CHAPTER 1

Bioinformatics Overviews

Ritu Pasrija

1 Background

In the 1970s, a Dutch theoretical biologist Paulien Hogeweg along with Ben Hesper, first coined the term bioinformatics. They were interested in accumulating information regarding biological systems. Their observation was that in addition to biochemistry and biophysics, it is worthwhile to recognise bioinformatics as a research area and has the potential to become 'biology of the future'. This became true as in these last 50 years, development of bioinformatics has happened at a very fast pace. Although for a particular interval, persons viewed bioinformatics as the software tools advancement method to support, accumulate, manoeuvre, and scrutinise biological information. Whilst this application is indeed a significant one in bioinformatics, this field has much more potential than that. Both 'bioinformatics' and 'computational biology' are instrumental in accumulating enormous information of several parts of natural science. So, it is important to understand the difference in these two terms. On the one hand, bioinformatics uses computer science,

R. Pasrija (✉)
Department of Biochemistry, Maharshi Dayanand University, Rohtak, India
e-mail: ritupasrija.biochem@mdurohtak.ac.in

S. Dutta and S. Gochhait (eds.), *Information Retrieval in Bioinformatics*,
https://doi.org/10.1007/978-981-19-6506-7_1

statistics to molecular biology and create computational & statistical techniques, which help in examination and management of biological data. On the other hand, computational biology uses computational simulation mode, mathematical models, and fundamentals in computer science, genetics, anatomy, biochemistry, and statistics among others. Amalgamation of these two has led to a new term called 'Systems biology', which combines organism-wide information of natural science for acquiring a comprehensive perception of a biological entity, like a bacterium. This led to creating synthetic genomes, and soon a synthesised cell would become a reality. Nonetheless, to understand the feasibility of this fancy hypothesis, it is important to revisit the key discoveries in biological sciences, which would also help in understanding the history of development of 'bioinformatics' as a separate branch.

2 History

During the 1950s, DNA and computers were not the important tools in research and in biochemistry, investigations were largely happening on mechanistic enzymes model. Many scientists in fact thought that proteins are the carriers of genetic information, as DNA seemed too simple to carry genetic information, whereas protein show a large number of alternatives and complexity.

This, the major turning point in bioinformatics has to be DNA being regarded as the genetic material. The first evidence for this came from experiments of Oswald Avery et al. (1944), who revealed that DNA regulates the characters in organisms, instead of proteins. This group studied the uptake of pure DNA from a virulent *Streptococcus pneumoniae (S. pneumonia)* bacterial strain, which has smooth round colonies (named S); which could bestow virulence to even a non-virulent strain (rough colonies, R) (Avery et al., 1944). Subsequent work by Alfred Hershey and Martha Chase (in 1952) validated these findings that DNA of bacterial cells infected by bacteriophages can be transmitted to other bacterium and alter the phenotype of recipient cell (Hershey & Chase, 1952). Later in 1953, James Watson, Francis Crick, and Rosalind Elsie Franklin finally proposed the double-helix structure of DNA (Franklin & Gosling, 1953; Watson & Crick, 1953). Further, it took additional 13 years in interpreting the amino acid codon and 24 additional years in improving the first DNA sequencing technique. Thus, 1970–1980 witnessed a paradigm shift from protein to DNA analysis. In 1970, Saul B. Needleman and

Christian D. Wunsch established their dynamic programming algorithm for the alignment of pair-wise protein sequences. After another decade of first multiple sequence alignment (MSA) algorithms was developed, its application was applied to other biological sequences (DNA and RNA). Therefore, development in DNA always lagged behind proteins. Comparison wise, the amino acids alignment is based on-identical, similar and dissimilar amino acids (based on their chemical properties), whereas only identical matches and mismatches are considered for DNA and RNA. The MSA and its use to sequence-structure-function relationship is so common and useful that from the 1980s onwards, the term 'bioinformatics' is mostly used to refer to the computational methods for genomic data analysis. This big transformation opened the possibility to sequencing whole genomes. Biologist, Fred Sanger, undertook the first genome sequencing in a bacterial virus, called bacteriophage $\phi \times 174$ (5368 base pairs). Later, Craig Venter in the 1980s sequenced the first organism, a bacterium *Haemophilus influenza*. Ernst Haeckel, in 1866, used DNA sequences in phylogenetic inference and reconstructed the first molecular phylogenetic trees from amino acids linear arrangements in proteins, which show the closeness and related ness among species during evolution.

