Advanced Techniques in Biology & Medicine

Kaushik et al., Adv Tech Biol Med 2016, 4:3
DOI: 10.4172/2379-1764.1000185

Mini Review Open Access

Genomic Databases and Softwares: In Pursuit of Biological Relevance through Bioinformatics

Mahima Kaushik[1,2]*, Swati Mahendru[2], Mohan Kumar[2], Swati Chaudhary[2] and Shrikant Kukreti[2]*

[1]*Cluster Innovation Centre, University of Delhi, Delhi, India*
[2]*Nucleic Acids Research Laboratory, Department of Chemistry, University of Delhi, Delhi, India*

Abstract

With the completion of human genome project, a plethora of information had been available for exploring various unanswered questions related to cell and molecular biology. Bioinformatics has been instrumental in unravelling the genetic, phenotypic, structural and functional aspects of the whole genome by using this information. Genomics and proteomics have become one of the most relevant fields after the advancements in computational analysis, interpretation, and modelling software. It can not only quite categorically describe the position of nucleotides and amino acids throughout genomes and proteomes respectively, but it also helps in performing the phylogenetic analysis, search for associated transcription factors, multiple sequence alignments, and many other relevant explorations/hunts. The advances in the knowledge of genetics acquired from molecular biology and bioinformatics are applied and point towards potential therapeutic strategies such as genome editing. This review has an aim of discussing some of the bioinformatics databases and software, which has been utilized for exploring the position of a DNA sequence, any associated single nucleotide polymorphism (SNP) related to a disease, on or nearby situated transcription factor binding sites followed by multiple sequence alignment of this sequence with other organisms. This study provides the insights in to the functional elements of any DNA, RNA or Protein sequence prior to exploring the structural polymorphism, which may regulate the gene expression. Also, this review briefly discusses the tools used for programmable nuclease–based genome editing technology.

Keywords: Bioinformatics; Single nucleotide polymorphism; DNA mutations; Transcription factors; Genome editing tools; CRISPR; Zinc finger nucleases

Introduction

In the last few decades, molecular biology has evolved and matured with the vast amount of biological data generated after the completion of Human Genome Project. In order to explicate the potential of this crucial information, storing and organizing this huge amount of data is a prerequisite. A large number of attempts have continuously been made to decipher the intertwined relationships of structure and function of the cell and its organelles. For getting the information on structure, function, mechanisms and pathways related to biomolecules, computational modelling had already been explored as a versatile tool. A large number of bioinformatics tools for discovering information from various genomic databases have quite recently been explored [1]. Bioinformatics is actually a branch of science in which biology and information technology merge with the computational algorithms and statistical data to form a single discipline and aims to determine the key biological mechanisms at both molecular and cellular level [2,3]. Comprehensive open access databases like GenBank, EMBL, DDBJ, and UNIGENE, have been the major storage of information in terms of DNA, RNA as well as their translated proteins, which is then utilized for getting the clues related to various cellular processes. Genetic testing is used for identifying chromosomes, genes, and proteins, while gene sequencing determines the exact position and order of all the nucleotides or bases in a DNA sequence. Strong evidences suggest that there are approximately 19,000 human protein coding genes which are well below the initial estimations of 100,000 genes [4]. So our target is to demystify the functioning of the complex human machinery by using such a small number of protein-coding genes. Computational modelling has quite widely been used for studying the molecular mechanisms at the cellular level even in a crowded environment having other cell organelles and membranes [5]. In addition, it is also used as a synthetic or systems biology technique for creating cell like circuits [6].

Bioinformatics softwares like Chem Genome 2.0 have been used on a large scale to identify genes along with their possible location across the genome. The position of small DNA, RNA or protein sequences has been identified using BLAST throughout the genome in various organisms. Many computer softwares like TFBSTools, TESS, Alibaba 2, etc. have been used for finding out the associated transcription factors with these sequences. Bioinformatics tools such as TransmiR and TFmiR analyze the post-translational regulatory potential of miRNAs. The availability of such regulatory units on or near these sequences determines their possible role in the control of gene regulation. Apart from the regulatory units on or near the DNA/RNA sequence, mutations or genetic variations impart clinical significance to them. Genetic variations cause the heritability of complex human traits, such as susceptibility to common diseases as well as phenotypic traits [7]. They can be searched through various single nucleotide polymorphism (SNP) databases like human gene mutation database (HGMD), dbSNP, COSMIC, etc. It is always very difficult to figure out SNPs in non-coding regions, as SNPs cannot make changes directly in the amino acids of the proteins in the non-coding region [8]. HaploReg is a new database developed by Ward and Kellis in 2012, for systematic mining of imputed variants, cell types, regulators and target genes responsible for complex human pecularities and disease amid the sea of available GWAS (Genome-wide association studies)

