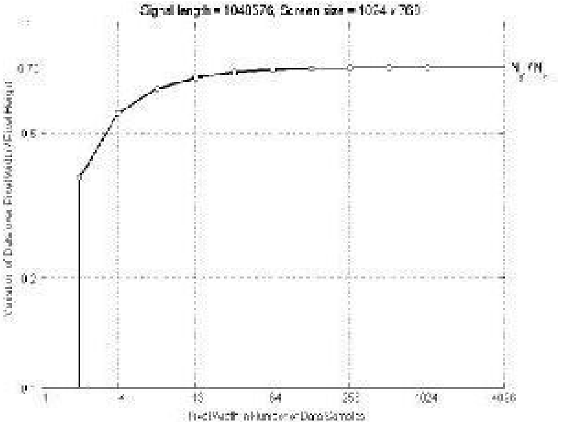

Fig. 2. Representability diagram for monotonous signals.

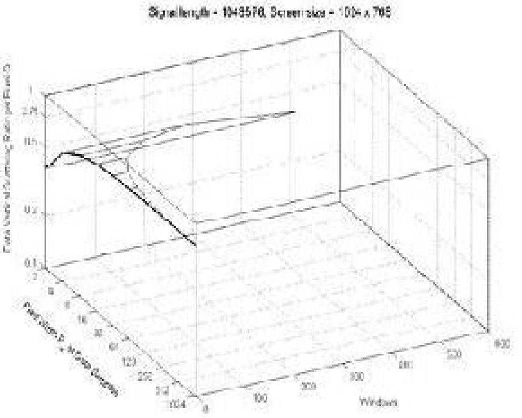

Fig. 3. 3D represent ability diagram for the signal in Fig. 2.

$$\tilde{V}_y^{(k)} = \frac{2^{k-1}}{L}\left(s[L]-s[1]-\left(h_{\max}^{(k)}-1\right)mean\left(|d|\right)_{\downarrow P_x^{(k)}}\right), \qquad \text{(xx)}$$

Where

$$mean\left(|d|\right)_{\downarrow P_x^{(k)}} = \underset{h=1,K,h_{\max}^{(k)}-1}{mean}\left(\left|d\left[h\,P_x^{(k)}\right]\right|\right) = \frac{1}{h_{\max}^{(k)}-1}\sum_{h=1}^{h_{\max}^{(k)}-1}\left|d\left[h\,P_x^{(k)}\right]\right| \qquad \text{(xxi)}$$

is the average absolute variation of the signal between samples at distance one, down-sampled at the pixel step $P_x^{(k)}$:

As a consequence, the average data scattering ratio for a monotonous signal is given by the equation:

$$\tilde{Q}^{(k)} = \frac{\tilde{V}_y^{(k)}}{\tilde{P}_y^{(k)}} = \frac{N_y}{N_x} \frac{s[L]-s[1]-\left(N_P^{(k)}-1\right) mean\left(|d|\right)_{\downarrow P_x^{(k)}}}{s[L]-s[1]-\left(N_S^{(k)}-1\right) mean\left(|d|\right)_{\downarrow S_x^{(k)}}}, \qquad \text{(xxii)}$$

where $N_P^{(k)}$ is the total number of pixels to represent the sequence s[i], i = 1, ..., L, for an horizontal pixel size $P_x^{(k)} = 2^{k-1}, N_S^{(k)} = N_P^{(k)} / N_x$ is the total number of screens necessary to represent the data at resolution k, and mean $\left(|d|_{\downarrow D}\right)$ is the average of the absolute values of the signal variation between successive samples d[i] = s[i+1] - s[i], down sampled with step D. As long as the sampling density is high enough,

$$mean\left(|d|\right)_{\downarrow S_x^{(k)}} \approx mean\left(|d|\right)_{\downarrow P_x^{(k)}} \approx \frac{s[L]-s[1]}{L-1}, \qquad \text{(xxiii)}$$