Similarly, many more genomes are already sequenced (more than 1000), including of human, called "The Human Genome Project (HGP)", which completed in April 2001 by two independent groups (McPherson et al., 2001; Venter et al., 2001). These recent life science data explosions—such as genotyping, transcriptomics, or proteomics—also became possible with the availability of genomes and opened gates for new studies. The gold mine of enormous data later became freely available at European Molecular Biology Laboratory (EMBL) site (www.ensembl.org). Along with this, various bioinformatics tools also became available including on this site, like—BLAST, Ensembl, primer synthesis, phylogeny that rely on the accessibility of the cyberspace. Investigations on genomic sequences information, including humans unlocked the applied prospects like—discovery of drug and their targets, as well as individualised therapy. Thus, biologists end up being increasingly dependent on computational scripts, written in scripting languages, such as C, C++, Shell, Python, R, and Ruby. Before, we look at the research challenges, algorithms, big data, retrieval of information, and application of bioinformatics; it is crucial to understand the basics of biology and bioinformatics (Porter et al., 2021).

Basics of Bioinformatics: We know that all surviving entities are composed up of units called cells, which contain the genetic material (nucleic acid), and passed from one generation to the next. Many living systems are made up of only one cell (unicellular), and one cell is the whole organism; whereas in developed species like plants and animals, a life form has more than billions of cells. Apart from that, cells can be of two major types: prokaryotic cells and eukaryotic cells. The prokaryotic cells do not have nucleus and genetic material lies open inside the cell, whereas in eukaryotic cells genetic material is present inside a structure, called nucleus. Prokaryotic cells are generally unicellular, whereas eukaryotic cells can be either unicellular, like baker's yeast *Saccharomyces cerevisiae* (*S.cerevisiae*) or multicellular organisms, like humans.

The nucleic acid is of two kinds in nature: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Most organisms' genetic material is DNA, although few cells might have RNA, as in certain viruses. Both forms of nucleic acid are polymers, assembly of repeated units, called nucleotides. A nucleotide involves three segments: a base, a pentose sugar (ribose), besides a phosphate group. These bases vary in different nucleotides and are of four types: Guanine (G), Thymine (T), Adenine (A) and Cytosine (C). For RNA, the bases are identical like DNA, except T replaced with Uracil (U). RNA is a single helix, whereas DNA is a double-helix molecule, in which bases lie parallel to each other and form bond with each other, called hydrogen bond base pairing. During bond formation, T always pairs with A base through two hydrogen bonds (double bond) and C always pairs with G through three bonds (triple bond). This removes steric hindrance and stabilises the structure. When an RNA strand duo with another strand, the pairing followed is $A = U$ and $C \equiv G$. These nucleotides join one after another linearly and generate polymer.

***The Central Dogma*:** DNA is the controlling element, but needs to pass on the information to a kind of RNA, termed as messenger RNA (mRNA), through a process called transcription and subsequently data in mRNA is passed via translation to proteins. Three DNA/RNA nucleotides code for one amino acid, which polymerise in a linear fashion to make proteins. Like UUU code for an amino acid, phenylalanine and AUG code for methionine. There are 20 standard amino acids, which are written as alphabets in capital. The sequence of three DNA nucleotides (called codon) decides the amino acid incorporated. The nucleotides (U, C, A and G) in random sequence of three can give 64 possible combinations, thus one amino acid may be coded by more than one combination and grouped together. Among these 64, three codons act as stop codon (UAA, UGA and UAG), and once incorporated, they stop the translation, as they do not code for any amino acid.

The genetic material has different genes on it, which control the different characters in an organism, its fitness, etc. Thus, scientist developed fantasy for this molecule, as genetic material manipulation is possible, and desired characters, fitness, chances of survival is achievable. Similarly, it also opened avenues for gene manipulation to correct any disease phenotype. All this became possible, as research has given insight into function of the genes. This led to immense advancement of techniques to study genetic material, termed 'genomics'. Interestingly, the different cells of an organism although have same DNA, but do not express all the genes present on their genetic material. Thus, all cells do express some common genes, called 'house-keeping' genes, whereas some genes are transcribed to RNA and to translated proteins in exclusive cells or time interval. Like in foetal stage, the haemoglobin synthesised is different from that of an adult and are product of different genes. Once born, the foetal haemoglobin gene is not transcribed any more. These differential DNA expressions pattern led to different RNA profile in various cells and is studied under 'transcriptomics'. Similarly, the functional molecules, proteins presence, and their measure are covered in 'proteomics'. Apart from that, three-dimensional (3D) modelling of biomolecules and biological systems is also of interest to biological scientist.