***Corresponding author:** Dr. Mahima Kaushik, Associate Professor, Cluster Innovation Centre, University of Delhi, Delhi, India, Tel: +911127666702; E-mail: mkaushik@cic.du.ac.in\kaushikmahima@yahoo.com

Prof. Shrikant Kukreti, Department of Chemistry, University of Delhi, Delhi, India, Tel: +911127666726; E-mail: skukreti@chemistry.du.ac.in\kukretishrikant@yahoo.com

Received July 17, 2016; **Accepted** July 28, 2016; **Published** August 04, 2016

Citation: Kaushik M, Mahendru S, Kumar M, Chaudhary S, Kukreti S (2016) Genomic Databases and Softwares: In Pursuit of Biological Relevance through Bioinformatics. Adv Tech Biol Med 4: 185. doi: 10.4172/2379-1764.1000185

[9,10]. It is very useful to explore the information of linked variants and their associated annotations [11]. SNP's related to different diseases like cancer, beta-thalassemia, muscular dystrophy, cerebellar ataxia etc. have already been studied quite in detail. Some of the SNP's have also been studied for exhibiting structural polymorphs like hairpin, cruciform, slipped structures, duplex, triple, quadruplex, etc. [12-14]. Most of these DNA structural polymorphs have already shown there *in vivo* existence and may further be responsible for controlling gene expression. Pairwise or multiple sequence alignment of these DNA sequences from various organisms is then performed using programs like Clustal, PSI/TM-Coffee, MAFFT, etc. for determining their interdependent relationship and conserved nature in different species [15]. Genome-wide association studies (GWAS) have newly been added to this field for studying the molecular processes at the gene level [16]. Exponential growth in the biological data obtained from various bioinformatics tools enhanced our understanding of role of genetics in human health by facilitating our basic knowledge of disease and its related mechanisms. It also, thereby improved and advanced the development of potential therapeutic strategies. Genome editing is one such therapeutic strategy that can amend the DNA/RNA inside cells and tissues, which are affected with certain diseases. Hence, genome editing has been explored for its potential not only for the treatment of monogenic disorders, but also for diseases which are infectious or have both genetic as well as environmental association. A summary of some of these bioinformatics databases and softwares has been elucidated in Figure 1 and some of these are briefly discussed in the following sections.

Covering such a broad topic in a review was really impractical; hence this review had focussed only on an aim of discussing a few of the bioinformatics tools. These tools have been utilized for the identification of any possible biological application of a DNA/RNA/Protein sequence. This information may further be utilized in identifying the relevance of the structural polymorphs of the biopolymers in the regulation of gene expression and its related cellular processes.

Basic Local Alignment Search Tool (BLAST)

The foremost indication about the characteristics of a freshly sequenced gene can be attained by sequence homology which is defined as comparison of a particular sequence with the sequences stored in a database. The most extensively used tool for sequence searching is BLAST, which is made available online at NCBI (National Center for Biotechnology Information; *http://www.blast.ncbi.nih.gov*), and was developed by Altschul et al. in 1990 [17-18]. BLAST provides primary information about the nucleic acid or protein sequence before carrying out any biological studies. It is employed to search an unknown sequence by performing sequence similarity with other sequences stored in a biological database at a very fast speed. BLAST is a user-friendly program which helps in determining the nucleic acid as well as amino acid sequences. *Megablast* is used for faster determination of closely related matches, while *Blastn* is comparatively slow and used for diverged sequences [19]. A nucleotide sequence can be searched against a nucleotide sequence database with the help of BLASTN, whereas BLASTP is used to compare a protein sequence against a protein sequence database. The nucleotide sequence can also be translated and searched against protein database with the help of BLASTX. Statistics of likelihood of a match can also be determined using BLAST. Genomes of distinctive species can be searched in BLAST, which also has a nucleotide database constituting approx. 44 billion bases of NCBI RefSeq and GenBank nucleotide sequence [20].

