International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 88-92
Published Online June 2019 in IJEAST (http://www.ijeast.com)

A STUDY OF BIOINFORMATICS APPLICATIONS AND MULTICORE ARCHITECTURE

Pradnya Borkar
Computer Science and Engg.
Priyadarshini J.L. College of Engineering

Vijaya Bhute
Computer Science and Engg.
Priyadarshini J.L. College of Engineering

Abstract— Bioinformatics is a discipline which combines two fields computer science and biology. This new discipline has grown rapidly and nowadays it has become the basis for every molecular biological study. This study comprise of computer implementation and algorithms for analyzing and integrating biological data and genetic macromolecules like deoxyribonucleic acid(DNA) , ribonucleic acids(RNA) , or proteins. To process the large data, genetic algorithm is useful for optimization and is being implemented using multicore architecture. In this field, genetic algorithm plays an important role when it comes to the enlargement and improvement such as prediction of RNA secondary structure. Genetic Algorithm is most popular type of evolutionary method and its basic goal is to find out the best optimum result in any given problem. This paper includes briefing of bioinformatics , parallel architecture, genetic algorithm.

***Keywords*— Bioinformatics, multicore architecture, Genetic Algorithm**

I. INTRODUCTION

Bioinformatics is becoming an increasingly popular platform for molecular biological study. Many human genome projects collect nucleic acid sequence data at an extraordinary rate of over one million nucleotides per day which provide large amount of genetic details for analysis. For analyzing biological data and genetic molecules like DNA (deoxy-nucleic acid) and RNA(ribo-nucleic acid), this discipline uses computer implementations and algorithms. The DNA has the information of building other components of cells such as proteins and RNA molecules whereas RNA is a nucleic acid which is meant for copying the genetic information of the DNA.

A. Definition of Terminologies

RNA is part of a group of nucleic acid molecule that provides a mechanism to copy the genetic information of DNA. DNA is made up of two long twisted strands that contains genetic information. The DNA has a direction on how to build other components of the cells, such as proteins and RNA molecules. Most biologically active RNA includes mRNA (messenger RNA), tRNA(transfer RNA), rRNA (ribosomal RNA), snRNA(small nuclear RNA) and other non-coding RNAs. Like DNA, RNA is made up of a long chain of components called nucleotides. The contents of nucleotides are a nucleobase, a ribose sugar, and a phosphate group. In RNA, encoding of genetic information is carried out by sequence of nucleotides. DNA is not used to make proteins directly. RNA carries the information from DNA to a ribosome, where the amino acids are brought together to form a protein. All cellular organisms use messenger RNA (mRNA) to carry the genetic information that gives direction of protein synthesis. RNA is used as genetic information material in many viruses instead of DNA[1]. In prokaryotic cells (cells without nucleus), cytoplasm is the place where RNA and proteins are both made whereas in eukaryotic cells(cells with nucleus), after copying DNA in the nucleus RNA moves to the cytoplasm where the proteins are made.

i. Transcription from DNA to RNA

Transcription is the process of transferring information from DNA to RNA. The chemical structure of RNA and DNA is quite similar as both are made up of four types of nucleotide subunits. Three of the bases of RNA are the same as in DNA: Adenine (A), Guanine (G) and Cytosine (C), and. However, the fourth base of RNA is Uracil (U), not Thymine which is found in DNA. Transcription produces a single-stranded molecule of RNA [2]. Transcription is different from replication in many ways. A single strand of RNA is produced by transcribing only one strand of DNA. After transcription is over, the RNA is released without staying attached to DNA. At the end of transcription, the DNA molecule closes. Replication and transcription involve passing along information that is coded in the language of nucleotide bases. To produce proteins, cells translate this language of nucleotide bases into the language of amino acids. Three specific bases are equal to one amino acid. The actual assembly of the amino acids in their proper sequence is the translation which takes place in the cytoplasm of a cell. The translation involves all three types of RNA i.e. in short period of time , from the same gene, many copies of RNA can be made. The comparison chart of RNA and DNA is shown as below[3].

