DNA AND BIOINFORMATICS: A REVIEW OF DIFFERENT ASPECTS AND APPLICATIONS

Shafique Ahmed[1], Bilal Ahmad[2], Idrees Ahmed Nasir[1], Muhammad Arshad[2], Muti-ur-Rehman Khan[3]

[1]Centre of Excellence in Molecular Biology (CEMB), Lahore

[2]Department of Bioinformatics and Biotechnology, International Islamic University, Islamabad

[3]Department of Pathology, University of Veterinary and Animal Sciences, Lahore

ABSTRACT: *Bioinformatics is becoming more and more popular in biology and biotechnology community. Applications of bioinformatics in amplification of DNA by PCR, predicting the correct boundaries of a gene by gene homology or prediction method, and for post sequencing steps in DNA studies like DNA mapping and assembly to construct a complete genome with repetition or missing the sequences, promoter analysis to check the expression level of a gene by analyzing TFBS, evolutionary analysis by phylogeny reconstruction, different genetic markers selection and their use in forensic science are the most popular and useful in DNA studies and research. In this review paper, applications of above mentioned bioinformatics application in DNA studies, online availability and efficiency of computational algorithms in the bioinformatics and reliability of bioinformatics' tool is considered.*

Key words: Gene prediction, DNA mapping, Phylogeny reconstruction, Promoter analysis, Sequence features detection.

I. INTRODUCTION

Bioinformatics is the blended form of two fields; biology and informatics. Major initiative for the creation of bioinformatics was the need to store biological information at beginning of "genomic revolution". Foundation stone of bioinformatics was laid in 1970 when the Needleman-Wunsch algorithm was created for amino acid sequences of two proteins. Hogeweg in 1979 named this new discipline "bioinformatics". Main focus of this discipline has been the development of tools for data management and analysis. By definition 'any application of computation to the field of molecular biology, including data management, algorithm development, and data mining is bioinformatics'. This discipline includes the use of tools and techniques from three distinct disciplines; molecular biology, computer science and the statistics. Traditionally the knowledge base in biology has resided within the heads of experienced biologists. This approach worked well when the amount of data was not so great. However, this situation has changed - many complete genomes are appearing each year and new experiments providing information in molecular biology. For example, one experiment can now produce data on the transcription level of 100,000 different mRNA types. Therefore, there is a need to establish systems that can apply this data to get information. It is not said that such systems could replace human experts; however, they could play a crucial role in filtering the flood of data to the point where human can easily handle this. This then raises many questions, in particular regarding how biological concepts and their relationships can be rendered in ways that make them computationally tractable.

II. GENE PREDICTION

While the sequences of genomes of various organisms have been determined over the last some years, converting such raw sequence data into knowledge remains a hard task. So gene prediction is one of the most important problems in computational biology [1]. Once the genome of a species has been sequenced, gene finding is one of the first and most important steps in understanding the genome function [2]. Also before analyzing the gene, one should have the correct and accurate coding sequence of that gene first. Gene finding in a genome is called gene prediction [3]. Finding genes in prokaryotes is an easier problem than in eukaryotes due to absence of introns and smaller the intergenic regions [4]. The widely used and recognized bioinformatics approaches for genome annotation are homology methods and gene prediction. Approximately 50% of the total genes can be found by homology to the other known genes or protein. Indeed approximately half of the genes can be found by homology to other known genes. In order to determine the 50% of remaining genes, the only solution is to turn to fastand accurate predictive methods [8]. Gene prediction methods require exonic and intronic regions of protein models or from huge quantities of unknown DNA sequence from different species [9]. To solve the gene prediction problem by homology based gene finding we align the genomic sequence, which is to be predicted, with similar sequences present in database and select the best matched sequences. With this information we can infer the possible gene boundaries by comparing with best aligned genes in the database [10]. In Metagenomic samples, many genes are identified by the process called homology to find genes by paying Basic local alignment search tool or BLAST [11]. Number of bioinformatics tools have been developed and employed for gene prediction and annotation of genomic sequences from single prokaryotic species such as GLIMMER and GeneMark [12,13]. CONTRAST uses the term phylogeny-free method to numerous in formatives de novo gene calculations. In this, a two-stage approach is used; designing of a set of two identifier to identify exonic region margins is collected with a worldwide exemplary of gene configuration**.** DIFFERENCE forecasts strict exonic region assemblies for 65% more human genes than the earlier approaches [14]. These gene prediction watch for 5'- upstream UTR, 3'- downstream, promoter region, TATA box and other cis-regulatory factor to determine the exact location of gene. Accuracy of predicting a gene depends upon the length of DNA sequences, larger the DNA sequence more inaccurate prediction will be and vice versa. Because of decline in gene concentration and the occurrence of huge non coding regions, gene prediction software is programmed for specific species such as Gene finder, a gene prediction program optimized for performance in *C. elegans* genomic sequences. It works best for *C. elegans* but not efficient for other species, so selection of software needs your more attention [15]. Some of the leading software's, currently used

for gene prediction; include GeneMark, TWINSCAN, and mGene, GENSCAN, ExoniPhy, ExonHunter and Glimmer.