PSI-BLAST and DELTA-BLAST are used for sensitive protein-protein searches. PSI-BLAST gathers information from protein search followed by the development of the position-specific scoring matrix (PSSM). Reverse position-specific BLAST (RPSBLAST) searches a PSSM database for a protein query at a very fast speed [21]. Recently, Ling and Benkrid had developed the improved version of the NCBI-BLAST known as GPU-BLAST which combines Graphics Processing Unit (GPU) with the BLAST [22,23]. It provides supercomputing ability to the computer, thus accelerating the use of parallel algorithms

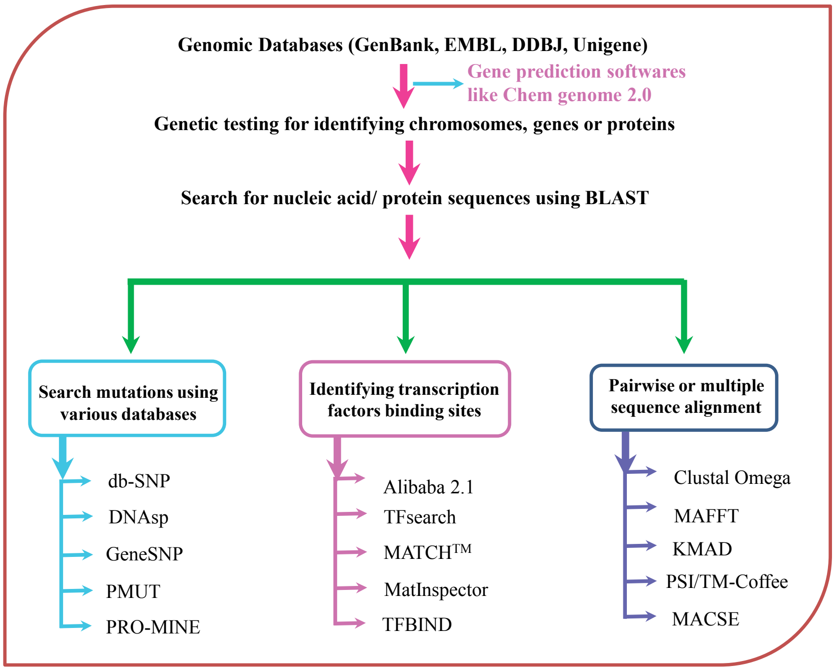
Genomic Databases (GenBank, EMBL, DDBJ, Unigene)
Gene prediction softwares like Chem genome 2.0
Genetic testing for identifying chromosomes, genes or proteins
Search for nucleic acid/ protein sequences using BLAST
Search mutations using various databases
Identifying transcription factors binding sites
Pairwise or multiple sequence alignment
db-SNP
DNAsp
GeneSNP
PMUT
PRO-MINE
Alibaba 2.1
TFsearch
MATCH™
MatInspector
TFBIND
Clustal Omega
MAFFT
KMAD
PSI/TM-Coffee
MACSE

Figure 1: Summary of some bioinformatics tools utilized for the predicting the biological applications of nucleic acids and proteins [42-44,50,51,54,68-74].

by researches. NCBI-BLAST can be speed-up by up to four times with the help of GPU-BLAST. The existing version of GPU-BLAST works only for protein sequence query (BLASTP). BLAST has been coined as the most important bioinformatics tool which enables the researchers to understand the function of proteins and genes present in the genome. Continuous improvement of BLAST website is in progress for developing more sensitive and fast bioinformatics tool for better and user-friendly search.

Regulatory Potential Analysis of DNA Sequences

The central dogma of molecular biology is that DNA produces RNA (transcription) which in turns produces the protein (translation). Transcription is the primary and central point of regulating the gene expression [24]. Interestingly, all genes are not expressed in each and every cell. It depends on intricate mechanisms of transcriptional regulation determined by the interaction of specific proteins known as the transcription factors with the particular DNA sequences. Thus transcription factors are recognized as a key developmental regulator which explains the basis of gene expression, cell differentiation and homeostasis [25]. There are quite a few bioinformatics databases which can help us to identify transcription factor binding sites (Figure 2).

TRANSFAC is a database of eukaryotic transcription factors along with their target genes and regulatory binding sites. Public release of this database is available at http://www.gene-regulation.com. It was developed more than two decades ago to represent the basic interaction between transcription factors and DNA binding site. Over the years its contents and features have been advanced, altered and augmented. Its latest version embodies the features such as expression patterns for transcription factors from *human* and *mouse*, *GENE Table* with links to *LocusLink*, *RefSe* and *OMIM* [26,27]. A number of bioinformatics tools such as TESS (*http://www.cbil.upenn.edu/tess*) [28], Match (http://www.gene-regulation.com/pub/programs.html#match) [27], AliBaba2.1 (*http://www.gene-regulation.com/pub/programs.html#alibaba2*) [29], are based on *TRANSFAC* database for predicting transcription factor binding sites (TFBS) in DNA sequences.

JASPAR is the principal open-access database of annotated, high-quality, matrix-based nucleotide profiles which describes the binding preference of transcription factors for eukaryotes. It is accompanied by a web interface for searching and sub setting, an online utility for sequence analysis as well as a collection of programming tools for genome-wide and comparative genomic explorations of regulatory regions. It can be accessed via *http://jaspar.genereg.net* [30,31]. *TESS* (*http://www.cbil.upenn.edu/tess*) [28] and *TFBSTools* (*http://bioconductor.org/packages/TFBSTools*) [32] are examples of bioinformatics tools, which make use of *JASPAR* database for speculating the transcription factor binding sites in DNA sequences.