All these advancements led to generation of a large amount of biological information among different organisms and species. Thus, biological scientists need to utilise computational and analysis tools for acquiring, understanding, interpreting and sharing of these data. These have led to the 'Bioinformatics' we know today, and became essential in organising information in modern biology, as well as treatment. Therefore, it is not exaggerating to conclude that bioinformatics is a multidisciplinary field, which connects biology with computer's knowledge, mathematics, statistics, and physics. This has led scientists in essentially acquiring a good knowledge of molecular biology, along with computer science for analysis of bioinformatics data.

The Human Genome: The name 'genome' exactly refers to total genes or whole of the DNA's content in an organism or a cell. A regular human cell encloses 23 couples of chromosomes (separate threads of DNA). The human genome has 22 autosomes, and female additionally has two copies of X chromosome (XX) and male has one X and one Y chromosome (XY). These 23 pairs of DNA threads have approximately ~ 20,000–25,000 genes in the human genome, which are generally protein coding, although sometimes only RNA is transcribed, which has regulatory role. Besides protein coding region, the controlling sequences are also there in genome, which includes—promoters, introns, intergenic (between-gene) regions, and repetitive sequences in the genome. On an average, more than half of the human genome is transcribed and translated, though a very small quantity of them is managed as mRNAs and study of all RNA transcripts is included under 'transcriptomics'.

3 Perspectives of Computer Science and Information Technology

As earlier stated, the sequencing techniques opened an era of loads of data, called big data and its usefulness relies on programming and software development, and building enormous datasets of biological information in research (Gochhait et al., 2021). This information is much more than easily and efficiently interpreted by a biology researcher. Further, different investigators may decipher the data in dissimilar ways, and even the same researcher may make varying explanations, resulting in erratic and non-uniform data processing. Sometimes, after interpretation, the rationale behind may be lost or only loosely remembered. Occasionally, a researcher exits a study group; the technique used to understand data also goes with them and vanishes. Lastly, the researcher's interpretation may be prejudiced towards getting a predetermined consequence. Thus, bioinformatics being a systematic study rule out all these anomalies in science and research. The important basis of bioinformatics includes.

Algorithms: These are the rules followed in computations and done on both Linux and Windows platform. Various tasks rely on particular algorithms, which are critical in both examining and accurately handling of the data. The biological scientists and bioinformaticians choose the computer science processes for sequencing, gathering, unravelling biological functions and relationships, which finally helps in interpretation of the information. These include DNA, RNA and protein alignments (could be local and global), gene prediction, phylogenetic tree construction database similarity search, motif detection, Markov chains or information entropy and sequence logos, molecular modelling, etc. The FASTA, BLAST, Spectral Forecast, Objective Digital Stains (ODSs), self-sequence alignment and Discrete Probability Detector (DPD) algorithm are few examples. A bioinformatics can use resources freely available over the internet and some are paid softwares. The literature references and various databases (genome, sequence, function structure) of molecular biology can be searched at PubMed, PubMed Central, NCBI, EBI, ExPASy, RSCB. Some important tools in Bioinformatics are introduced below:

1. **Recovery and exploration of sequence**: Linear sequence of DNA, RNA and proteins is used for sequence alignment and provides lots of information. It depends on particular algorithm like FASTA BLAST, CLUSTAL X/W, etc. These are helpful for identity, similarity and dissimilarity in sequences (homology), phylogeny search analysis and phylogeny tree construction for relatedness among species.

FASTA is a text-based arrangement for representing either base sequences or amino acid (protein) order, in which base or amino acids are symbolised with single-alphabet codes and blanks in between alphabets are not permitted. Matches are displayed in black and red are treated as mismatches in nucleotide alignment. It was developed by David J. Lipman and William R. Pearson in 1985 and is used in many programming languages like Python, PERL, Ruby, etc. A multiple sequence FASTA layout is obtainable by concatenating various single FASTA sequence in a single file. The FASTA sequence first line starts with a '>' (greater-than) symbol as shown below for two genes: Green fluorescent protein (GFP) from jellyfish and insulin from humans.