So that equation (xxii) becomes:

$$\tilde{Q}^{(k)} = \frac{N_y}{N_x} \frac{P_x^{(k)}-1}{P_x^{(k)}-1/N_x} \qquad \text{(xxiv)}$$

From (xxiv) it results that all monotonous signals have almost the same represent ability characteristic drawn in Fig. 3 as a line. The circles correspond to experimental data for various monotonous signals like linear, parabolic of various degrees, logarithmic and exponential of various bases etc. Monotonous signals are practically a best case in what concerns the representability characteristic. As results from (xxiv) and from the experimental data in Fig. 2, for large values of pixel width P_x, the representability characteristic tends asymptotically towards the aspect ratio of the screen:

$$\tilde{Q}^{(k)} \xrightarrow[2^{k-1} \square 1]{} \frac{N_y}{N_x}. \qquad \text{(xxv)}$$

Fig. 3 gives a 3D representability diagram for a monotonous signal of L = 2 (Wood D.W. 2001) samples as the one analyzed in Fig. 2. The average data vertical scattering ratio per pixel Q is shown for all the $N_w = L / (P_x N_x)$ windows (screens) of N_x pixels in which the sequence is divided for each pixel width P_x. A 2D representability diagram $\tilde{Q} = f(P_x)$, like the one in Fig. 2, which gives the average for all the windows for a certain P_x, is also shown on the left vertical plane. It can be noticed from Fig. 3 that the average data vertical scattering ratio per pixel is the same for all the windows in which the signal is split at a certain resolution given by the number of samples in a pixel (Cristea 2002).

12. b. Uniformly distributed random signals

The theoretical worst case from the representability purpose of read is associate hypothetic signal that the variation between two successive samples is equal the screen height. A sensible worst case is provided by a random signal uniformly distributed on the screen height. The representability characteristic may be found in closed kind for this case. the typical variation of the information for the set of samples reminiscent of a component, i.e., The typical of the distinction between the biggest and therefore the smallest values of the samples within the set of random variables uniformly distributed across the screen height expressed in pixels, without delay from the relation (A3) within the annex:

$$\tilde{Q}^{(k)} = \frac{\tilde{V}_y^{(k)}}{\tilde{P}_y^{(k)}} = N_y \frac{P_x^k - 1}{P_x^k + 1}. \qquad \text{(xxvi)}$$

The representability characteristic is shown in Fig. 4.The line has been computed analytically exploitation the equation (xxvi), whereas the circles represent information from a town experiment that simulates the uniform distribution of the samples in an exceedingly vary adequate the screen height in information units. For big values of the component dimension, they represent ability characteristic asymptotically approaches Ny – the vertical size of the screen in pixels:

$$Q^{(k)} \xrightarrow[2^{k-1}\square 1]{} N_y \qquad \text{(xxvii)}$$

Fig. 5 gives the 3D represent ability diagram for the same uniformly distributed random signal of L = 2 (Wood D.W. 2001) samples. Notice the small fluctuations of the average data vertical scattering ratios per pixel in the various windows for each pixel width Px. The 2D represent ability diagram in Fig. 4, which gives the average for all the windows at a certain Px, is also shown on the left vertical plane of Fig. 5.The monotonous signals and the uniformly distributed random signal provide the practical best case and the practical worst case for the represent ability of signals. To illustrate the behavior of non-monotonic signals in (Cristea 2004) are given the represent ability characteristics of several functions with period's form 2 (Cristea 2004) to 2 samples (Kawai J. 2001). As expected, the several signals behaves as a monotonous signal the best case, when its period is longer times the width of the screen in number of samples, and as the worst case, when the period is lower than the width of the pixels. Two aliasing effects occur in the vicinity of the limiting cases, at levels of the average data scattering ratio equal to twice the best case and half the worst case, respectively. In-between these two levels, the average data scattering ratio vary almost linearly with respect to the pixel width.

