Journal of ISOSS
2020 Vol. 6(2), 40-44

ENCRYPTION OF BIG DATA VIA Z-DNA TO ANALYSIS OF B-DNA THROUGH NEURON PROCESSOR

Naila Rozi
Sindh Madrassatul Islam University, Karachi. Email: nrozi@smiu.edu.pk

ABSTRACT

With the rapidly growing demand for DNA analysis, the need for storing and processing large scale genome data has presented significant challenges. This paper describes how the genome analysis Toolkit can be deployed to an elastic cloud, and defines policy to drive elastic scaling of the application. We extensively analyses the GATK to expose opportunities for resource elasticity, demonstrate that it can be practically deployed at scale in cloud environment, and demonstrate that applying elastic scaling improves the performance to cost tradeoff achieved in a simulated environment. The applications of biological computing are numerous and growing from patient diagnosis to drug discovery and design large scale computing has never been important in the medical field. Modern biological application can be time sensitive for example, genetic mutation identification may be applied to determine the appropriate therapy for a particular patient, In this Context the users cannot tolerate high or it is a part of the maintenance fees but the only electricity consumption is price shows the total actual execution time to complete all .10TB backup within 25,600s.By developing advanced analytics on the top of big data. Methods like encryption and decryption helps in address many issues related to Molecular DNA big data before making the use of it. Analyzing and finalizing better technical solution for specific application is all time astonishing issue from the developers and analysts perspective..By the time data is growing tremendously, so there is need of more and more advance way of storing the data..One evolving idea to store this data is making the use of z- DNA .Let by George Harvad University a team of researchers have recently figured out a way of whopping 5.5 megabytes=700 terabytes within a single gram of DNA .Rather than enciphering binary data on magnetic drives, scientist are leveraging strand DNA to microcode data.DNA ,capable of storing 96 bits ,is synthesized with each other of the ATGC bases represent binary value T & C representing 0 and A & G representing 1.It is found that DNA is very compact to store one bit base with each base only few atoms large.DNA is also volumetrically meaningful to store in a beaker or other incarnation rather than a Hard disk. Finally while advanced storage system need to be kept in sub zero vacuums. Big data can also help to advance our understanding of DNA.

KEYWORDS

DNA, Encryption, Big Data

INTRODUCTION

The undecidability of DNA uses a technique called diagonalization discovered by George Cantor in1873.Cantor was concerned with the problem of measuring the size of infinite sets of sequencing. If we have two infinite sets, how can we tell whether one is larger than the other or whether they are of same size? For finite sets, of course, answering these questions is easy. We simply count the element of an infinite set; we will never finish bit size of DNA. So we can't use the counting method to determine the relative sizes of infinite set of DNA. In crystal structure can be found invivo and play a role in transcription regulation compared to B-DNA ,Z-DNA is thermodynamically disfavored under normal cellular conditions and it forms only under specifications such as high salt concentration is needed for in vitro studies. However, recent evidence indicates that anions can also influence B –to Z-DNA transition and also a pure elctrostatistical model fails to accurately explain the role of salts in Z- DNA FORMATION. In vivo ,Z-DNA is assumed to form as a result of mechanical constraints like negative super coiling but also of Z-DNA binding proteins such as a double stranded RNA editing enzymes ,which stabilizers DNA in its Z-conformation .The presence of Z -DNA amidst the widely predominant B-DNA structure is dependent on the formation of B-Z DNA junctions which require the presence of more weakly bound AT base pairs. Monitoring of the B-to Z-DNA transitions up until now relied almost exclusively on circular dichorism (CD) measurements and allowed the determination of high resolution Z-DNA and B-Z DNA junction structures.