GFP gene sequence of *Aequorea victoria* in FASTA format

```
>AGTAAAGGAGAAGAACTTTTCACTGGAGTTGTGACAATTCTTGTTGAATTAGATGGTGAT
GTTAATGGTCACAAATTTTCTGTTAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAAC
TTACCCTTAAATTTATTTGTACTACTGGAAAACTACCTGTTCCCTGGCCAACACTTGTTAC
TACTTTGACTTATGGTGTTCAATGTTTTTCAAGATACCCAGATCACATGAAACGGCACGAC
TTTTTCAAGAGTGCAATGCCCGAAGGTTATGTACAAGAAAGAACTATTTTTTTCAAAGATG
ACGGTAACTACAAGACACGTGCTGAAGTTAAGTTTGAAGGTGATACCCTTGTTAATAGAAT
CGAGTTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCTTGGACACAAATTGGAATAC
AACTATAACTCACACAATGTATACATTATGGCAGACAAACAAAAGAATGGAATCAAAGTTA
```

```
ACTTCAAAATTAGACACAACATTGAAGATGGAAGTGTTCAACTAGCAGACCATTATCAACA
AAATACTCCAATTGGCGATGGCCCTGTTCTTTTACCAGACAACCATTACCTGTCCACACAA
TCTGCTCTTTCTAAAGATCCCAACGAAAAGAGAGACCATATGGTGCTTCTTGAGTTTGTAA
CAGCTGCTGGTATTACACACGGTATGGATGAACTATACAAACACCATCACCATCACCATCA
CTAG
```

Humans Insulin gene sequence in FASTA format

```
>AGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCT
TTGCGTCAGGTGGGCTCAGGATTCCAGGGTGGCTGGACCCCAGGCCCCAGCTCTGCAGCAGG
GAGGACGTGGCTGGGCTCGTGAAGCATGTGGGGGTGAGCCCAGGGGCCCCAAGGCAGGGCACC
TGGCCTTCAGCCTGCCTCAGCCCTGCCTGTCTCCCAGATCACTGTCCTTCTGCCATGGCCCTG
TGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCC
TTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAA
CGAGGCTTCTTCTACACACCCAAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACT
GCCCATTGCTGCCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTCCCACCCAGCATGG
GCAGAAGGGGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAGTT
CTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGAGGA
CGGTGTTGGCTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCCACTCGC
CCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAGATTCAAGTG
TTTTGTTAAGTAAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGG
GCGAACACCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAGTGGGCCAGACCCCTGT
CGCCAGGCCTCACGGCAGCTCCATAGTCAGGAGATGGGGAAGATGCTGGGGACAGGCCCTGGG
GAGAAGTACTGGGATCACCTGTTCAGGCTCCCACTGTGACCTGCCCCGGGGCGGGGGAAGGAG
GTGG
GACATGTGGGCGTTGGGGCCTGTAGGTCCACACCCAGTGTGGGTGACCCTCCCTCTAACCTGG
GTCCAGCCCGGCTGGAGATGGGTGGGAGTGCGACCTAGGGCTGGCGGGCAGGCGGGCACTGTG
TCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCGCCGCTGTTCCGGAACCTGCTCTGCGC
GGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTGCAGC
CCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCT
GCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCC
GCCGCCTCCTGCACCGAGAGAGATGGAATAAAGCCCTTGAACCAGC
```

Similarly, protein sequence in FASTA format for both proteins can be written and provided below—

GFP protein sequence from *Aequorea victoria* in FASTA format

```
>MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVT
TFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVNRIEL
```

```
KGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPI
GDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK
```