12. c. Represent ability of the phase of genomic signals

Aspects related to the graphical representation of the cumulated and the unwrapped phase of genomic signals can be found in (Cristea 2004). In Fig. 6 is given the average data scattering ratio for the 6,311,978 base pairs long contig of the first chromosome of Homo sapiens NT004424 (National Center for Biotechnology Information ; J. Craig Venter 2001; Consortium 2001) analyzed in Fig. 5. The results are typical for many other prokaryotes and eukaryotes genomic signals. The screen size has been considered 1024 x 768 pixels. In the special case of one pixel per sample, for which the variation inside a pixel is zero, the scattering ratio cannot be represented on the logarithmic plot. This case corresponds to an error free graphic, disregarding the smoothness of the resulting line. For pixels comprising two samples and up to about 16 samples, i.e., for DNA segments comprising up to 16384 base pairs, both the cumulated and the unwrapped phase have an average data scattering ratio in the range 5...8, so that the data should be presented taking into account their dispersion. In most cases, this can be done by tracing a couple of lines showing the minimum and maximum values, respectively. When there are only several points apart from the others, the representation can be made by a line corresponding to the typical value in a window with the dimension of a component, among two alternative lines delimiting some confidence interval outlined in terms of the signal dispersion. The remarkable feature for the analysis of enormous scale deoxyribonucleic acid options is that the indisputable fact that the common vertical scattering quantitative relation of the signal on a component becomes less but the important value one, i.e., the variation of the signal for the set of samples described by a component becomes less that the component height, once the component dimension is larger than concerning 1450 samples. The cumulated phase analysis displays a comparatively tiny variation and, once described independently, remains with a rather vital dispersion of the samples that needs a presentation kind of like the one used for applied mathematics data.

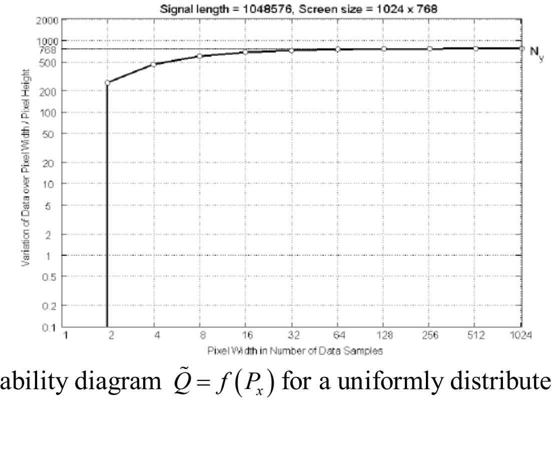

Fig. 4. Representability diagram $\tilde{Q} = f(P_x)$ for a uniformly distributed random signal.

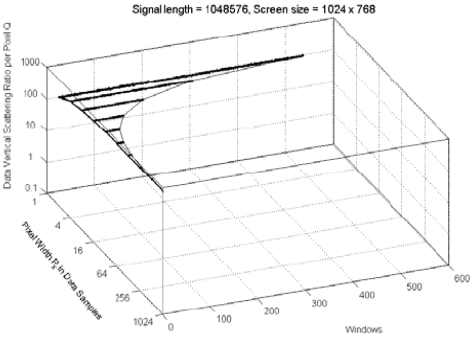

Fig.5. 3D Representability diagram $\tilde{Q} = f(P_x)$ for a uniformly distributed random signal.

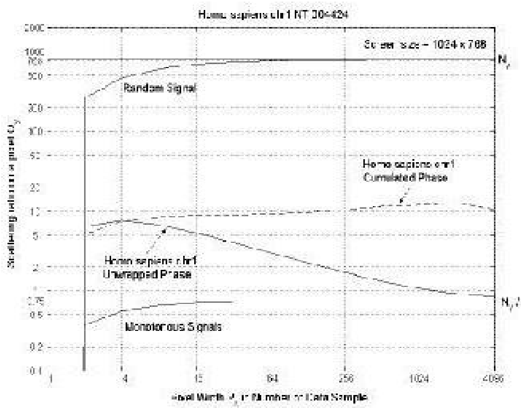

Fig.6. Representability diagram $\tilde{Q} = f(P_x)$ for the cumulated and unwrapped phase of the contig NT004424 (National Center for Biotechnology Information ; J. Craig Venter 2001; Consortium 2001) of Homo sapiens chromosome 1.