METHODOLOGY

A considerably variety of EST projects have been spawned sequencing short fragment s typically 200-300 bp of c DNA clones from brain, skeletal muscle, lymphocytes ,liver, heart. Introns and non-coding DNA flanking genes are generally so highly diverged that alignment of orthologous sequences from the two species can be extremely difficult unless the comparison is confined to sequences which are located close to exons. Thus, very short introns less than 200 bp can be aligned and compared but larger introns accounting for the great majority of introns are progressively more difficult to compare because of the very high sequence divergence. TCR cluster ,where sequencing of about 100 kb in mouse and humans reveals a sequence identity of approximately 70% ,even though only about 6% of DNA is coding DNA. This is likely to be related to the very unusual mechanisms for expressing TCR and immunoglobulin genes. DNA is conserved humans indeed the TTAGGG repeats are conserved throughout vertebrates, presumably because of selection pressure to ensure continued recognition by the enzyme. However, highly repetitive DNA sequences in general are among the most rapidly diverging sequences because of selection pressure to ensure continued recognition by the enzymes. However, a highly repetitive DNA sequence there is a poor conservation of hyper variable miniStellite and microsatellite at orthologous locations in humans. The sequence of an introns in the ZFY gene on the Y chromosome revealed no differences when the Y chromosome revealed no differences of 38 different sample taken. Gene tracking involves, at least one who is heterozygous for the diseases and so many or may not have passed on the disease gene to the consultand. The investigation always goes through three stages. The DNA marker sequence used for gene tracking is not the sequence which causes the disease there is always the possibility of recombination between the diseases and the marker. The recombination fraction can be estimated from family studies by standard linkage analysis. RC 8 turned out to show 20% recombination with the DMD locus. The investigation always goes through 3 main stages. Suppose RNA send message to Z- DNA , then by using encryption matrix E is given as:

E=$\begin{bmatrix} ATCGGAT=-1 & AGTCGAAT=-3 \\ CGTATTCG=2 & TAGTACGA=5 \end{bmatrix}$

When we multiply this matrix to $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$(negatively identity) then above matrix will become

E= $\begin{bmatrix} 1 & 3 \\ -2 & -5 \end{bmatrix}$

This process will work if the matrix C is the product of matrices E and M so C= E.M. We then decode the message stored in matrix C by multiplying by matrix D. Since matrices D and E are inverses. Medium of ultra compact information storage: Very large amounts of data that can be stored in compact volume .A gram of DNA contains 1021 DNA bases = 108 Tetrabytes of data. RSA asymmetric cryptographic algorithm suggested by RIVEST in 1977 .The algorithm uses two keys ,public (which is announced to all) and private (which is kept secretly with the receiver).Algorithm of encryption is given as

C= P^{e} modulo N

This generates P = C^{d} modulo N

First two very large prime number p(Genetic Sequence) and q(Large Genetics algorithm GA) are arbitrary chosen

Calculate N= P times multiply q

Φ = (Minus one sequencing of initial ATCG) multiplying effect (Minus one sequencing GA)

Such that 1< e < ϕ Where common difference of each GLA must be equal to 1

Distinguish the 2 chromosomes

In the early days of gene tracking ,this was the limiting step .Not many markers were known and those were mostly two allele RFLPs, for which at most 50 % of people would be heterozygous maximum value of 0.5 at p = q= 0.5this limitation has largely disappeared. That is, work out which marker allele segregates with the disease allele. This is done by typing suitable relatives .Best relative to type depends on the mode of inheritance and the structure .For X-linked diseases the father is always best, if available, because he has only one X chromosome which he necessarily passes on to his daughter. Which introduce a delicate aspect of gene tracking. Where deductions depends critically on paternity, it seems prudent to check paternity However it is important that the genetic team should have a clearly thought out policy.

Discover Pro bound

Use the marker to identify which chromosome is transmitted to the consultand. Recombination between marker and disease can never be ruled out. Even for every linked marker. The same Principle applies to gene tracking in autosomal diseases as to X-linked with the following analysis.

a. Distinguish the two chromosomes in the relevant parents.
b. Determine phase
c. Workout which chromosome?

To Provide security against brute force attacks, a key length should not be less than 2^{100}.In the proposed algorithm there are five keys, four of them belong to the primary values and the control coefficient for the logistic function, while the fifth is a series of DNA. Assuming that accuracy of each key from the first four keys are equal to 10^{-19} as well as the series of DNA, which has a length of 2^n

Huntington Disease

Figure illustrate both gene tracking and direct testing in family with HD .The same logic applies to any dominant diseases, although the late onset of HD causes special problem for counseling. Measuring the size of the $(CAG)_N$ trinucleotide repeat in a DNA sample .This is more reliable than gene tracking and avoids involving the rest of the family. DNA sufficient for PCR typing can usually be extracted from the dried blood spot. Again, using multi-allele microsatellites would make most families informative, provided the pedigree structure was appropriate The family illustrates how genetic testing often involves much more than just the science. Laboratory workers must remember that they are testing people, not DNA. The error rates shown are the chance of having a clinically affected child when the prediction was of an unaffected child or vice versa .They are obtained as follows.