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 88-92
Published Online June 2019 in IJEAST (http://www.ijeast.com)

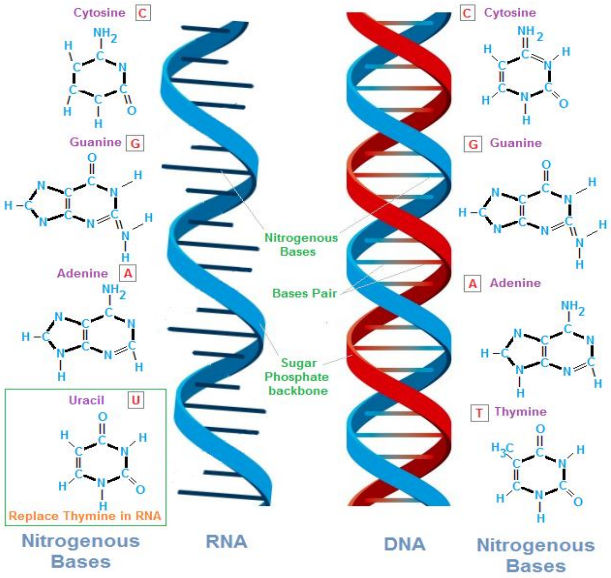

Fig.1: Structure of RNA and DNA

Table I: Difference between DNA and RNA

	DNA	RNA
Stands For	• DeoxyriboNucleicAcid.	• RiboNucleicAcid.
Definition	• Contains genetic information used in the development and functioning of all living organism.	• Carries genetic information from DNA which is used in protein synthesis.
Function	• Provides long term stable storage and transmission of genetic information	• Transfers genetic code needed for the creation of proteins from the nucleus to the ribosome.
Structure	• Double-stranded. • Structure consists of its phosphate group, five-carbon sugar(the stable 2-deoxyribose) and four nitrogen containing nucleobases : Adenine, Thymine , Cytosine and Guanine.	• Single-stranded. • Structure consists of its phosphate group, five-carbon sugar(the stable ribose) and four nitrogen-containing nucleobases: Adenine, Uracil, Guanine and Cytosine.
Base Pairing	• Adenine links to Thymine (A-T) and Cytosine links to Guanine (C-G).	• Adenine links to Uracil (A-U) and Cytosine links to Guanine (C-G).
Location	• Found in the nucleus of a living cell and in mitochondria.	• Depending on the type of RNA, this molecule is found in a cell's nucleus, its cytoplasm, and its ribosome.
Stability	• Deoxyribose sugar in DNA is less reactive because of C-H bonds. • Stable in alkaline conditions. • DNA has smaller grooves, which makes it harder for enzymes to "attack."	• Ribose sugar is more reactive because of C-OH (hydroxyl) bonds. • Not stable in alkaline conditions. • RNA has larger grooves, which makes it easier to be "attacked" by enzymes.
Propagation	• DNA is self-replicating.	• RNA is synthesized from DNA.
Unique Features	• The helix geometry of DNA is of B-Form. • DNA is protected in the nucleus, as it is tightly packed. • DNA can be damaged by exposure to ultra-violet rays.	• The helix geometry of RNA is of A-Form. • RNA strands are continually made, broken down and reused. • RNA is more resistant to damage by Ultra-violet rays.

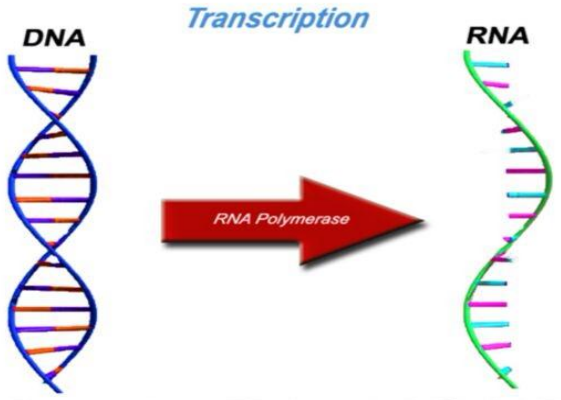

Fig. 2: Transcription of DNA to RNA

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 88-92
Published Online June 2019 in IJEAST (http://www.ijeast.com)