III. PCR AND PRIMER DESIGNING

PCR is the basic technique to amplify a DNA fragment for various reasons e.g. cloning, checking Short Tandem Repeats (STRs), Single Nucleotide Polymorphism (SNP), parentage identification and genetic mutation. The development of the PCR has often been linked to the development of the Internet. Both inventions have emerged in the last 20 years to the point where it is difficult to imagine life without them [16]. Major components of the PCR are the Taq polymerase enzyme, DNA template, primers, and the PCR machine (Thermo cycler) which maintains the optimum temperature for each step in every cycle [17]. For a specific PCR amplification, one must know the exact nucleotide sequences which lie on either side of the region of interest in DNA. These sequences are used to design two synthetic DNA oligonucleotides (primers) each one complementary to the 5' end of the two strands. The primers are typically 20–30 nucleotides long [18]. The use of specific primers for PCRs intensification of identified or unidentified gene families was first stated about a period of ten years ago and has been generally adopted [19]. There are several classes of primers, including gene specific, SSR specific (Simple sequence repeats), SNP genotyping primers and DNA sequencing primers [20]. Primers are best designed through bioinformatics' tools. In order to design a good primer several parameters need to be considered such as product size, melting temperature (Tm), GC content, primer length, 3' end stability, self-complementarity, dimer possibility and position constraints [20]. Main problems encountered after primer designing are; failure of PCR products or due to mis pairing of primers**.** One should have the genomic data moreover the mark DNA sequence, this genomic information include repetitive DNA elements, protein coding and non-coding boundaries, and SNPs, these are achieved from different databases to get a best optimized primer. Then these all information is collected to form template sequence of primer designing. [21]. UniPrime2 is efficient software, it works by automatically retrieving and aligning homologous sequences from GenBank, identifying regions of conservation within the alignment, and generating suitable primers that can be used to amplify variable genomic regions. Some of the leading bioinformatics tools available for primer designing are Primer Design Pipeline, Primer3, Primer3Plus, RExPrimer, BatchPrimer3 and UniPrime2. Primer Design, Primer3 and Primer3Plus calculate melting temperature(Tm) by nearest neighbor thermodynamic theory and string-based alignment scores to estimate complementarity to pick the best optimum primers [23]. REx Primer uses Primer3 algorithm as core program but it annotated to numerous databases like SNP databases and pseudo gene database. These connected databases allow this program to avoid mis-priming problem [21]. BatchPrimer3 is very efficient for designing sequencing primer. It also uses Primer3 as its core program to compute primer sequence [20,22,24].

IV. DNA MAPPING AND ASSEMBLY

DNA mapping is a significant diagnostic method in different fields of medical, to identify the genomic sequence and pathogenicity of different microbes**.**[25]. DNA mapping is widely used strategy to study structures and organizations of genomes [26]. Genome assembly refers to the process of taking a large number of short DNA sequences, all of which generated by a shotgun sequencing project and putting them back together to create a representation of the original chromosomes from which the DNA originated [27]. Growing interest in comparative genomics has created a need for technologies that can rapidly and efficiently characterize a genome, particularly larger genomes. DNA is assembled by keeping mapping information of the DNA. Genome mapping depends on sequence of the genomes to deliver elementary resources, these materials include DNA probes and localized marker for identification of DNA probes, all these information then built the genomic maps. A big resolution map plays an important role for the construction of complete genome association**.** If someone wants to know the benefits and applications of this method, he should has the knowledge of structure and function of DNA and genes, about codons and anticodons and all the factors that are very important in gene expressions**.**[28]. Gel electrophoresis has been widely used for detection of sequence motifs for the Human Genome Project [29]. The HAPPY mapping technique (Haploid DNA samples that are amplified by the PCR) depends onaccidental DNA splintering and resolving of connection, optical mapping for relatively small genomes, up to 400 Mb so far, Direct Linear Analysis (DLA) based on the analysis of individual DNA molecules bound with sequence-specific fluorescent tags [26,30,31]. Complex genomes contain many repetitive sequences that make it tougher to assemble the reads into the core sequence. Bioinformatics is involved deeply in DNA mapping and assembly. An important step of the assembly process is to generate a set of read-read alignments i.e. aligning sequenced nucleotides for checking the structural matches and mismatches. If we generate only true alignments in this step,then we could produce the optimal assembly of the sequence data. Today, sequencing and assembly methodologies can be applied to entire mammalian genomes by the virtue of bioinformatics [32]. Computer assisted assembly process starts by building a de-novo assembly from the reads and at the same time aligning the same reads to related genomes. Most of the DNA assemblers use anoverlap–layout–consensus computational framework to generate an assembly [33]. More accurately, assembler computational algorithms track this rout: placing reads on a reference genome, grouping reads, enlarging consigns, joining scaffolds, correcting misassembles, and smoothing the assembly [34]. Minimums, a small bioinformatics algorithm, perform well on several small assembly tasks, including the assembly of viral genomes, individual genes, and BAC (Bacterial Artificial Chromosome) clones [35]. Some commonly used bioinformatics tools for genome assembly are Celera assembler, ARACHNE and Eulerian[36-38].

V. PHYLOGENY RECONSTRUCTION

Phylogeny reconstruction involves incorporation of DNA or protein sequences from modern organisms into an

evolutionary model to estimate the corresponding sequence of an ancestor that no longer exists [39]. Phylogenies are extremely useful tools, not only for establishing genealogical relationships among a group of organisms or their parts (e.g. genes), but also for a variety of research once the phylogenies are estimated. Number of uses for phylogenetic information are there among them discovery of drug resistance to reconstructing the common ancestor to all of life is of peak importance [40]. Development of this technique is based upon genomic sequences being known and recent advances in DNA synthesis [41]. Estimating the sequence of an ancestor generally involve two methods, the parsimony and maximum likelihood estimation (MLE) [42]. MLE is used for fitting a statistical model to data, and providing estimates for the model's parameters. While parsimony method is a non-parametric statistical model commonly applied in computational phylogenetic toestimate phylogenies. Nowadays, quartet-based phylogeny reconstruction methods i.e. building a local phylogeny for every subset of 4 species, have acknowledged significant attentions in the bioinformatics community. The accuracy of a phylogeny reconstruction method is measured by simulations on synthetic datasets with known true phylogenies [43]. Currently, all computer algorithms for solving these problems are heuristics without performance guarantee. As a result drawing the phylogenetic tree is not a trivial task since it is not possible to know the exact evolutionary history for a set of organisms [44]. The biological importance of these problems calls for developing better algorithms with assurance of finding either optimal or approximate solutions [45]. Existing phylogeny reconstruction methods tends to calculate the true phylogeny in one of the two ways: by an explicit algorithm that leads to the determination of a phylogeny or by defining a measurement for the quality of generated phylogenies and searching for an optimal phylogeny [43]. Multiple sequence alignment or MSA is a sequence alignment of three or more biological sequences, generally nucleotide (DNA or RNA) or peptide (protein) sequences. In many cases, the input set of query sequences is assumed to have an evolutionary relationship by which they share a common lineage. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Results of MSA can be best viewed and interpreted by using PFAAT or CINEMA software's. They actually add some color schemes like shading, navigation schemes through results by selecting specific column and rows and curationi.e. adding biochemical factor to the mathematically produced result, to the result [46]. But MSA result may contain errors if the input sequences are very different or diverge, otherwise for relatively less diverge sequences. ClustalW, the most common software for multiple sequence alignment, is used for aligning promoter sequences and BaseML from the Paml package used for ancestral sequence reconstruction [47]. Other bioinformatics tools being used extensively are BATWING BEAST, Geneious, MOLPHY (Molecular phylogenetic based upon protein or nucleotide) and PhyloQuart.