DNA-binding domain (DBD) is a database which browses through all the publicly available proteomes and speculates all the sequence-specific DNA-binding transcription factors (TFs) from them. The current version of DBD consists of over 700 proteomes. It provides genome-wide transcription factor predictions for organisms. Transcription factor predictions in this database can be accessed through *http://transcriptionfactor.org* [33,34].

Regulatory Potential Analysis of micro RNA Sequences

MicroRNAs (miRNAs), endogenous small noncoding regulatory RNAs (~ 24 nt), present in the biological system regulate the transcription of a large number of human protein-coding genes associated with biological processes such as regulation of gene expression, cellular metabolism, development, tissue homeostasis and processing of information [35]. MicroRNAs (miRNAs) control gene expression at the posttranscriptional level via annealing to transcripts of protein-coding genes, resulting in cleavage or translation inhibition of the target mRNAs [36] (Figure 2). Some databases like *miRecords* (*http://c1.accurascience.com/miRecords/*) [37] and *DIANA-TarBase* (*http://www.microrna.gr/tarbase*) provide information regarding the target genes of miRNA (miRNA → genes) are [38]. The expression of miRNAs can be regulated by transcription factors (TFs). *TransmiR* (*http://cmbi.bjmu.edu.cn/transmir*) is a database which provides information on which transcription factors regulate miRNA (TFs → miRNA). In addition, transcription factors and miRNA can work in a synergistic way or form feedback loops to regulate gene expression [39]. *TFmiR* (*http://service.bioinformatik.uni-saarland.de/tfmir*) explores four different types of interactions, TF → gene, TF → miRNA, miRNA → miRNA and miRNA → gene [40].

Evaluation of Genetic Variants

DNA mutations are considered as potential targets to understand

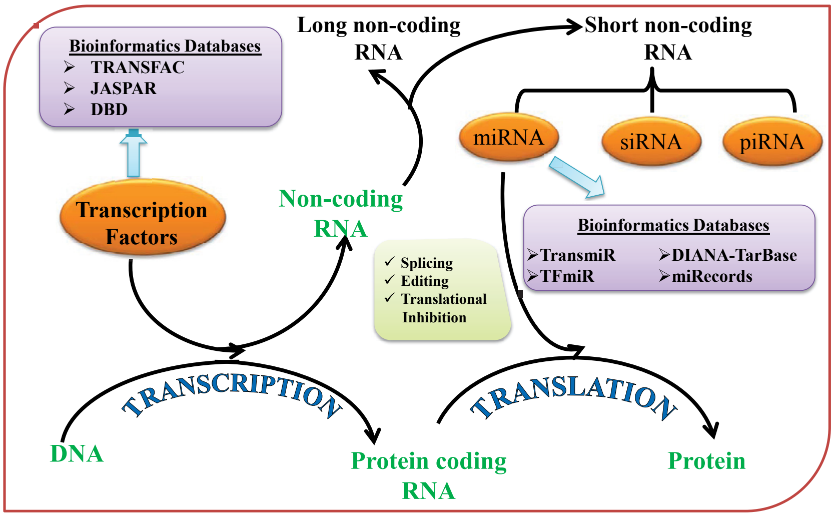

Figure 2: Central Dogma of molecular biology describing the bioinformatics databases for assessing the regulatory units at transcriptional and translational level.

the functional role of specific genomic regions. Developing bioinformatics tools to understand DNA polymorphisms are proved to be a relevant area of research. Various algorithms and tools have been developed to retrieve a huge data from DNA mutation analysis. *DNAsp* (DNA sequence polymorphism) is a bioinformatics tool used for the analysis of DNA polymorphisms. Conserved DNA regions can also be identified using the latest version of DNAsp. Also, multiple DNA sequences can be analyzed at the same time [41]. *GeneSNPs* is another database to gather information about DNA mutations [42]. Similarly, mutations found in proteins can be analysed using *PMUT* bioinformatics tool. It can help in detecting disease-related mutations. This software gathers information from the local databases and then examines the polymorphism present in a particular protein [43]. *PRO-MINE* (Protein Mutations in Neurodegeneration) is a database having a massive amount of data regarding disease-related mutations found in the TDP-43 gene [44]. Information about DNA repair mechanisms can be retrieved from *REPAIRtoire* database which provides data about the correlation between mutations found in genes and human diseases. Other databases related to DNA repair are *KEGG* (Kyoto Encyclopedia of Genes and Genomes) and repairGENES. Cross-linked databases such as KEGG PATHWAY, KEGG DISEASE, KEGG ORGANISMS and KEGG GENES collectively form KEGG software. DNA repair pathways including *NER*, *MMR*, and *BER* are involved in KEGG. On the other hand, *repairGENE*S database include genes encoding proteins which participate in DNA repair mechanisms [42].