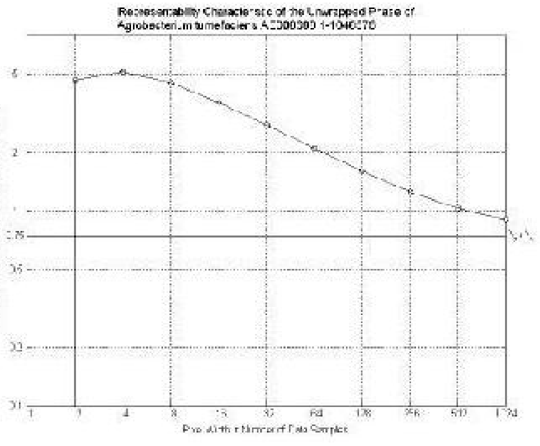

Fig. 7. 2D representability diagram of the unwrapped phase of the genomic signal for 1st 1048576 bp from the circular chromosome of Agrobacterium tumefaciens (AE008688) (Wood D.W. 2001).

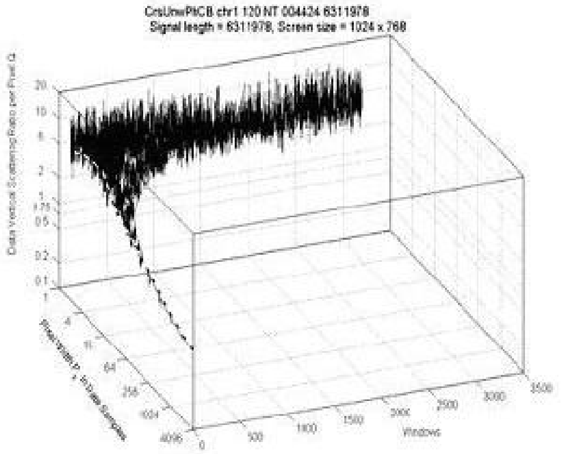

Fig. 8. 3D Representability diagram of the Unwrapped phase of the Homo Sapiens chr. 1 cont. NT004424 also analysed in Fig. 6.

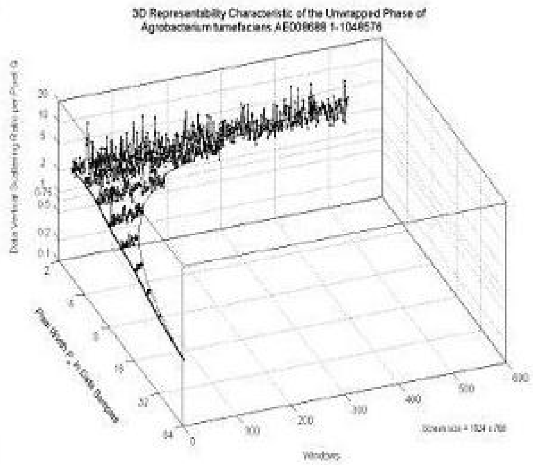

Fig. 9. 3D Representability diagram of the unwrapped phase of the genomic signal in Fig. 7.