VI. BIOINFORMATICS APPLICATIONS IN PROMOTER ANALYSIS

Promoter is present upstream of the transcription start site (TSS) that helps in of transcription initiation and it is a cis-acting element. [48]. Gene expression depends on the interaction between transcription factors and cis-acting sequence elements in transcriptional regulation (promoter) regions [49]. Core and Proximal promoter are the main groups of promoter. Core promoter is found adjacent to TSS, while proximal promoter finds its place one or more kb upstream of the gene. Accurate discovery of these fundamentals is a precondition to decoding the multi-faceted controlling systems that helps in different protein gene expression patterns like tissue specific and lineage specific[8]. Transcription factor binding sites (TFBS) range from 6 to 12 base pairs, these are progressive sequences, due to this reason, the promoter cannot detect TFBSincomplex DNA sequence [50]. Gene sequence, gene expression, and binding data are used to develop the predicting tools for transcription regulating cites [51]. Most of the algorithms used in bioinformatics for promoter analysis use cross-species comparison to screen TFBS calculations and to recognize possibly active controlling essentials, these controlling essentials are ConSite and rVista 2.0 [52,53]. This phylogenetic foot printing method includes CONREAL and Footer is webs based and are used in gene predictions studies**.** [54,55]. Osteopromoter Database (OPD) is an innovative promoter database that has many different of genes that arbitrate osteoblast cells in propagation and distinction of the cells. InOPD, the genes and promoters are presented; these are retrieved from the PUBMED, different bibliographic reference databases and complementary DNA sequences hubs**.** This data is prepared in alphabetical order and have cross-references with other new or old databases [48].For promoter analysis,PromAn is used that is web based software. PromAn delivers computerized exploration of genomic areas with slight previous information of the genomic sequence**.** Prophecy and experimental catalogs are joined together to localize the promoter site within a large genomic input sequence [50]. Other promoter site databases are also available such as DBTSS (Database of Transcriptional Start Sites) and EPD (Eukaryotic Promoter Database) [56,57]. In future the software's and databases designed for promoter prediction will also incorporate tissue-specific, transcriptomic and interatomic data comparative to transcription factors to put in new filtration methods to improve the TFBS. These filtration methods are used to help the users to access regulatory motifs and modules to express the transcription factors interacting**.** [50]. Presence of multiple TSS's often induces limitation into the result of promoter prediction software's. But results of software can be benchmarked by comparing them against the promoter present in EPD [58]. Other efficient software's being used to locate the promoter region are Dragon Promoter Finder, Eponine Transcription Start Site finder, NNPP, Promoter Prediction are Regulatory Sequence Analysis Tools.

VII. SEQUENCE FEATURES DETECTION

Finding location of features in nucleotide sequences is one of the most common tasks in sequence data analysis [59]. The

identification of features in large and complex datasets is an important step towards gaining insight in the processes underlying the data and extracting knowledge from complex biological data [60]. A feature is a sequence pattern with some functional significance, such as start and stop codon, splice sites, and sequences that are recognized by protein in order to regulate gene expression [61].Several DNA features have been discovered and the list is still being populated as the research is going on. Some of them are listed in table 1. The borders between introns and exons are termed as splice sites [63]. A range of computational methods have been developed for detecting splice site and other features. These computational methods can be grouped into different categories, which include probabilistic approaches, the neural network, the support vector machine approaches, and the methods based on discriminate analysis and the information theoretic approaches [64-67]. There are more than 150 Bioinformatics software are present to identified the regulatory bindings sites.[68]. As input, these algorithms e.g. Features canneedtwo sequences, a patternofsequence and a sequence target. Explore setting are put by selecting a definite DNA factors and anentrancerate**.** These outcomes are showed in FASTA format. ENSEMBL, National Centre for Biotechnology Information (NCBI) and University of California, Santa Cruz (UCSC) are used as the external databases [62]. Many different strategies have been developed and still developing with their own limitation and advantages. To customize the multiple methods into one method, BEST and EMD two methods are adopted to benefit the bioinformatics communities. [69,70]. Techniques that incorporate further sources of information have also made recently. Such as PhyloGibbs, PhyME, and WeederHlinkjointly sequences from controlling regions of linked organisms [71-73]. Other software programs being used in bioinformatics to identify sequence feature detection are MotifViz, ORF Finder, POBO, PredictRegulon, RepeatMasker, rVista, TRANSFAC (database on eukaryotic transcriptional regulation), Web Weeder and SeqVISTA. MotifViz uses three motif discovery programs, Clover, Rover and Motifish, covering most available algorithms for detecting motifs OR tool identifies all open reading frames using the standard or alternative genetic codes. POBO uses bootstrap analysis to detect significantly over- or underrepresented promoter regions. PredictRegulon, a web server, constructs the binding site recognition profile based on ungapped MSA of known binding sites to detect the operons and protein binding cites in prokaryotic genomes

Repeat Masker screens DNA sequences for interspersed repeats and low complexity DNA sequences by sequence comparison through cross match, RMBlast and Decypher. The rVistaand Web Weedertool combines sequence comparisons, TFBS predictions and cluster analysis to identify junk and noncoding DNA regions that are conserved evolutionarily.SeqVISTApresent a graphical view of different features detected by different tools. It display results from different sequence analysis tools in an integrated style and aims to provide unity to the bioinformatics resources present on internet.