Genome-wide association studies (GWAS) furnish a huge data having genetic variants with familiar phenotypes. A confusing point regarding this type of study is linkage disequilibrium (LD), which authorize multiple variants at the same locus to be affiliated with single phenotype [9]. *HaploReg* is a tool which is used for analyzing annotations of the non-coding genome at variants on haplotype blocks [45]. HaploReg revamps on SNAP, by supplying LD calculation of 1000 Genomes Project, and monitor insertion or deletion of bases in the DNA associated with query SNPs [9]. HaploReg elucidates the SNPs with progressive constraint measures, predicts the chromatin states, and how SNPs make variations in the Positional Weight Matrices (PWM) of a particular transcription factor. HaploReg annotation of GWAS has been used for haplotype fine-mapping and enrichment analysis [46,47]. HaploReg v4 is the updated version of the HaploReg, which is having the higher number of chromatin states map reference epigenomes from ENCODE (Encyclopedia of DNA Elements) 2012 and Roadmap Epigenomics, integrating regular binding data [10].

Multiple Sequence Alignment

Availability of vast amount of high-quality data for human as well as other large number of genomes has turned the interest of a lot of researchers towards comparing biological sequences (DNA, RNA or protein) from two or more genomes. The fundamental procedure used for comparing is known as Multiple Sequence Alignment (MSA) and plays an essential role in exploring the highly conserved sub-regions among a set of biological sequences, thereby, identifying functionally or structurally important sites and understanding the evolutionary history of some species from their associated sequences [48]. The alignment of sequences is shown in the form of a matrix with each row denoting a single sequence. The aim of such alignment is to position the sequences in such a way that it contains the least number of mismatched nucleotides and place the homologous nucleotides in the same column while representing the insertions and deletions in the form of gaps [49]. Several widely used packages are available for MSA, e.g. Clustal, MAFFT, KMAD, PSI-TM-Coffee etc.

Clustal is one of the oldest of the currently most widely used programs. Clustal Omega is the latest version of Clustal released in 2011. The web interface of Clustal Omega is available at *http://www.ebi.ac.uk/Tools/msa/clustalo/*. It is considered better than other packages because of its lesser execution time, better quality and more accuracy. It permits the reuse of the previously used alignments and hence reduces recomputing time of the same alignment for second exercise. New sequences can be made just by adding another set of sequences on the existing alignment itself. Also, existing alignments can be used to help align new sequences [50,51].

A multiple sequence alignment program, ***MAFFT*** was developed by Katoh et al. in 2002 [52]. It drastically reduced the execution time, however, there was a greater risk of over-alignment i.e. alignment of unrelated sequences due to increasing number of low-quality sequences in protein databank. New feature to suppress this over alignment is introduced in MAFFT version 7.263 and higher by utilizing agreement between the local segments and the entire sequence to determine which residues should be aligned and gapping the dissimilar segment by using a variable scoring matrix (VSM). MAFFT web server is accessible at http://mafft.cbrc.jp/alignment/software/ [53].

Knowledge-based Multiple Sequence Alignment (*KMAD*) had been designed for specifically aligning intrinsically disordered proteins (IDPs), lacking a stable tertiary structure as well as an active site. Short linear motifs (SLiM) which are found in abundance in the disordered region are being used for facilitating interaction between IDPs and other proteins. KMAD incorporates the main function determinants of IDPs, that is, SLiM, domain and posttranslational modifications (PTM) annotations in the alignment procedure. KMAD web server is accessible at *http://www.cmbi.ru.nl/kmad/* [54].

PSI/TM-Coffee is designed for multiple sequences Alignment of proteins based on combined strategies of homology extension along with the consistency-based alignment approach. The PSI / TM-Coffee web server is part of the T-Coffee web platform accessible from *http://tcoffee.crg.cat/tmcoffee*. It features the T-Coffee homology extension procedure [55] for both regular proteins (Position Specific Iterative T-Coffee, PSI-Coffee) and trans membrane proteins (TM-Coffee) [56].