The genomic signal cumulated phase and unwrapped phase will be place in relation with the distribution of bases and base-pairs, severally. It is shown that the reorientation of deoxyribonucleic acid segments involves the synchronic reversal of the order and also the complementing of the nucleotides (A with T and C with G) within the inverse coding regions. Applied mathematics scale regularity and applied mathematics structure of the ester distribution mirrored within the nearly piece-wise linear variation of the cumulated section for prokaryotes appears to point a particular biological function at the size of whole chromosomes. This feature is typical for the chromosomes, each linear and circular, whereas the plasmids don't have it. The regularity shown by the nucleotide sequences obtained when concatenating the reoriented coding regions suggests the existence of a reputed primary ancestral genomic material having a quite uniform massive scale applied mathematics structure. This feature is additionally determined solely within the chromosomes, and isn't found within the plasmids. This review presents additionally some new results on data representability, primarily applied for genomic data. The cumulated section and unwrapped section are often depicted adequately as easy graphic lines for terribly low and huge scales, whereas for medium scales (thousands to ten thousands base pairs) applied mathematics descriptions have to be compelled to be used (Wood D.W. 2001).

Application of digital signal processing:

The application of Digital Signal process in Genomic Sequence Analysis has received nice attention within the previous few years, providing a brand new insight within the resolution of varied issues like

- Detection of coding regions in genomic sequences
- Based on spectral analysis.
- Reading frame identification.

- Detection of periodicities in genomic sequences,
- Detection of CpG islands,
- Detection of palindromes,
- Finding various signals and options in genomic sequences.
- Studies on proteins,

On the opposite hand, the most DSP tools that have found application in these fields are

- Digital filters (IIR, FIR),
- Discrete transforms (Fourier, Cosine, Walsh Hadamard, Wavelet),
- Parametric models (mainly autoregressive),
- Information Theory concepts (entropy),
- Fractals.

Other recursive tools that are applied in Bioinformatics though not self-addressed during this paper are thought-about typically as neighboring areas. This is the case of Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Fuzzy Sets and Genetic Algorithms.

A recent development closely associated with the impact of DSP on Bioinformatics is that the new field of Genomic Signal processing (GSP). An early survey on this could be found in (Dougherty E.R. 2005). A proper definition of GSP was given by (Schonfeld D 2008) as "the analysis, processing, and use of genomic signals for gaining biological information and also the translation of that data into systems-based applications" (Qiu P 2007). Remark this interest in exploiting the DSP strategies to get data from genomic and proteomic information to create models of molecular biological systems. This may permit getting a deeper understanding of the structure and functions of living systems and can facilitate in developing new diagnostic tools, therapeutic procedures and pharmacologic drugs. Associate application example in cancer classification and prediction can be seen in (Chen J 2003).

Finally, it is attention-grabbing to note that Bioinformatics is additionally having associate influence on new developments, as are seen in (Tsaftaris SA 2004) (Janevski A 2012). Characterizing copy-number variation is a basic method to profile normal and diseased tissue samples, but challenges remain in accurately interpreting the data from a single genome and comparative measurements from groups of sequenced genomes (Huang H 2012), these studies several variations of copy-number analysis approaches to assess the significance and impact of each methodology choice (Doderer M 2012), proposes C2Maps platform as an online bioinformatics resource providing biologists with directional relationships between drugs and genes/proteins in specific disease contexts based on network mining, literature mining, and drug effect annotations (Evani US 2012), analyzes pathway consolidation approaches and provides a user-friendly web-accessible tool that can enable users to extract functional relations of genes across multiple pathway databases. A case study of performing personal genome analysis on a cloud computing environment is presented in (Max A. 2013). The approach can assist researchers in applying existing cloud computing technologies to analyze enormous amount of data generated by next-generation sequencing technologies.

Recent advances in our ability to observe the molecular and cellular processes of life in action such as atomic force research, optical tweezers and forster light resonance energy transfer raise challenges for digital signal processing (DSP) of the ensuing experimental information. This text explores the distinctive properties of such biophysical statistic that set them aside from different signals, like the prevalence of abrupt jumps and steps, multi-model distributions and auto correlate noise. It exposes the issues with classical linear DSP algorithms applied to the current reasonably information, and describes new nonlinear and non-Gaussian algorithms that are ready to extract data that is of direct connection to biological physicists. It is argued that these new strategies applied in this context typify the emerging field of biophysical DSP. Useful experimental examples are provided in (Max A. 2013).