VIII. SEQUENCE POLYMORPHISMS

The genomes of individuals from the same species vary in nucleotide sequence as a result of different evolutionary processes known as sequence polymorphism. Examining the patterns of sequence polymorphism is called sequence polymorphism analysis [74]**.** There are different methods that are used in Polymorphism sequence like DNA sequencing, restriction fragment length polymorphism (RFLP), single strand conformation polymorphism (SSCP), randomly amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) [75-78].Single nucleotide polymorphisms (SNPs) are widely used in natural populations studies now a day's[79]. Parallel sequencing (sequencing multiple target polynucleotide motifs in a sample) technologies have turned into significant and commonly used methods in the study of polymorphisms sequence on

genome-wide level [80].Researchers generate large amount of sequence data with the help of high throughput technology, with the help of this technology, the whole genome sequencing is used to find the polymorphism sequence as well as phenotypic consequences. With the advancement of modified software to predict data created by these technologies has lagged behind [81]. One commonly used bioinformatics tool for the detection of sequence polymorphism is Smith-Waterman algorithm. Smith-Waterman algorithm gives the best results but it takes much time due to the number of computations required for the search [82]. DNA sequencing platform"Illumina"can now produce about 100 million sequence reads of up to 75-nt every in a one run [81]."*Galign",* a web based algorithm is used to find polymorphisms between sequence retrieved by Illumina technology [83]. We do not use Smith-Waterman algorithm in galign alignment method for sequence analysis. Instead of Smith-Waterman algorithm, we use simple algorithm to read the parse sequence.

The *galign* method gives us polymorphism location, nucleotide alterations and amino acid changes.[81]. Multiple methods have been developed for SNP prediction and filtering, such as GMAP and Maq mapping software [84-86]. A database has also been developed and publically available for sequence polymorphism known as Polymorphic [87].

Table (1): Nucleic Acid Features

Feature name	Briefing
1. LTR	Long Terminal Repeat is a sequence that is directly repeated at both ends of a definite sequence.
2. Enhancer	A cis-acting sequence rises the process of eukaryotic promoters
3. GC signal	GC box is a preserved GC-rich region sited upstream of the start end of eukaryotic transcription sites.
4. -35 signal	Itis a preservedhexameraround 35 bps upstream of the start end of bacterial transcription point.
5. CAAT signal	CAAT box; piece of a conserved DNA sequence located about 75 bp up-stream of the starting point of eukaryotic transcription units which may be involved in RNA polymerase binding
6. N_region	Extra nucleotides inserted between rearranged antibody coding DNA segments
7. N_region	Codes for the variable amino terminal part of an antibody
8. sig_peptide	Signal peptide coding sequence
9. primer_bind	Non-covalent primer binding site for initiation of transcription, or reverse transcription
10. promoter	Region on a DNA molecule where RNA polymerase bind
11. rep_origin	Starting site for duplication of nucleic acid to give two duplicates
12. TATA signal	TATA box is a conserved AT-rich septamer found about 25 bps before the start point of each eukaryotic gene feature
13. repeat region	Region of genome containing repeating segments
14. STR	Short tandem repeat

IX. SEQUENCE RETRIEVAL AND SUBMISSION

It is becoming gradually popular for research individuals to interchange new biological data and update new sequences by directly uploading the data on the Web based databases. GenBank is a comprehensive database that contains publicly available nucleotide sequences for more than 1000 organisms (http://www.ncbi.nlm.nih.gov/guide/all/), obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including WGS and environmental sampling projects [88].Daily data exchange with the European Molecular Biology Laboratory (EMBL), Nucleotide Sequence Database in Europe and the DNA Data Bank of Japan ensures worldwide submission and retrieval [89,90].A repository of primary nucleotide sequences is an essential requirement for computational analysis and genome research. "Webin" has been designed to allow rapid submission of single, multiple or very large numbers of sequences (bulk submissions) at EMBL and is available athttp://www.ebi.ac.uk/embl/Submission/webin.html.GenBank is a quite useful repository service provided by NCBI is GenBank. All findings records are entered in GenBank by direct electronic submissions (www.ncbi.nlm.nih.gov/Genbank/index.html), with the majority of authors using BankIt or Sequin programs. Bankit is used when the data, which is to be submitted, is a single sequence, a simple set of sequences or a small batch of different sequences but Sequin is applied when complex submissions are to be made containing long sequences, multiple annotations, or phylogenetic and population studies [88]. The sequences and biological annotations in GenBank, and the collaborating databases EMBL and DDBJ, are submitted primarily by individual authors to one of the three databases, in the categories of EST (Expressed sequence tags), STS (Sequence-tagged sites), GSS (genome survey sequences), HTC (High-throughput genomic) and WGS (Whole genome shotgun sequence).GenBank personnel gives an accession number to a sequence submission. That is shared between the three large collaborating databases GenBank, DDBJ (DNA Data Bank of Japan) and EMBL (European Molecular Biology Laboratory) and remains same over the lifetime of the record. The accession number serves as a citation reference. [91].the sequence records in GenBank are accessible through Entrez (www.ncbi.nlm.nih.gov/Entrez/),an excellent database retrieval system that covers over 30 biological databases [92]. The user can retrieve the nucleotide sequence through different references such as getting sequence records with respect to the sequencing projects and BLAST search [93]. The EMBL SRS server maintains a comprehensive collection of specialized databanks along with the main nucleotide and protein databases [94]. User can search the sequence by similarity method or individual taxonomic division [95]. The most commonly used computational algorithms for sequence retrieval are FASTA and WU-BLAST [96,97]. Other than these, bioinformatics tools being used commonly in sequence retrieval and submission are Alias Server, EMBOSS, Gene Lynx, SeqHound, Sequin, Gene Lynx and PubCrawler.