Insulin protein in Humans in FASTA format

```
>MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

BLAST: Its full form is **B**asic **L**ocal **A**lignment **S**earch **T**ool and this procedure catches the region of similarity between sequences. This was first proposed in 1990 by David J. Lipman and his team, and one of highly cited paper with more than 65,000 citations (Altschul et al., 1990). This program can compare both nucleotide (n-BLAST) and protein primary sequences (p-BLAST) in two different variants. The evaluation of sequence is finally used to calculate the statistical significance of matches. It is freely available on internet at 'https://blast.ncbi.nlm.nih.gov/Blast.cgi'. It can be performed on various operating systems like UNIX, Linux, Mac, and MS Windows and is written in C and C+ language. BLAST deduce useful and evolutionary relationships between linear arrangements of bases/amino acids, as well as help identify members of gene families. Like haemoglobin gene sequence in humans can be compared with that of mouse. The input sequence is generally provided in FASTA or gene bank format; whereas the output formats, include HTML, plain text, and XML. Besides online, the program is also available in free and paid download versions (BLAST+). The megablast and discontiguous megablast are other variants with separate applications.

Clustal W/X: is an algorithm for multiple sequence analysis (MSA) of DNA or proteins. It produces meaningful multiple sequence alignment of divergent species, and computes best matches for the chosen sequences and line them up so that the resemblances and the disparities can be understood by viewing the cladograms and phylograms, both of which are part of the Clustal W algorithm.

2. **Graph theory:** Graph theory, also called as graph methodology, takes the help of graphs for comparison. Graph is a set of vertices (lines), connected by edges (exist between two vertices), which could be directed (with arrow), undirected, weighed, etc. It is useful in showing networks or flow of communication and simplifies complex relationship.

3. **Artificial intelligence (AI):** Here the bioinformatics functions impersonate the brainpower of the human with computers. Its usefulness is understandable by realising the need and importance in whole genome sequencing, sequence reconstructions, and gene finder. It also generates vital tools for data processing.
4. **Data mining:** Data mining is vital for extrapolation and meaningful significant information can be extracted from huge datasets. The multifaceted arrays of data can be used as input and utilise a variety of arithmetical and numerical techniques to unearth surprising inconsistencies, like grouping and bundling algorithms for any process. An example includes gene annotation in whole genome sequence, domain, and motif discovery. Similarly, mass spectroscopy can classify proteins, although hindrance in data mining may come from variances in complexity, scale, number and the lack of an accepted ontology.
5. **Soft computing:** Many devices are upgraded for storing the biomedical information; still computers have its significant role and hold a special place with researchers and biologists. This has the unique progress of expressing the data of the gene. In addition, it also expresses the bioinformatics data. This evolves with neural network model and artificial neural networks. Similarly, this is the easiest and reliable method for analysing the process. The ultimate factor of this method is genomic and proteomic applications. This is useful for the scientists to do the experiments that result in a vast amount of data.
6. **Simulation and modelling:** The term simulation refers to computation, especially for advanced algorithms and softwares, which are used for curation and analysing the sequence, functions, and structures. Computer simulation is consistent, accommodating, and movable method to support the information. The properties of biomolecules, their interaction like protein–ligand interaction, drug target analysis, enzyme catalysed reactions and protein folding are very well understood with simulation-based methods. Thus, the working of these molecules mimics the actual physiological events. These simulation studies rely on mathematics, physics, biophysics, and chemistry. Like, quantum mechanics and molecular mechanics (QM and MM) and lively mock-ups of proteins are amalgamated with multiple-scale approaches like diffusion models with cellular automata (CA) in brain tumour replication experiments. Another

example of antifungal protein chitosanase/glucanase, soil bacterium *Paenibacillus* sp. shows binding with potential compounds at C-terminus of proteins, which became possible with ligand dockings and free energy estimates (affinity predictions).

Drug development has benefitted a lot from simulation studies. It is well-known that drug discovery and progress require almost 10–15 years, which is not even costly, but also labour intensive. The time and cost input can be significantly reduced with use of computational tools. Drug development is one of the foremost goals of bioinformatics and is popularly known as Computer-Aided Drug Designing (CADD) (Dong & Zheng, 2008; Macalino et al., 2015). It starts with in-silico structure-activity relationships (SAR) studies, which include binding approximation of potential drug molecules on various target sites, which could be: either enzymes, receptors, ion channels, or transporters, inside the cellular membranes. Computers are used in depiction of shapes of reference molecules 2D (2-dimensional) and 3D (3-dimensional) structure, followed by evaluation of their active pharmacophores (important groups involved in binding) and strength of these interactions at target sites, termed 'molecular docking' (Looger et al., 2003).