Genome Editing Tools

Genome editing is the technique in which a gene sequence is precisely targeted and then the nucleotide sequence of the genome is altered with high efficiency. This technology is highly useful in treatment of monogenic disorders, infectious diseases as well as diseases that have both genetic as well as environmental association. Initially, nuclease-based genome editing creates a specific double strand break in the genome and then cell's own endogenous repair machinery is permitted to repair the break. This break can be repaired by cell, utilizing any one of these two basic mechanisms: non-homologous end-joining (NHEJ) or homologous recombination (HR) [57].

Genome editing tools based on programmable nucleases can be broadly categorized on the basis of their DNA recognition mode in two classes. Meganucleases, Zinc Finger Nucleases (ZFNs) and Transcription Activator–Like Effector Nucleases (TALENs) come under the nuclease class, which specifically binds to DNA via DNA-protein interactions. The Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated nuclease Cas9 targets specific DNA sequences by a short RNA guide molecule that forms RNA–DNA Watson-Crick base pairing as well as by protein-DNA interactions [58,59].

CRISPR/Cas9: CRISPR (Clustered Regularly Interspaced Short

Palindromic Repeats) is a gene editing tool to facilitate genome engineering in plants and animals. It was established to provide adaptive immunity of bacteria and archaea to plasmids and viruses. Cas9, a CRISPR-associated protein is a nuclease that can be governed by short RNA sequences to carry out cleavage at a particular genomic location in human and mouse cells [60,61]. RNA duplex containing guide sequence conserves two characteristics such as 5'-end sequence that help in determining specific target site of DNA and a RNA-duplex structure formed at 3'-end that provides binding site to Cas9 [62]. In any organism of choice, DNA sequences can be modulated or edited easily with the help of this tool. This method enables scientists to interpret the functional and structural framework of the genome and it also relates the genetic variations with biological phenotypes [63]. A simple software tool, CRISPRdirect (http://crispr.dbcls.jpl/) has been designed to effective selection of RNA targets for CRISPR/Cas systems [64]. Apart from efficient cleavage of specific target DNA sites, off-target sites also gets cleaved sometimes, which is a major problem for researchers. SgRNAcas9 is latest software which diminishes the effects of off-targets and allows more specific selection of DNA target sites. It is freely available at BioTools website (www.biotools.com) [65].

Zinc-finger nucleases: Zinc-finger nucleases (ZFNs) are designed restriction enzymes which are responsible for the catalytic cleavage of DNA. It has an exceptional capability to perform genomic modifications or expression. Various databases and softwares are available to study the potential zinc finger nucleases. Zinc Finger Consortium has developed software ZiFiT; a tool for finding the potential zinc finger nuclease sites in a particular sequence [66]. It is available on http://zifit.partners.org/ZiFiT/. The latest version ZiFiT Targeter Version 4.2 includes the identification of CRISPR/CAS targeted sites and reagents. ZFNGenome (http://bindr.gdcb.iastate.edu:88/) is a GBrowse-based tools which effectively can visualize and identify target sites for OPEN-generated zinc finger nucleases. It provides information of potential zinc finger nucleases target sites which also include their chromosomal location and their position relative to transcription initiation sites [67].

Outlook and Future Directions

Bioinformatics and computational modelling for finding out the interdependent relationships between different genes and their protein products for solving puzzles related to replication, transcription, translation and other cellular processes have now paved its way in to the advanced stage. For dealing with the challenges related to the understanding of various biological mechanisms and pathways, bioinformatics is playing an exceptionally crucial role. It not only provides the information related to the position of any nucleic acid sequence throughout the gene but also elaborates upon all the proteins which may be binding near or on the same, ultimately leading to control of gene expression. For correlating it with any functional relevance, meticulous approaches of systems biology have now been paid significant attention. One such approach is genome editing which holds great promise in recent advances made in the field of precise gene targeting and modification with the help of programmable nucleases. All of these bioinformatics, computational and system biology related studies have not only been utilized in exploring the complicated circuitry of the biological systems but have also been providing clues for finding any alternative approach of more effective intervention in the field of medicine.

Acknowledgement

Authors would like to acknowledge the Research and Development Grant (2015-2016) of University of Delhi, Delhi.