X. FORENSIC BIOINFORMATICS

The science of DNA-based human identification is known as forensics [98].Human recognition depends on specific properties like DNA evidence, blood sample, saliva sample, etc.DNA is the most reliable source of human identification due to its genetic differences that are expressed or unexpressed DNA sequences can be used as markers to differentiate between individuals [99]. Bioinformatics and DNA based forensic methods are inter disciplinary and

illustrate their methods from figures and computer sciences, while, computational forensics (CF) integrates expertise from computational science and forensic sciences which is based upon computer-based modeling, simulation, analysis, and recognition in studying and solving problems posed in various forensic disciplines [100]. Now days, STR markers are used for the identification of criminals using 13-17 nuclear STR markers. These markers are made by the Combined DNA Index System (CODIS)[101]. Orchid Biosciences has pioneered relatively new forensic method. In this method personal identification is based upon Single Nucleotide Polymorphisms [SNPs][102]. A bioinformatician can determine haplotypes by using different bioinformatics algorithm, such as Haploview, on sequence data [103]. Another facility provided by bioinformatics to forensic sciences is the establishment of STR databases which facilitate the estimation of the probability of matching or mismatching of two DNA profiles [99]. One of the most popular STR database is ENFSI DNA WG STR Population Database (http://www.str-base.org/index.php). Bioinformatics also help in mass disaster identification in which we have to identify parent-child relationship and other kinships like that. There are some potential drawbacks of using computational programs as error in comparing sequences for culprit identification may wrongly accuse an innocent. Commonly used software's on this front are Mass Disaster Kinship Analysis Program (MDKAP) and Mass Fatality Identification System (M-FYSis) [104]. Bayesian network, application of bioinformatics, is being used for statistical inference on the DNA data produced during process of analysis [105].

XI. CONCLUSION: IS BIOINFORMATIC ANALYSIS RELIABLE?

Bioinformatics software's are built on the bases of statistical and computational principles. Although they are given with good biological learning but sometimes results of a biological phenomenon or reaction differ a lot from set statistical rules. That is why results produced from computational tools should be taken with caution. For example in phylogeny reconstruction, ClustalW reconstructed alignments are highly uncertain in their details. Only very closely related sequences can produce accurate alignments by bootstrap method, while many sequence sets of biological interest are expected to produce reconstructed alignments with error in more than half of their aligned columns. This method does not gives accurate phylogenetic tree but it only gives information about the stability of the tree topology (the branching order), and it helps assess whether the sequence data is adequate to validate the topology [106]. But due to the upcoming programs and benchmarks e.g. PREFAB results are flattering gradually more consistent, but they do predict only.

REFERENCES:

1. Flicek, P., "Gene prediction: compare and CONTRAST," *Genome Biol,* **8:** 233(2007)
2. Korf, I.,"Gene finding in novel genomes," *BMC Bioinformatics*, **5**: 59 (2004)
3. Alexandre, L. Vardges, T.H. Yury, O. Chernoff and Mark, B., "Gene identification in novel eukaryotic genomes by self-training algorithm," *Nucleic Acids Res*, **33**: 6494–6506. (2005)
4. Catherine, M. Marie-France, S. Thomas, S. and Pierre, R., "Current methods of gene prediction, their strengths and weaknesses," *Nucleic Acids Res*, **30**: 4103-4117 (2002)
5. Cho, Y. and Walbot, V., "Computational methods for gene annotation: the Arabidopsis genome," *Curr. Opin Biotec*, **12**: 126–130 (2001)
6. Borodovsky, M. Rudd, K. E. and Koonin, E. V., "Intrinsic and extrinsic approaches for detecting genes in a bacterial genome," *Nucleic Acids Res*, **22**: 4756–4767 (1994)
7. Fickett, J.W., "The gene identification problem: an overview for developer," *Comput. Chem*, **20**: 103–118 (1996)
8. Fickett, J.W.," Finding genes by computer: the state of the art," *Trends Genet*, **12**: 316–320 (1996)
9. Besemer, J. and Borodovsky, M., "Heuristic approach to deriving models for gene finding," *Nucleic Acids Res*, **27**: 3911–3920 (1999)
10. Gross, S.S. and Brent, M.R., "Using multiple alignments to improve gene prediction," *J. Comput. Biol*, **13**: 379-393 (2006)
11. Altschul, S.F. Gish, W. Miller, W. Myers, E.W. and Lipman, D.J., "Basic local alignment search tool," *J. Mol. Biol*, **215**:403-410 (1990)
12. Delcher, A.L. Harmon, D., Kasif, S., White, O. and Salzberg, S.L., "Improved microbial gene identification with GLIMMER," *Nucleic Acids Res*, **27**: 4636-4641 (1999)
13. Lukashin, A. and Borodovsky, M., "GeneMark.hmm: new solutions for gene finding," *Nucleic Acids Res*, **26**: 1107-1115 (1998)
14. Gross, S.S. Do, C.B. Sirota, M. and Batzoglou, S.,. "CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction," *Genome Biol*, **8**: 269 (2007)
15. Kevin, L.H. Tom, C. and Richard, D., "GAZE: A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming," *Genome Res*, **12**: 1418-1427 (2002)
16. Bartlett, J.M.S. and Stirling, D., "A Short History of the Polymerase Chain Reaction," *Methods Mol. Biol*, **226**: 3-6 (2003)
17. Sambrook, J. and David, W.R., "Molecular Cloning: A Laboratory Manual Chapter 8: In vitro Amplification of DNA by the Polymerase Chain Reaction," CSHL Press, (2001)
18. Hamed, S.N. Noorossadat, T. and Mahmood, C., "Designing multiple degenerate primers via consecutive pairwise alignments," *BMC Bioinformatics*, **9**: 55(2008)
19. Zhensheng, P. Richard, B. Alexey, L. and Mikhail, S., "Selection strategy and the design of hybrid oligonucleotide primers for RACE-PCR: cloning a family of toxin-like sequences from Agelenaorientalis," *BMC Mol. Biol*, **8**: 32 (2007)
20. Frank, M.Y. Naxin, H. Yong, Q.G. Ming-cheng, L. Yaqin, M. Dave, H. Gerard, R.L. Jan, D. and Anderson, O.D., "BatchPrimer3: A high throughput web

application for PCR and sequencing primer design," *BMC Bioinformatics*, **9**: 253(2008)