Different computational softwares used for docking are AutoDock 4, FLIPDock, Vina, SwissDock, UCSF DOCK, FRED, EADock, SWISS MODEL, LOMETS, PatchDock, HADDOCK 2.2, FIND SITE and ClusPro, etc. (Grosdidier et al., 2011; Pagadala et al., 2017). Some of these are paid; however, AutoDock 4 and FRED are freely available. These SAR studies generate large amount of data and can be useful in predicting a lead compound, which has the potential to develop as a drug in future. The success stories include—Sotalol (brand name Betapace) is used to treat a type of fast heartbeat called, sustained ventricular tachycardia. It slows the heartbeat by acting on potassium channels (Brugada et al., 1990). Similarly, Amlodipine (brand name Norvasc) is used for high blood pressure and coronary artery diseases, by inhibiting calcium ion influx across cell membranes (Calcium channel blocker) (Fares et al., 2016). Similarly, Daliresp® (Roflumilast) is a phosphodiesterase (PDE) 4 inhibitor and is used for treating chronic bronchitis, psoriasis, and neuroinflammation by reducing inflammation (Dong & Zheng, 2008).

Although this method has its own limitations, as side effects may not be entirely be predicted and thus requires subsequent validation with actual laboratory experiments and clinical trials.

Image processing: It is essential and assists the biologists and scientists to view their research. This method displays every stage of accomplishments practically.

With the various steps/techniques involved in bioinformatics, we can now proceed with application of bioinformatics.

4 Application of Bioinformatics

Genome Applications: It starts with DNA sequencing, genome assembly, annotation of genes and prediction of gene function (based on the similarity to known genes), sequence analysis for comparative exploration, evolutionary studies, etc. Algorithms such as BLAST, Clustal W and FASTA provide clarification in sequence investigation and examination. This has benefitted the proteomics, transcriptomics, and metabolomics studies. Similarly, functional genomics studies involve RNA-sequence alignment and differential expression analysis. Comparative genomics and computational evolutionary biology shed light on major events in evolution and divergence. Besides them primer designing, restriction enzymes map analysis, RNA fold, dot plot are other genome-based applications.

In predicting protein structure: The RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank) provided the first open access digital platform for researchers and is available at 'https://www.rcsb.org/'. It supports retrieval of 3D-structure data of biological molecules, including proteins and involves protein sequence retrieval, followed by virtually establishing the similarities with sequences of known structures, present in the PDB (homology modelling) (Berman et al., 2000).

Biomedical: In the biomedical field, bioinformatics tools have an overpowering effect on the understanding of genome, molecular medicine, personalised medicine, and preventive medicine. Novel information on the molecular mechanism of any ailment makes it easier to efficiently treat and prevent the disease. This makes it easier to investigate genes straightforwardly associated with numerous diseases. For all ailments, alike drug is given to patients, but different people have different genotype, so it is important to consider the variations, even subtle but significant (called single nucleotide polymorphism, SNP) in patients' response to drug. It is not exaggeration that if DNA profile of a patient is analysed, the medication would be as unbeaten and efficient as possible, especially in chemotherapy. Like Tamoxifen, (commercially known as Nolvadex), is used in breast cancer treatment, which works by binding on oestrogen receptor in breast tissue. This drug works when a person is positive for oestrogen receptor (ER +ve) and is ineffective in patients lacking this receptor (ER -ve). Similarly, polymorphisms (different forms) of gene for human leukocyte antigen (HLA) alters the immune response and metabolism of drugs. Many other drugs effectiveness has been found related to the genotype and often studied under genome-wide association studies (GWAS).

This proves that success rate in treatment depends on patient's sensitivity/insensitivity to particular drugs, which is sometime dependent on variable forms of genes. This, jeopardy of failure during therapy can be lessened by performing GWAS tests beforehand.

Human microbiome and metagenomics: The word 'microbiome' refers to the entire population of microbes that live inside any specific ecosystem. The human microbiome refers to all the microorganisms living inside human body, without causing the disease, including intestine/gut and includes various bacteria and fungi. Their gene and environmental interactions affect the human physiology. The variations in terms of dissimilar bacterial populations in different humans create a 'unique microflora', which has been studied in population groups. Like autistic children, harbour significantly fewer types of gut bacteria than healthy children do (e.g. *Prevotella* and *Coprococcus* species) (Kang et al., 2013). *P. copri* is involved in breakdown of protein and carbohydrate foods.