References

1. Teufel A, Krupp M, Weinmann A, Galle PR (2006) Current bioinformatics tools in genomic biomedical research (Review). Int J Mol Med 17: 967-973.
2. Mahendru S (2012) Bioinformatics: A bridge between genetics and chemistry. Research News For U 8: 92-94.
3. Lio P, Bishop MJ (2005) Nucleic acid and protein sequence analysis and bioinformatics. Reviews in Cell Biology and Molecular Medicine.
4. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, et al. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23: 5866-5878.
5. Im W, Liang J, Olson A, Zhou HX, Vajda S, et al. (2016) Challenges in structural approaches to cell modeling. J Mol Biol.
6. Wagner HJ, Sprenger A, Rebmann B, Weber W (2016) Upgrading biomaterials with synthetic biological modules for advanced medical applications. Adv Drug Deliv Rev.
7. Peters DT, Musunuru K (2012) Functional evaluation of genetic variation in complex human traits. Hum Mol Genet 21: R18-R23.
8. Appasani K (2015) Genome-wide association studies: From polymorphism to personalized medicine. Cambridge University Press, India.
9. Ward LD, Kellis M (2012) HaploReg: A resource for exploring chromatin states, conservation and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 40: 930-934.
10. Ward LD, Kellis M (2016) HaploReg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res 44: 877-881.
11. Zeggini E, Morris A (2015) Assessing rare variation in complex traits: Design and analysis of genetic studies. Springer, USA.
12. Kaushik M, Kukreti R, Grover D, Brahmachari SK, Kukreti S (2003) Hairpin–duplex equilibrium reflected in the A? B transition in an undecamer quasi-palindrome present in the locus control region of the human ß-globin gene cluster. Nucleic Acids Res 31: 6904-6915.
13. Kaushik M, Kukreti S (2006) Structural polymorphism exhibited by a quasipalindrome present in the locus control region (LCR) of the human ß-globin gene cluster. Nucleic Acids Res 34: 3511-3522.
14. Kaushik M, Kaushik S, Roy K, Singh A, Mahendru S, et al. (2016) A bouquet of DNA structures: Emerging diversity. Biochem Biophys Rep 5: 388-395.
15. Abd-Elsalam KA (2003) Web-based bioinformatics resources for protein and nucleic acids sequence alignment. Afr J Biotechnol 2: 714-718.
16. Lehne B, Lewis CM, Schlitt T (2011) From SNPs to genes: Disease association at the gene level. PLoS One 6: e20133.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
18. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, et al. (2008) NCBI BLAST: A better web interface. Nucleic Acids Res 36: W5-W9.
19. Ladunga I (2009) Finding homologs in amino acid sequences using network BLAST searches. Curr Protoc Bioinformatics Chapter 3: Unit 3.
20. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, et al. (2013) BLAST: A more efficient report with usability improvements. Nucleic Acids Res 41: W29-W33.
21. Madden T (2013) The BLAST sequence analysis tool. The NCBI Handbook [Internet] 2nd edition. Bethesda (MD): National Center for Biotechnology Information, USA.
22. Ling C, Benkrid K (2010) Design and implementation of a CUDA-compatible GPU-based core for gapped BLAST algorithm. Procedia Comput Sci 1: 495-504.
23. Vouzis PD, Sahinidis NV (2011) GPU-BLAST: Using graphics processors to accelerate protein sequence alignment. Bioinformatics 27: 182-188.
24. Latchman D, Latchman DS (2010) Eukaryotic transcription factors. Academic press, USA.
25. Hughes TR (2013) A handbook of transcription factors. Springer Science & Business Media, USA.

26. Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Res 24: 238-241.

27. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, et al. (2003) TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378.

28. Schug J (2008) Using TESS to predict transcription factor binding sites in DNA sequence. Curr Protoc Bioinformatics Chapter 2: Unit 2.

29. Grabe N (2002) AliBaba2: Context specific identification of transcription factor binding sites. In Silico Biol 2: S1-S15.

30. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, et al. (2014) JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res 42: 142-147.

31. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-D94.

32. Tan G, Lenhard B (2016) TFBSTools: An R/bioconductor package for transcription factor binding site analysis. Bioinformatics 32: 1555-1556.

33. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD--taxonomically broad transcription factor predictions: New content and functionality. Nucleic Acids Res 36: D88-D92.

34. Kummerfeld SK, Teichmann SA (2006) DBD: A transcription factor prediction database. Nucleic Acids Res 34: D74-D81.

35. Ultsch A, Lötsch J (2014) What do all the (human) micro-RNAs do? BMC Genomics 15: 976.

36. Wójcicka A, Kolanowska M, Jażdżewski K (2016) Mechanisms in endocrinology: MicroRNA in diagnostics and therapy of thyroid cancer. Eur J Endocrinol 174: R89-98.

37. Xiao F, Zuo Z, Cai G, Kang S, Gao X, et al. (2009) miRecords: An integrated resource for microRNA-target interactions. Nucleic Acids Res 37: D105-D110.

38. Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, et al. (2015) DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA: mRNA interactions. Nucleic Acids Res 43: D153-D159.

39. Wang J, Lu M, Qiu C, Cui Q (2010) TransmiR: A transcription factor-microRNA regulation database. Nucleic Acids Res 38: D119-D122.