21. Ittima, P. Chumpol, N. Anunchai, A. Pongsakorn, W. Payiarat, S. Uttapong, R. Gallissara, A. and Sissades, T., "RExPrimer: an integrated primer designing tool increases PCR effectiveness by avoiding 3' SNP-in-primer and mis-priming from structural variation," *BMC Genomics*, **10**: 4(2009)
22. Kelvin, L. Anushka, B. Timothy, B.S. Karen, B. Tina, C.M. Dana, B. Steve, F. Sean, M. and Samuel, L., "Novel computational methods for increasing PCR primer design effectiveness in directed sequencing," *BMC Bioinformatics*, **9**: 191(2008)
23. Andreas, U. Harm, N. Xiangyu, R. Bisseling, T. René, G. and Jack, A. M. L. ,"Primer3Plus, an enhanced web interface to Primer3," *Nucleic Acids Res*, **35**: 71–74(2007)
24. Robin, B. Nicola, S. Michaël, B. and Emma, C. T., "UniPrime2: a web service providing easier Universal Primer design," *Nucleic Acids Res*, **37**: 209–213 (2009)
25. Ming, X. Angie, P. Connie, H. Ting-Fung, C. Dongmei, C. Lucinda, L. Eunice, W. Amy, L.K. Joseph, L.D. Paul, R.S. and Pui-Yan, K., "Rapid DNA mapping by fluorescent single molecule detection," *Nucleic Acids Re*, **35**: e16 (2007)
26. Eugene, Y. C. Nuno, M. G. Rebecca, A. H. Amie, J. H. Jonathan, W. L. Anthony, M. M. Gregory, R.Y. Eugene, D. C. Martin, F. Gordon, G. W. Steven, R. G. and Rudolf, G., "DNA Mapping Using Microfluidic Stretching and Single-Molecule Detection of Fluorescent Site-Specific Tags," *Genome Res*, **14**: 1137–1146 (2004)
27. Anderson, S., "Shotgun DNA sequencing using cloned DNase I-generated fragments," *Nucleic Acids Res*, **9**: 3015–3027 (1981)
28. Phillips J.A., "DNA mapping in growth and developmental disorders" *Horm. Res*, **41**: 157-168(1994)
29. Soderlund, C. Humphray, S. Dunham, A. and French, L., "Contigs built with fingerprints, markers, and FPC V4.7," *Genome Res*, **10**: 1772–1787(2000)
30. Dear, P. H. and Cook, P. R., "Happy mapping: a proposal for linkage mapping the human genome," *Nucleic Acids Res*, **17:** 6795–6807 (1989)
31. Zhou, S. Bechner, M. C. Place, M. Chris, P. C. Louise, P. Sally, A. L. Rod, R. Steve, G. Miron, L. and David, C. S., "Validation of rice genome sequence by optical mapping," *BMC Genomics*, **8**: 278-293(2007)
32. Andreas, S. Mostafa, R. Haixu, T. Pavel, P. and Serafim, B., "Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies," *PLoS ONE*, **2**: 484 (2007)
33. Anton, V. David, C.S. Shiguo, Z. and Michael, S. W., "An algorithm for assembly of ordered restriction maps from single DNA molecules," *Proc. Natl. Acad. Sci*, **103**: 15770–15775 (2006)
34. Sante, G. Eric, S.L. Kerstin, L.T. and David, B.J.," Assisted assembly: how to improve a de novo genome assembly by using related species," *Genome Biol*, **10**: 88 (2009)
35. Daniel, D.S. Arthur, L.D. Steven, L.S. and Mihai, P., "Minimus: a fast, lightweight genome assembler," *BMC Bioinformatics*, **8**: 64 (2007)
36. Myers, E.W. Sutton, G.G. Delcher, A.L. Dew, I.M. Fasulo, D.P. Flanigan, M.J. Kravitz, S.A. Mobarry, C.M. Reinert, K.H.J. and Remington, K.A., "Whole-genome shotgun assembly and comparison of human genome assemblies," *Sci*, **287**: 2196–2204(2000)
37. David, B.J. Jonathan, B. Sante, G. Evan, M. Kerstin, L.T. Jill, P.M. Michael, C.Z. and Eric, SL., "Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2," *Genome Res*, **13**: 91–96 (2003)
38. Pevzner, P.A. Tang, H. and Waterman, M.S., "An Eulerian path approach to DNA fragment assembly," *Proc. Natl. Acad Sci*, **98**: 9748-9753 (2001)
39. Guoliang, L. Jian, M. and Louxin, Z., "Greedy Selection of Species for Ancestral State Reconstruction on Phylogenies: Elimination Is Better than Insertion," *PLoS One*, **5**: 8985(2010)
40. Clement, M. Posada, D. and Crandall, K.A.," TCS: a computer program to estimate gene Genealogies," *Mol. Ecology*, **9**: 1657–1659 (2000)
41. Thornton, J.W., "Resurrecting ancient genes: experimental analysis of extinct molecules," *Nat. Rev. Genet*, **5**: 366–375 (2004)
42. Felsenstein, J.J., "Using the quantitative genetic threshold model for inferences between and within species," *Philos. Trans. R. Soc. Lond. B. Biol. Sci*, **360**: 1427–1434 (2005)
43. Gang, W. Ming-Yang, K. Guohui, L. and Jia-Huai, Y., "Reconstructing phylogenies from noisy quartets in polynomial time with a high success probability" *Algorithms Mol. Biol* **3**: 1(2008)
44. Hyun, J.P. and Tiffani, L.W., "A Fitness Distance Correlation Measure for Evolutionary Trees," *Bioinformatics and Computational Biology Lecture Notes in Computer Science*, **5462**: 331-342(2009)
45. Elias, I. and Tuller, T., "Reconstruction of ancestral genomic sequences using likelihood," *J Comput Biol*, **14**: 216-237 (2007)
46. Procter, J.B. Thompson, J., Letunic, I.,Creevey, C.,Jossinet, F. and Barton, G.J., "Visualization of multiple alignments, phylogenies and gene family evolution," *Nat. Methods*, **7**:S16-25 (2010)
47. Edgar, R.C. and Batzoglou, S., "Multiple sequence alignment," *Curr. Opin. Struct. Biol*. **16**: 368-373 (2006)
48. Grienberg, I. and Benayahu, D., "Osteo-Promoter Database (OPD) – Promoter analysis in skeletal cells," *BMC Genomics*,**6**: 46 (2005)
49. Leonardo, M.R. Kannan, T. John, L.S. and David, L., "Promoter Analysis: Gene Regulatory Motif Identification with A-GLAM," *Methods Mol. Biol*, **537:** 263–276 (2009)
50. Aurélie, L. Frédéric, C. Laurent, B., José-Alain, S., Thierry and Olivier, P., " PromAn: an integrated knowledge-based web server dedicated to promoter analysis," *Nucleic Acids Res*, **34**: 578–583 (2006)
51. Seon-Young, K. and Yong, S.K., "Genome-wide prediction of transcriptional regulatory elements of