An outline of the foremost relevant applications of DSP algorithms within the analysis of genomic sequences, showing the most results obtained by exploitation these techniques, analyzing their relative benefits and downsides, and providing relevant examples. We tend to finally analyze some views of DSP in Bioinformatics, considering recent analysis results on pure mathematics structures of the ordering, that recommend alternative new DSP applications in this field, in addition as this field of Genomic Signal processing.

Software Challenges:

We would like to style such novel software system with the assistance of that all the mathematical models are going to be analyzed within the same software package and chance of comparison among them is additionally possible. The most expression of this software system coming up with is to see the accuracy of the results that were analyzed by numerous software's. The complete situation of the ordering signal analysis is obtained in brief time and within the same platform. The results are often taken with any kind of biological systems.

XIII. Conclusion

The conversion of genomic sequences into digital genomic signals offers the chance to use powerful signal processing strategies for the analysis of genomic data. The study of genomic signal reveals native and world options of chromosomes that may be troublesome to spot by exploiting solely the symbolization utilized in genomic databases. Independent component analysis has been accustomed to characterize the variability of the strains.

Moreover, a range of comparative maps of the chromosomal locations of orthologous human and mouse genes are offered on the World-Wide Web. These permit biological information for one species to be readily extrapolated to the other. The unexpected availableness of genomic sequence information opens the possibility of higher-resolution genomic comparisons right all the way down to the ester level. A tiny low range of websites are presently offered to assist a life scientist seeking to check homologous (i.e. orthologous or paralogous) genomic sequences. In such cases, it should be fruitful to check one among the class sequences with the sequence from a lot of distant vertebrate, like a bird or a fish. Conversely, in quickly evolving regions it will happen that the sole detectable human–mouse matches are in protein-coding regions. Below those conditions, it could be potential to seek out, say, restrictive regions by comparison sequences of two nearer species, like human and monkey or mouse and rat.

Another tool for analyzing genome sequences as separate sets of numbers, with every component representing a nucleotide deals with a large set of data; an intuitive transformation of characters to numbers has been performed. Consequently, the sequences are often thought of as discrete-time signals and studied exploitation typical digital signal processing approaches. The numerical assignments powerfully have an effect on the results of the analysis. The tool is predicated on associate improvement of the deoxyribonucleic acid walk tested on many deoxyribonucleic acid sequences and also the results are verified to match results reportable within the literature.

On the opposite hand deoxyribonucleic acid sequences will be regenerated into genomic signals by employing a nucleotide advanced illustration derived from the nucleotide tetrahedral illustration. The genomic signal cumulated phase and unwrapped phase will be place in relation with the statistical distribution of bases and base pairs, respectively. It is shown that the reorientation of deoxyribonucleic acid segments involves the synchronic reversal of the order and therefore the complementing of the nucleotides (A with T and C with G) within the inverse committal to writing regions. The big scale regularity and applied mathematics structure of the ester distribution mirrored within the virtually piece-wise linear variation of the cumulated section for prokaryotes appears to point an explicit biological perform at the dimensions of whole chromosomes. This feature is typical for the chromosomes, each linear and circular, whereas the plasmids do not have it. The regularity shown by the nucleotide sequences obtained on concatenating the reoriented coding regions suggests the existence of a reputed primary ancestral genomic material having a quite uniform large-scale applied mathematics structure. This feature is additionally determined solely within the chromosomes, and isn't found within the plasmids. This presents conjointly some new results on information represent ability, primarily applied for genomic information. The cumulated section and unwrapped section are often drawn adequately as easy graphic lines for terribly low and huge scales, whereas for medium scales (thousands to ten thousands of base pairs) applied mathematics descriptions have to be compelled to be used.

Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgement

Authors would like to thank to anonymous reviewers for their helpful comments and suggestions.