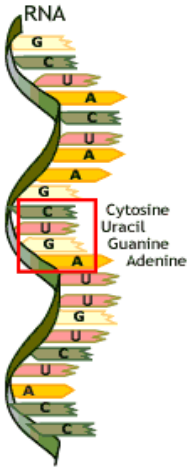

Fig. 3(a): Single stranded RNA Structure

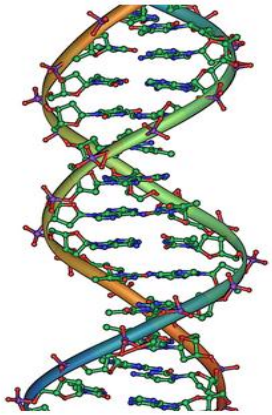

Fig. 3(b): Double Stranded RNA Structure

In recent times, the size of data in various fields such as medical sciences, computational biology and bioinformatics has increased to such proportions that much time is needed to process this agglomerated data. To manage this voluminous data in a short span of time, concurrent processing is needed where the application runs in parallel and the result can be achieved in short time. Nowadays, to obtain greater performance than the state-of-the-art architecture, a new dimension is added to the design by many advancement in hardware capability by implanting multiple processors on a chip. The large scale applications need enormous information processing which lead to the use of parallel architecture. The parallel architecture is also called multi-core architecture which is discussed in next section.

B. Multicore Architecture

A multicore is an architecture design that places multiple processors on a single die (computer chip). As chip capacity increases, placing multiple processors on a single chip becomes practical. These designs are known as Chip Multiprocessors (CMPs). Recently, the CMP has become the preferred method of improving overall system performance. CMPs consist of two or more cores such as two processors (dual core), four processors (quad core), and eight processors (octa-core) configurations and so on Cantu-Paz Erick(1998). Parallel computer architectures have demonstrated structural divergence, generally innervated by specific higher level parallel programming models.

C. Parallel Programming Models

Parallel random access machine (PRAM) executes different instruction by sharing a common clock in each cycle .PRAM permits concurrent access to different memory locations. Depending on memory access, PRAMS are categorized into four types:

i. Exclusive-read, exclusive-write (EREW) PRAM: This type allows exclusive memory accesses i.e. read or write operation cannot be carried out concurrently. This is the weakest PRAM model.
ii. Concurrent-read, exclusive-write (CREW) PRAM: This type allows multiple read accesses whereas multiple write accesses to a memory location are serialized.
iii. Exclusive-read, concurrent-write (ERCW) PRAM: This type allows multiple write accesses whereas multiple read accesses to a memory location are serialized.
iv. Concurrent-read, concurrent-write (CRCW) PRAM: In this type multiple read and write accesses to a common memory location is allowed. This is the most powerful PRAM model.

Apart from above mentioned models, there are two other parallel models shared memory model and message passing model.

D. Shared Memory Model

In Shared Memory Model, a set of independent processors and memory modules are connected through an interconnection network. A memory is globally shared by all processors. The processes running on different processors communicate by writing to and reading from the global memory. Shared memory multiprocessors are usually bus-based or switch-based.

Data coherence and performance degradation due to contention are the main problems address during design of shared memory system. If the multiple processors try to access shared memory simultaneously this may lead to performance degradation. Generally , caches are used to solve the contention problem. Coherence problem is caused due to sharing of multiple copies of data having same value. The copy of data become inconsistent, if one of the processors writes over the value of one of the copies as it is no longer equal to the value of the other copies.