human promoters using gene expression and promoter analysis data,". *BMC Bioinformatics*, **7**: 330.
52. Loots, G.G. and Ovcharenko, I., "rVISTA 2.0: evolutionary analysis of transcription factor binding sites," *Nucleic Acids Res*, **32**: 217–221(2004)
53. Sandelin, A. Wasserman, W.W. and Lenhard, B., "ConSite: web-based prediction of regulatory elements using cross-species comparison," *Nucleic Acids Res* ,**32**: 249–252 (2004)
54. Corcoran, D.L. Feingold, E. and Benos, P.V., "FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic foot printing," *Nucleic Acids Res*, **33**: 442–446 (2005)
55. Berezikov, E. Guryev, V. and Cuppen, E., "CONREAL web server: identification and visualization of conserved transcription factor binding sites," *Nucleic Acids Res*, **33**: 447–450 (2005)
56. Schmid, C.D. Perier, R. Praz, V. and Bucher, P., "EPD in its twentieth year: towards complete promoter coverage of selected model organisms," *Nucleic Acids Res*, **34**: 82–85 (2006)
57. Yamashita, R. Suzuki, Y. Wakaguri, H. Tsuritani, K. Nakai, K. and Sugano, S., "DBTSS: database of human transcription start sites, progress report," *Nucleic Acids Res*, **34**: 86–89 (2006)
58. Rajeev, G. and Pankaj, S., "Human pol II promoter prediction: time series descriptors and machine learning," *Nucleic Acids Res*, **33:** 1739 (2005)
59. Kerstin, Q. Kornelie, F. Holger, K. Edgar, W. and Thomas, W., "Matind and Matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data," *Nucleic Acids Res*, **23**: 4878-4884 (1995)
60. Yvan, S. Sven, D. Dirk, A. Pierre, R. and Yves, V. P., "Feature selection for splice site prediction: A new method using EDA-based feature ranking," *BMC Bioinformatics*, **5**: 64 (2004)
61. Cynthia, G. and Jambeck, P., "Developing Bioinformatics Computer Skills," *O'Reilly Media Inc*,169 (2001)
62. Igor, V.D. Björn, B., Daniel, W. Yulia, M.K. Alexander, E.K. Helmut, B. and Gerhard, K., "FeatureScan: revealing property-dependent similarity of nucleotide sequences," *Nucleic Acids Res*, **34**: 591-595(2006)
63. Akma, B. Chang, B.C.H. Halgamuge, S.K. and Jason, L., "Splice site identification using probabilistic parameters and SVM classification," *BMC Bioinformatics*, **7**: 15(2006)
64. Pertea, M. Lin, X. and Salzberg, S.L., "GeneSplicer: a new computational method for splice site prediction," *Nucleic Acids Res*, **29**: 1185–1190 (2001)
65. Degroeve, S. Saeys, Y. Baets, B.D. Rouze, P. and Peer, Y.V.D., "SpliceMachine: predicting splice sites from high-dimensionallocal context representations," *Bioinformatics*, **21**:1332–1338 (2005)
66. Zhang, M., "Identification of protein coding regions in human genome by quadratic discriminant analysis," *Proc. of International conference on Genome Informatics*, **13**:192–200 (1997)
67. Arita, M. Tsuda, K. and Asai, K., "Modeling splicing sites with pairwise correlations," *Bioinformatics*, **18**: 27–34 (2002)
68. Daniel, Q. Kathryn, D. Mohammad, S. Dhundy, B. and Hesham, A., "MTAP: The Motif Tool Assessment Platform," *BMC Bioinformatics*, **9**: 6 (2008)
69. Che, D. Jensen, S. Cai, L. and Liu, J., "BEST: Binding-site Estimation Suite of Tools," *Bioinformatics*, **21**: 2909–2911 (2005)
70. Hu, J. Yang, Y. and Kihara, D., "EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences," *BMC Bioinformatics.*,**7**: 342 (2006)
71. Siddharthan, R. Siggia, E.D. and van Nimwegen, E., "PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny," *PLoS Comput. Biol*, **1**: 67 (2005)
72. Sinha, S. Blanchette, M. andTompa, M., "PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences," *BMC Bioinformatics*, **5**:170 (2004)
73. Pavesi, G. Zambelli, F. and Pesole, G., "WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences," *BMC Bioinformatics*, **8**: 46 (2007)
74. Richard, M.C. Gabriele, S. Christopher, T. Stephan, O. Georg, Z. Paul, S. Norman, W. Tina, T.H. Glenn, F. David, A.H. Huaming, C. Kelly, A.F. Daniel, H.H. Bernhard, S. Magnus, N. Gunnar, R. Joseph, R.E. and Detlef, W., "Common Sequence Polymorphisms Shaping Genetic Diversity in Arabidopsis thaliana," *Sci*, **317**: 338-342 (2007)
75. Bagyalakshmi, R. Senthilvelan, B. Therese, K.L. Murugusundram, S. and Madhavan, H.N.," Application of polymerase chain reaction (pcr) and pcr based restriction fragment length polymorphism for detection and identification of dermatophytes from dermatological specimens," *Indian J. Dermatol*, **53**: 15-20 (2008)
76. Igor, V., "Medical Biomethods Handbook," *As minhaspublicações*, 73-77 (2005)
77. Gu, W**.** Post, C.M. Aguirre, G.D. and Ray, K., "Individual DNA bands obtained by RAPD analysis of canine genomic DNA often contain multiple DNA sequences," *The J. Hered*, **90**: 96-98 (1999)
78. Schlotterer, C., "The evolution of molecular markers - just a matter of fashion?," *Nat. Rev. Genet*, **5**:63-69 (2004)
79. Brumfield, R.T. Beerli, P. Nickerson, D.A. and Edwards, S., "The utility of single nucleotide polymorphisms in inferences of population history," *Trends. Ecol. Evol*, **18**: 249-256 (2003)
80. Rogers, Y.H. and Venter, J.C., "Genomics: Massively parallel sequencing," *Nat*, **437**: 326-327 (2005)
81. Shai, S., "galign: A Tool for Rapid Genome Polymorphism Discovery," *PLoS One*, **4:** 7188 (2009)
82. Farrar, M., "Striped Smith-Waterman speeds database searches six times over other SIMD implementations," *Bioinformatics*, **23**: 156-61 (2007)
83. Whiteford, N. Skelly, T. Curtis, C. Ritchie, M.E. Löhr, A. Zaranek, A.W. Abnizova, I. and Brown, C., "Swift:

primary data analysis for the IlluminaSolexa sequencing platform," *Bioinformatics*, **25**: 2194-2199 (2009)

84. Wu, T.D. and Watanabe, CK., "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics*, **21**: 1859–1875 (2005)
85. Li, H. Ruan, J. and Durbin, R., "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Res,"***18**: 1851–1858 (2008)
86. David, L.H. Steven, B.C. Qijian, S. Nathan, W. Edward, W.F. Randy, C.S. James, E.S. Andrew, D.F. Gregory, D.M. and Perry, B.C., "High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence," *BMC Genomics*, **11**: 38 (2010)
87. Eric, B. Laurent, D. Simon, P. and Nicolas, G., "Polymorphix: a sequence polymorphism database," *Nucleic Acids Res*, **33**: 481-484 (2005)
88. Dennis, A.B. Ilene, K.M. David, J.L. James, O. and Eric, W.S., "GenBank," *Nucleic Acids Res*, **38**: 46–51 (2010)
89. Sugawara, H. Ogasawara, O. Okubo, K., Gojobori, T. and Tateno, Y., "DDBJ with new system and face," *Nucleic Acids Res*, **36**: 22-24 (2008)
90. Tamara, K. Philippe, A. Nicola, A. Wendy, B. Kirsty, B. Paul, B. Alexandra, V.D.B. Guy, C. Karyn, D. Ruth, E. Nadeem, F. Maria, G.P. Nicola, H. Carola, K. Rasko, L. Quan, L. Vincent, L. Rodrigo, L. Renato, M. Michelle, M. Francesco, N. Ville, S. Peter, S. Guenter, S. Mary, A.T. Katerina, T. Robert, V. Dan, W. Weimin, Z. and Rolf, A., "The EMBL Nucleotide Sequence Database," *Nucleic Acids Res*, **32**: 27–30 (2004)
91. Dennis, A.B. Ilene, K.M. David, J.L. James, O. and David, L.W., "GenBank," *Nucleic Acids Res*, **36**: 25–30 (2008)
92. Dennis, A.B. Ilene, K.M. David, J.L. James, O. and David, L.W., "GenBank," *Nucleic Acids Res*, **35**: 21–25 (2007)
93. Zhang, Z. Schäffer, A.A. Miller, W. Madden, T.L. Lipman, D.J. Koonin, E.V. and Altschul, S.F., "Protein sequence similarity searches using patterns as seeds," *Nucleic Acids Res*, **26**: 3986-3990 (1998)
94. Zdobnov, E.M. Lopez, R. Apweiler, R. and Etzold, T., "The EBI SRS server—new features," *Bioinformatics*, **18**: 1149–1150 (2002)
95. Guenter, S. Wendy, B. Alexandra, V.D.B. Maria, G.P. Carola, K. Tamara, K. Rasko, L. Quan, L. Vincent, L. Rodrigo, L. Renato, M. Francesco, N. Peter, S. Mary, A.T. Katerina, T. and Robert, V., "The EMBL Nucleotide Sequence Database: major new developments," *Nucleic Acids Res*, **31**: 17–22 (2003)
96. Pearson, W.R., "Using the FASTA program to search protein and DNA sequence databases," *Methods. Mol. Biol*, **24**: 307–331 (1994)
97. Smith, R.F. and Waterman, M.S., "Comparison of biosequences," *Adv. Appl. Math*, **2**: 482–489 (1981)
98. Paul, J. and Robin, W., "Genetics and Forensics: Making the National DNA Database," *Sci Stud*, **16**: 22–37 (2003)
99. Lucia, B. and Pietro, L.," Forensic DNA and bioinformatics," *Briefings in Bioinformatics*, **8**: 117-128 (2009)
100. Franke, K. Srihari, and Sargur., "Computational Forensics: Towards Hybrid-Intelligent Crime Investigation". *Third International Symposium on Information Assurance and Security IAS*, 383–386 (2007)
101. Mark, B., "DNA typing in forensic medicine and in criminal investigations: a current survey," *Naturwissenschaften*, **84**: 181–188(1997)
102. Howard, D.C. Jonathan, W.H. and Amy, J. S., "Development under extreme conditions: Forensic bioinformatics in the wake of the world trade center disaster," *Pac. Symp. Biocomput*, 638-653 (2003)
103. Barrett, J.C. Fry, B. Maller, J. and Daly, M.J.,"Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, **21**: 263-265(2005)
104. Leclair, B. Shaler, R. Carmody, G. R. Eliason, K. Hendrickson, B. C. Judkins, T. Norton, M. J. Sears, C. and Scholl, T., "Bioinformatics and human identification in mass fatality incidents: the world trade center disaster," *J. Forensic Sci*, **52**: 806-819 (2007)
105. Friedman, N. Linial, M. Iftach, N. and Dana, P., "Using Baysian Networks to analyze expression data," *J. Computational Bio*, **7:** 601-620 (2000)
106. Susan, H., "Bootstrapping Phylogenetic Trees," *Theory and Methods. Statistical Science*, **18**:241-255 (2003)