Preventive Medicine and Gene Therapy: Preventive medicine focus is on general health, well-being and simultaneously preventing diseases, disability and death (Gochhait, & Omale, 2018). In some cases, the patient's body becomes a laboratory for drugs trial, as gene responsible for disease is still not known, or it could be a multi-gene disorder. Our genetic makeup, environmental conditions, disease agents, lifestyles and genetic predisposition decide the probability of acquiring a disease in future and many people die every year from preventable diseases. These could be chronic respiratory diseases, cardiovascular disease, diabetes, certain infectious diseases, etc. Here, genetic testing can also be done to screen for mutations for disease-related gene polymorphism (single nucleotide polymorphism or SNP) that cause genetic disorders or are predisposed to certain diseases like certain type of cancers or diabetes.

Similarly, gene-editing tools, can be useful in correcting the genotype and alleviating the disease phenotype. Although method is still under research, optimistically it would soon become a reality.

Forensic Analysis and Bioinformatics: Forensic analysis is largely based on DNA-related data, which is also used for personal identification and relatedness. Genomic tests are extensively used as legal evidences in paternity disputes, cadaver recognition, insurance business frauds, and other crimes. This is the reason that many countries are instituting forensic databanks of frequent/serial lawbreakers and criminals. It is possible due to DNA sequencing, evidence orderliness, and machine learning algorithms to establish a Probabilistic Graphical Model (Bayesian networks) (Bianchi & Liò, 2007). The fingerprints, DNA samples, retinal/iris scan (unique patterns of a person's retina blood vessels) and tongue prints are various methods, which are equally effective like signature verification, voice recognition, and face recognition (Radhika et al., 2016)**.** Their biometric databases and similarity searches are possible due to bioinformatics only.

Microbial Genome and Climate change studies: This refers to exploring the details of the genetic material of microbes and isolating the genes, which give them an unparalleled ability to survive in extreme conditions. This application can have ample implications in the improvement of ecosystem, well-being, energy and industrial benefits. A well-suited example is *Pseudomonas putida*, a bacterium with synthesised genome, where DNA of four different species is combined to develop a petroleum-degrading phenotype during notorious oil-spill in ocean in 1980, which successfully cleaned the hydrocarbons floating in water and putting underwater animals at risk. In 1994, the US Department of Energy initiated the Microbial Genome Project, for the sequencing of microbes useful for environmental cleanup, energy production, industrial treatment, and minimising toxic waste. Similarly, projects that are more ambitious can be planned like global climate change, which is chiefly due to increasing levels of carbon dioxide emissions and one way to reduce atmospheric CO_2 is possible by exploring the genomes of microorganisms that uses CO_2 for carbon.

Biotechnology*:* There are many uses of bioinformatics to fasten research in the field of biology, like automated sequencing of genome, gene mapping, protein configuration, organism identification, drug development and vaccine design. Still, there is more scope as biotechnology field is a side street. Few examples are discussed below-

Crop Improvement: It is agreed that the growth in population led to worldwide temperature changes and resulted in declined crop yields. Thus, a major challenge is that people should not die of starvation and this is feasible with a sustainable agricultural production model. Here, comparative genomics aid to understanding genes functions, across plant species. In order to achieve high-quality crops in a short period, bioinformatics databases are used to design new technologies and varieties with better productivity.

Veterinary Science: Mere sufficient food consumption is not adequate for people to survive and stay fit. Our health is dependent on nutrients up take from food. At times, for consuming nutrients, people today depend on livestock as well. A great success achieved in modifying the animal's genotype through bioinformatics, which reduced the risk of possible infection and increasing production.

Innovation in examining the animal species helps in understanding the system genetics of complex traits and provides accurate prediction. Specifically, focuses was given on sequencing genome of animals including—horse, cow, pigs, and sheep. This led to the development in total production, as well as health of livestock. Moreover, bioinformatics has helped researchers in discovering new tools for the discovery of vaccine targets.

Canine cancer approximately affect one in every three dogs and exist as one of the leading causes of death, despite advances in conventional therapies. The cancer pathophysiology is linked with alteration of cellular gene expression, and bioinformatics toolbox promises to put forward innumerable insights into molecular mechanisms, diagnostics and novel therapeutic interventions for cancer. In dogs, osteosarcoma (OSA) affect the lives of 85–90% of affected, within two years of identification, despite uncompromising surgery and chemotherapy. Here, the death is more frequently due to metastasis and thus studying the transcriptome of tumours in bones, scientists have identified a possible therapeutic target. OSA cells express a protein, called her2/neu, on their cell membrane, and metastatic cells express this protein at much higher levels than tumour primary cells, and thus a target for immunotherapy. Scientists created a vaccine, which improves the immune response with obliteration of her2/neu OSA cells. In a first phase of clinical trial, this vaccine show significant improvement over other treatment options.