The shared memory systems can be categorized in the following categories:-

i. UMA : In the UMA system, a shared memory is accessible by all processors through an interconnection network. All processors have equal access time to any

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 88-92
Published Online June 2019 in IJEAST (http://www.ijeast.com)

memory location. The UMA system is also known as SMP (symmetric multiprocessor) systems. Each processor has equal opportunity to read/write to memory, including equal access speed.

ii. In the NUMA system, the shared memory access is non-uniform, as the access time varies with the location of memory word. Each processor has its own local memory and the collection of all local memories forms a global memory which is shared by all processors Ra'ed M (2010). An access to local memory is faster as compared to remote memory. Communication between processors is carried out through interconnection network.

Shared memory multiprocessors are usually based on bus structure or switch structure. In both of these cases, each processor has equal access to the global memory which is shared by all processors. Communication among processors is carried out by writing to and reading from memory. Locks and barriers are used for synchronization among processors.

D. Message Passing Architecture

In Message passing systems, communication among multiprocessors is carried out by passing messages as there is no global memory. Each processor has its local memory and all processors are connected with interconnection network.

For evaluating the performance of parallel program, certain measures are required such as scalability, speed up, cost, time complexity etc. Scalability is used as a measure of the system's ability to effectively utilize the processing resources. Scalability is the manifestation of enhancing the system performance in terms of speed, size, efficiency etc. A parallel architecture is said to be scalable, if it can be expanded to a larger system with a linear increase in its performance. Performance of parallel architecture can be determined by calculating the parameters such as speedup, cost and efficiency Banerjee Utpal(1988).

i. Speedup : The speedup factor of a parallel system can be defined as the ratio between the time taken by a single processor to solve a given problem instance to the time taken by a parallel system consisting of n processors to solve the same problem instance. Speedup is represented by

$$S(n) = T_s / T_p$$

where $S(n)$- Speedup , T_S - Time taken by a single processor ,T_P- Time taken by a parallel processor

ii. Cost : The cost of any algorithm is described as solving a problem on a parallel system. Cost reflects as the product of parallel runtime and the number of processing elements used. The cost of algorithm is defined as the total number of steps required for complete execution by n number of processors. Thus the cost can be computed by using formula.

Cost(C_P) = Time Complexity * Total Number of Processors

iii. Efficiency : Efficiency is defined as the ratio of the time required for running sequential algorithm and the cost of the parallel algorithm. The efficiency should be less than or equal to 1. If efficiency is greater than 1 then it indicates that the parallel algorithm is slower than the sequential algorithm.

Efficiency = Running time of Sequential Algorithm / Cost of Parallel Algorithm

$$E = T_s / C_p$$

E. Genetic Algorithm

In Genetic Algorithm, there are several individuals called chromosomes and the set of these chromosomes is called a population. In a population, pairs of chromosomes with good fitness values produce new chromosomes with some operation such as crossover or mutations. After applying these operations , next generation is produced by these new chromosomes according to their fitness values Cantu-Paz Erick(1998). Genetic algorithm uses two operators based on natural genetics to explore the search space i.e. crossover and mutation.Crossover is the primary exploration mechanism in genetic algorithm. This operator takes two random individuals to form the next generation and exchanges the random substrings between them. Mutation is a secondary search operator. The main functionality of mutation operator is to restore diversity because it may be lost due to the repeated application of selection and crossover operation. Mutation is just changing a single point of any string without interchanging. This operator alters some random value within the string. The probability of applying mutation is very low in genetic algorithm, but the probability of crossover is usually high. Many crossover techniques exist for organisms which use different data structures to store themselves.

i. One-point crossover : In one point crossover, a single point is selected on both the parents and all data beyond that point in either parent is swapped which results into the new organism i.e. new children.