Reference

[1]. Anastassiou, D. (2000). "Frequency-domain analysis of bimolecular sequences." Bioinformatics **12**: 1073-1081.
[2]. Arneodo A., e. a. (1998). "what can we learn with wavelets about DNA sequences?" Physica A(249): 439 - 448.
[3]. C. Elegans (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-2018.
[4]. Chargaff, E. (1951). "Structure and function of nucleic acids as cell constituents " Fed. Proc. **10** 654-659.
[5]. Chen J, L. H. (2003). "How will bioinformatics impact signal processing research?" IEEE Sign Proc Mag **6**: 106-126.
[6]. Consortium, I. H. G. S. (2001). "Initial sequencing and analysis of the human genome " Nature **409** 860-911.
[7]. Cosic, I. (1994). "Macromolecular bioactivity: is it resonant interaction between macromolecules-Theory and application." IEEE, Trans Biomed Eng **41**: 1101-1114.

[8]. Cristea, P. D. (2001). Genetic Signal Analysis, Proc. of ISSPA 2001. The Sixth International Symposium on Signal Processing and its Applications. Kuala Lumpur, Malaysia**:** 703-706.

[9]. Cristea, P. D. (2002). "Conversion of Nitrogenous Base Sequences into Genomic Signals." Journal of Cellular and Molecular Medicine **6**: 279-303.

[10]. Cristea, P. D. (2002). Large Scale and Global Features of Complex Genomic Signals, SFM-02 Saratov Fall Meeting, Saratov, Russia Proceedings of SPIE, Optical Technologies in Biophysics and Medicine.

[11]. Cristea, P. D. (2002). Whole Chromosome Features of Genomic Signals. 6th Seminar on Neural Network, Applications in Electrical Engineering, Belgrade, Yugoslavia.

[12]. Cristea, P. D. (2003). "Large Scale Features in DNA Genomic Signals." Signal Processing **83**(Special Issue on Genomic Signal Processing): 871-888.

[13]. Cristea, P. D. (2004). Genomic signals of chromosomes and of concatenated reoriented coding regions SPIE Conference, BIOS 2004 San Jose, CA, USA, Proceedings of SPIE

[14]. Cristea, P. D. (2004). "Genomic Signals of Re-Oriented ORFs, EURASIP." Journal on Applied Signal Processing, Special Issue on Genomic Signal Processing **1**: 132-137.

[15]. Cristea, P. D. (2004). Invariants of DNA Genomic Signals . SPIE International Symposium, AU04 - Smart Materials, Nano, and Micro-Smart Systems, AU104 Biomedical Applications of Micro and Nanoengineering II, Biocomputation & Biomodelling Sydney, Australia.

[16]. Cristea, P. D. (2004). Multiresolution Phase Analysis of Genomic Signals. ISCCSP 2004 - 1st International Symposium on Control, Communications and Signal Processing, Signal Processing in Biological Sciences. Hammamet, Tunisia**:** 743-746.

[17]. Cristea, P. D. (2005). Genomic signal analysis of HIV variability Proceeding of SPIE, Photonic West San Jose CA

[18]. Cristea, P. D. (2008). Nucleic acid structural properties identified by genomic signal analysis. Proceedings of the 9th WSEAS International Conference on Mathematics & Computers in Biology & Chemistry, Bucharest, Romania.

[19]. Doderer M (2012). "Pathway Distiller - multisource biological pathway consolidation." BMC Genomics **13**((Suppl 6)): S18.

[20]. Dougherty E.R., e. a. (2005). "Research issues in genomic signal processing." IEEE Sign Proc Mag, **22** 46-48.

[21]. Evani US, e. a. (2012). "Atlas2 Cloud: a framework for personal genome analysis in the cloud. ." BMC Genomics **13**((Suppl 6)): S19.

[22]. Felsenstein, J. (1999). PHYLIP. D. o. G. S. D. o. Biology, http://evolution.genetics.washington.edu/. **University of Washington**.