Another example is equine metabolic syndrome (EMS), an endocrine disorder in horses is linked with obesity, resistance to insulin. Studies show that this syndrome is not only linked with carbohydrate intake, lack of exercise, but there is also a susceptible genotype in horses.

All these examples suggest the wide and ever-growing applications of bioinformatics in biological science and medicine.

5 Conclusions and Future Prospects

Human evolution came a long way, and credit goes to advances in important fields under biotechnology, as diagnostics, drug inventions, clinical health, and agriculture have heightened our financial and social standards. Although a lot is achieved, but still bioinformatics can further help biotechnology in reaching new heights, and helping humankind, but we must set the ethical boundaries and inventions should happen within moral limits. Thus, besides inventions, a string surveillance and regulatory system is need of the hour. Industry practitioners and academicians have to look forward towards the adaptation of AI (Artificial Intelligence) in gaining momentum through big data analysis, machine learning, social media analysis, algorithm decision-making, simulation modelling, and other techniques that is used for bioinformatics visibility in the global market (Varsha et al., 2021).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine, 79*(2), 137–158.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Bianchi, L., & Liò, P. (2007). Forensic DNA and bioinformatics. *Briefings in Bioinformatics, 8*(2), 117–128. https://doi.org/10.1093/bib/bbm006

Brugada, P., Smeets, J. L., Brugada, J., & Farré, J. (1990). Mechanism of action of sotalol in supraventricular arrhythmias. *Cardiovascular Drugs and Therapy, 4*(Suppl 3), 619–623. https://doi.org/10.1007/BF00357040

Dong, X., & Zheng, W. (2008). A new structure-based QSAR method affords both descriptive and predictive models for phosphodiesterase-4 inhibitors. *Current Chemical Genomics, 2*, 29–39. https://doi.org/10.2174/1875397300802010029

Fares, H., DiNicolantonio, J. J., O'Keefe, J. H., & Lavie, C. J. (2016). Amlodipine in hypertension: A first-line agent with efficacy for improving blood pressure and patient outcomes. *Open Heart, 3*(2), e000473. https://doi.org/10.1136/openhrt-2016-000473

Franklin, R. E., & Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature, 171*(4356), 740–741. https://doi.org/10.1038/171740a0

Gochhait, S., & Omale, D. (2018, June). Analytical solution to the Mathematical models of HIV/AIDS with control in a heterogeneous population using homotopy perturbation method (HPM). *Advances in Modelling and Analysis A, 55*(1). ISSN 1258-5769.

Gochhait, S., et al. (2021). Data interpretation and visualization of COVID-19 cases using R programming. *Informatics in Medicine Unlocked, 26*(6). ISSN: 0146-4116.

Grosdidier, A., Zoete, V., & Michielin, O. (2011). SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Research, 39*(Web Server issue), W270–277. https://doi.org/10.1093/nar/gkr366

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology, 36*(1), 39–56.

Kang, D.-W., Park, J. G., Ilhan, Z. E., Wallstrom, G., LaBaer, J., Adams, J. B., & Krajmalnik- Brown, R. (2013). Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. *PLoS ONE, 8*(7), e68322. https://doi.org/10.1371/journal.pone.0068322

Looger, L. L., Dwyer, M. A., Smith, J. J., & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature, 423*(6936), 185–190. https://doi.org/10.1038/nature01556

Macalino, S. J. Y., Gosu, V., Hong, S., & Choi, S. (2015). Role of computer-aided drug design in modern drug discovery. *Archives of Pharmacal Research, 38*(9), 1686–1701. https://doi.org/10.1007/s12272-015-0640-5

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., Fulton, R., Kucaba, T. A., Wagner- McPherson, C., Barbazuk, W. B., Gregory, S. G., Humphray, S. J., French, L., Evans, R. S., Bethel, G., … Max-Planck-Institute for Molecular Genetics. (2001). A physical map of the human genome. *Nature, 409*(6822), 934–941. https://doi.org/10.1038/35057157

Pagadala, N. S., Syed, K., & Tuszynski, J. (2017). Software for molecular docking: A review. *Biophysical Reviews, 9*(2), 91–102. https://doi.org/10.1007/s12551-016-0247-1

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics, 22*, 256. https://doi.org/10.1186/s12859-021-04180-x

Radhika, T., Jeddy, N., & Nithya, S. (2016). Tongue prints: A novel biometric and potential forensic tool. *Journal of Forensic Dental Sciences, 8*(3), 117–119. https://doi.org/10.4103/0975-1475.195119

Sharma, A., Ghosh, D., Divekar, N., Gore, M., Gochhait, S., & Shireshi, S. (2021). Comparing the socio-economic implications of the 1918 Spanish flu and the COVID-19 pandemic in India: A systematic review of literature. *International Social Science Journal, 71*, 23–36. https://doi.org/10.1111/issj.12266

Varsha, P. S., Akter, S., Kumar, A., Gochhait, S., & Patagundi, B. (2021). The impact of artificial intelligence on branding: A bibliometric analysis (1982–2019). *Journal of Global Information Management (JGIM), 29*(4), 221–246. https://doi.org/10.4018/JGIM.20210701.oa10

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., & Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature, 171*(4356), 737–738. https://doi.org/10.1038/171737a0