- Four meioses are involved in producting the two offspring. A prediction that the RNAs were affected would be invalidated by a recombination in any one of them, so the error rate is 4ψ.
- If the prediction was that the RNA was a carrier, only half the possible 4 ψ error would be clinically important, so the practical error rate is 2 ψ.
- Two recombination's would be needed to make the fetus affected when it was predicted to be homozygous normal, giving a negligible practical error rate of ψ^2

Risk in Gene Tracking

Unlike direct testing, gene tracking always involves a calculation. Factors to be taken into account in assessing the final risk include: The Probability of diseases-marker and marker –marker recombination, as in the autosomal recessive .Uncertainty, due to imperfect pedigree structure or limited. There are several classes which involve sequence exchange between allelic or non allelic sequence often involving repeated sequences. Variable number tandem repeat (VNTR) can occur in case of repeated units that are very short, intermediate or large. The total population of germ cell in human embryos rises to an estimate maximum of 6.75555555549321 x 10^{9} during the third month. The technical questions in a population screening are fairly simple. Unexpectedly, Perhaps, false positive test result can pose a more serious problem than false negative.

Properties of Chromosome bands with standard Giemsa

The Chromosome 21 is in fact smaller than chromosome 22 in size. The observed metaphase chromosome lengths range between 2µm(chromosome 21) and 10 µm(chromosome 1),whereas the fully uncoiled DNA strands would be expected to measure between 1.7 and 8.5 cm. The presence of extensive heterochromatic regions on the Y chromosome, at the secondary constrictions of chromosomes 1q,9q,16q and on the short arms of the acrocentric chromosomes 13,14,15,21 and 22.

Table 1
Chromosome bands

Dark Bands (Corresponds to G)	Pale Bands (Corresponds to R)
Stain strongly with dyes that bind preferentially to AT –rich regions, such as Giemsa. May be comparatively AT-rich DNAs insensitive. Condense early during the cell cycle but replicates late Gene poor. Line Rich ,but may be poor in ALU repeats.	Stain in weakly with Giemsa May be Comparatively GC-rich DNAs sensitive Condense late

CONCLUSION

With new Encryption algorithm model of Z- DNA based on power of 2^n with respect to AT –rich regions DNA sequences lists the main ways ,the overall message is that one should not be native when speculating about the gene defect underlying a clinical syndrome in which changes to a gene can reduce or destroy its function. In general, it is not HD gene associated with HD expansion of the trinucleotide repeat. No missense, nonsense, frame shift splice mutation have been found. An added complication comes from the fact that most people with recessive conditions whose parents are unrelated are compound hetrozygotes , with two different mutation. For a classification of mutations by their nature and location in the gene it has arisen independently more than once illustrates the relatively high frequency of slipped strand miss pairing. In theory of linkage can be detected between 40c M loci across the gene.

REFERENCES

1. Gehani,T.H La Bean, J.H .Reif, DNA –based cryptography ,in 5th DIMACS series in Discrete Mathematics and Theoretical Computer Science MIT ,Vol.54 .1999,pp.233-249.
2. Smith AJH,De Sousa MA ,Kwabi-Addo B,Happell-Parton A,Impey H ,Rabbits p(1995)Nature Genetics,9,376-385
3. Soni ,A ,Acharya,A, “A Novel image Encryption Approach using an index based chaos and DNA Encoding and its Performance Analysis”Int.J.Comp Applications,vol.47,No.23 June 2012
4. Wang,Q,Zhang,Q,Wei,X,’ image encryption algorithim based on DNA biological Properties and Chaotic systems, “IEEE 5th Int Conf.Bio –inspired Computing: Theories and Application(BIC-TA),2010,PP.132-136