[23]. Fickett, J. W. (1982). "Recognition of protein coding regions in DNA sequences " Nucleic Acids Research **17**: 5303- 5318.

[24]. Freeman, J. M. (1998). "Patterns of Genome Organization in Bacteria." Science **279**: 1827-1832.

[25]. Gregory J Dick, A. F. A., Brett J Baker, Sheri L Simmons, Brian C Thoma, A Pepper Yelton and Jillian F Banfield (2009). "Community-wide analysis of microbial genome sequence signatures." Genome Biology,(10): R85.

[26]. Grigoriev, A. (1998). "Analyzing genomes with cumulative skew diagrams " Nucleic Acids Research **10**: 2286-2290.

[27]. Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows. Nucl. Acids. Symp, .

[28]. Himberg, J., Hyvarinen, A., Esposito, F. (2004). "Validating the independent components of neuroimaging time series via clustering and visualization." Neuroimage **22**(3): 1214-1222.

[29]. Huang H (2012). "C2Maps: a network pharmacology database with comprehensive disease-gene-drug connectivity relationships." BMC Genomics **13**((Suppl 6)): S17.

[30]. Hyvarinen, A. e. a. (2001). "Independent Component Analysis." Nature **409**: 860-911.

[31]. Istepanian RS, S. A., Nebel JC. (2011). "Comparative analysis of genomic signal processing for microarray data clustering " IEEE Trans Nanobioscience, **10**(4): 225-238.

[32]. J. Craig Venter, M. D. A., Eugene W. Myers, (2001). "The Sequence of the Human Genome." Science **291**: 1304-1351.

[33]. Janevski A, e. a. (2012). "Effective normalization for copy number variation detection from whole genome sequencing " BMC Genomics **13**((Suppl 6)): S16.

[34]. Juan V. Lorenzo-Ginori, A. R., Ricardo Grau Ábalo and Robersy Sánchez Rodríguez (2009). "Digital Signal Processing in the Analysis of Genomic Sequences." Current Bioinformatics

[35]. **4**: 28-40.

[36]. Kawai J., e. a. (2001). "Functional annotation of a full-length mouse cDNA collection." Nature **409** (6821): 685-690.

[37]. Lalovic, D. a. V., V. (1990). "The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential." Biosystems **23**: 311-316.

[38]. Lobry, J. R. (1996). "Asymmetric substitution patterns in the two DNA strands of bacteria." Molecular Biology and Evolution **13**: 660-665.

[39]. Max A., e. a. (2013). "Signal processing for molecular and cellular biological physics: an emerging field." Phil. Trans. R. Soc. **A 371**: 20110546.

[40]. National Center for Biotechnology Information. "National Institutes of Health, National Library of Medicine, ." National Center for Biotechnology Information, .

[41]. Peng C., e. a. (1992). "Long-range correlations in nucleotide sequences." Nature **356**: 168 - 170.

[42]. Qiu P, e. a. (2007). "Genomic Processing for Cancer Classification and Prediction " IEEE Sign Proc Mag **24** 100-110.

[43]. Schonfeld D, G. J. (2008). "Introduction to the Issue on Genomic and Proteomic Signal Processing " IEEE J Select Topics Signl Proc **2**: 257-259.

[44]. Thomas Wiehe, R. G., Webb Miller (2000). "Genome sequence comparisons:Hurdles in the fast lane to functional genomics." Henry Stewart publications 1467-5463. Briefings in bioinformatics **4**: 381-388.

[45]. Tsaftaris SA (2004). "How Can DNA Computing be Applied to Digital Signal Processing? ." IEEE Sign Proc Mag **21** 57-61.

[46]. Tsui, I. F., Chari, R., Buys, T. P., Lam, W. L., (2007). "Public databases and software for the pathway analysis of cancer genomes." Cancer Inform **3**: 379-397.

[47]. Wood D.W., e. a. (2001). "The Genome of the Natural Genetic Engineer Agrobacterium tumefaciens C58." Science **294**(5550): 2317-2323.