40. Hamed M, Spaniol C, Nazarieh M, Helms V (2015) TFmiR: A web server for constructing and analyzing disease-specific transcription factor and miRNA co-regulatory networks. Nucleic Acids Res 43: 283-288.

41. Librado P, Rozas J (2009) DnaSP v5: Software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

42. Milanowska K, Rother K, Bujnicki JM (2011) Databases and bioinformatics tools for the study of DNA repair. Mol Biol Int 2011: 475718.

43. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) PMUT: A web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176-3178.

44. Pinto S, VlahoviÄek K, Buratti E (2011) PRO-MINE: A bioinformatics repository and analytical tool for TARDBP mutations. Hum Mutat 32: E1948-1958.

45. Ritchie GR, Flicek P (2014) Computational approaches to interpreting genomic sequence variation. Genome Med 6: 87.

46. Lee MN, Ye C, Villani AC, Raj T, Li W, et al. (2014) Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science 343: 1246980.

47. Roadmap Epigenomics Consortium; Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317-330.

48. Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. J Comput Biol 1: 337-348.

49. Elias I (2006) Settling the intractability of multiple alignment. J Comput Biol 13: 1323-1339.

50. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

51. Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 1079: 105-116.

52. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

53. Katoh K, Standley DM (2016) A simple method to control over-alignment in the MAFFT multiple sequence alignment program. Bioinformatics 32: 1933-1942.

54. Lange J, Wyrwicz LS, Vriend G (2016) KMAD: knowledge-based multiple sequence alignment for intrinsically disordered proteins. Bioinformatics 32: 932-936.

55. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205-217.

56. Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, et al. (2016) PSI/TM-Coffee: A web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic Acids Res 44: W339-343.

57. Porteus MH (2015) Towards a new era in medicine: therapeutic genome editing. Genome Biol 16: 286.

58. Cox DB, Platt RJ, Zhang F (2015) Therapeutic genome editing: prospects and challenges. Nat Med 21: 121-131.

59. Jasin M (2016) Gene editing 20 years later: Genome Editing. Springer, New York.

60. Cong L, Ran FA, Cox D, Lin S, Barretto R, et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339: 819-823.

61. Mali P, Yang L, Esvelt KM, Aach J, Guell M, et al. (2013) RNA-guided human genome engineering via Cas9. Science 339: 823-826.

62. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346: 1258096.

63. Hsu PD, Lander ES, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157: 1262-1278.

64. Naito Y, Hino K, Bono H, Ui-Tei K (2015) CRISPRdirect: Software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics 31: 1120-1123.

65. Xie S, Shen B, Zhang C, Huang X, Zhang Y (2014) sgRNAcas9: A software package for designing CRISPR sgRNA and evaluating potential off-target cleavage sites. PLoS One 9: e100448.

66. Sander JD, Maeder ML, Reyon D, Voytas DF, Joung JK, et al. (2010) ZiFiT (Zinc Finger Targeter): An updated zinc finger engineering tool. Nucleic Acids Res 38: W462-468.

67. Reyon D, Kirkpatrick JR, Sander JD, Zhang F, Voytas DF, et al. (2011) ZFNGenome: A comprehensive resource for locating zinc finger nuclease target sites in model organisms. BMC Genomics 12: 83.

68. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, et al. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31: 3576-3579.

69. Cheng JC, Chang HM, Leung PC (2013) Egr-1 mediates epidermal growth factor-induced downregulation of E-cadherin expression via Slug in human ovarian cancer cells. Oncogene 32: 1041-1049.

70. Meng X, Kondo M, Morino K, Fuke T, Obata T, et al. (2010) Transcription factor AP-2beta: A negative regulator of IRS-1 gene expression. Biochem Biophys Res Commun 392: 526-532.

71. Akhtar H, Islam G, Jan SU, Nawaz A, Akhtar S, et al. (2015) Identification of essential regulatory elements responsible for the explicit expression of IL-28Ra and their effect on critical SNPs using *in-silico* methods. Pak J Pharm Sci 28: 1523-1532.

72. Taylor-Douglas DC, Basu A, Gardner RM, Aspelund S, Wen X, et al. (2014) Evaluation of hypothalamic murine and human melanocortin 3 receptor transcript structure. Biochem Biophys Res Commun 454: 234-238.

73. Liguori R, Quaranta S, Di Fiore R, Elce A, Castaldo G, et al. (2014) A novel polymorphism in the PAI-1 gene promoter enhances gene expression. A novel pro-thrombotic risk factor? Thromb Res 134: 1229-1233.

74. Liu Y, Li Y, Zhang L, Li M, Li C, et al. (2015) NF-Î°B downregulates Cbl-b through binding and suppressing Cbl-b promoter in T cell activation. J Immunol 194: 3778-3783.