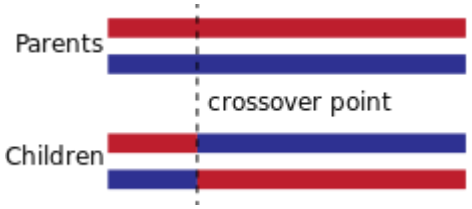

Fig. 4: One Point Crossover

ii. Two-point crossover : In two-point crossover , two points are to be selected on the parent strings. Everything between the two points is exchanged of two parent organisms which provides two child organisms.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 2, ISSN No. 2455-2143, Pages 88-92
Published Online June 2019 in IJEAST (http://www.ijeast.com)

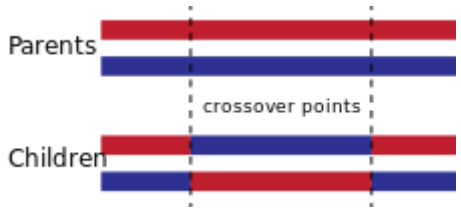

Fig. 5 :Two Point Crossover

iii. Cut and splice : The another crossover variant is the cut and splice approach that results in a change in length of the children strings. Each parent has different crossoverpoint and the data after this point and before this point get exchanged of first and second parent accordingly producing new children.

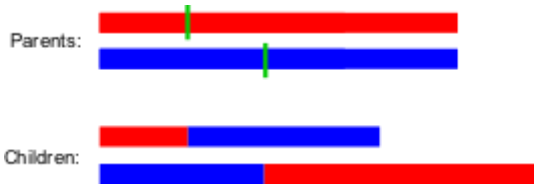

Fig.6: Cut and Splice Crossover

iv. Uniform Crossover and Half Uniform Crossover : In Uniform Crossover a fixed ratio between two parents are used for crossover. The Uniform Crossover facilitate the parent chromosomes to contribute the gene level rather than the segment level. If the mixing ratio is 0.5, then the children take half of the genes from first parent and other half from second parent though the crossover points are chosen randomly as shown below in Fig.7. In half uniform crossover, exactly half of the string organism (bits) between two parents are swapped.

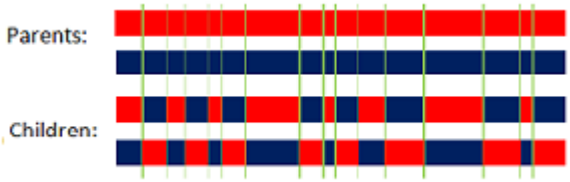

Fig.7 Uniform and Half Uniform Crossover

II. CONCLUSION

In this paper, we have discussed the basics terminologies of bioinformatics such as DNA, RNA, transcription of DNA to RNA. The volumnious data of bioinformatics can be processed using parallel programming and applying optimization using genetic algorithm on multicore architecture.

III. REFERENCES

[1]http://www.apelslice.com/books/9780618843175NIMAS/HTMLOU T/HTML/c_id4625971.html.

[2]https://www.pinterest.com/pin/303711568595038635/.

[3]http://chemistry.about.com/od/lecturenoteslab1/a/Dna-Versus-Rna.htm.

[4] Cantu-Paz Erick(1998): Survey of Parallel Genetic Algorithm Calculateurs paralleles, reseaus et systems repairs, vol.10, no.2,pp.141-171.

[5] Ra'ed M. Al-Khatib, Abdullah Rosni and Rashid Nur'Aini Abdul(2010): A Comparative Taxonomy of Parallel Algorithms for RNA Secondary Structure Prediction Evolutionary Bioinformatics:6 27–45.

[6] Banerjee Utpal(1988): Depedence analysis for Supercomputing, Kluwer Academic Publishers, Norwell, MA,.

[7] Castanotto D, Rossi JJ(2009). The promises and pitfalls of RNA- interference based therapeutics Nature.;457(7228):426–33.

[8] Osborne R. (2007):Companies jostle for lead in RNAi, despite uncertainties.

[9] Thomas D, Frank S.(2009): RNAi and microRNA-oriented therapy in cancer: rationales, pomises and challenges.

[10] Buratti E, Dhir A, Lewandowska MA, Baralle FE.(2007): RNA structure is a key regulatory element in pathological ATM and CFTR pseudoexon inclusion events. Nucl Acids Res. 26;35(13):4369–83.

[11] Reeder J, Steffen P, Giegerich R. pknotsRG(2007): RNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic acids research.;35(Web Server issue):W320.