Introduction to the Concepts of Agr-Informatics

Sidharth Singh, M. Aafikul Haque, and Om Silakari

Introduction

What is Bioinformatics? Defining a term like Bioinformatics is a tedious task especially because it essentially has multiple meanings covering a wide range of topics which includes but not limited to DNA data storage, mathematical modeling, understanding the mechanism behind complicated human diseases, etc. In basic terms bioinformatics can be defined as an interdisciplinary research area which amalgamates biological and computational science [1]. How? In the modern world, computers are being used by biologists equally as compared to the individuals of any other profession; for example, Bankers or pilots. Apart from sending regular emails, filling up spreadsheets, listening to music, biologists are trained to do certain specific tasks like storing complex biological information into various databases, developing algorithms to retrieve meaningful data from these databases, designing mathematical models to predict the outcome, etc. [2]. In our personal understanding we can say that Bioinformatics is the science which deals with the computational management of all kinds of molecular biological information. Now the question arises what is molecular biological information and how computational science is managing it? To understand this, we have to look at a wider picture on the biological work which is being done in the world. At present most labs are generating data which is related to the mechanism of a particular disease, new finding in already established concepts, developing new drugs, new vaccines, etc. So, we need a place

S. Singh (✉)
Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, India

M. A. Haque
Department of Pharmaceutical Analysis, Anurag University, Hyderabad, India

O. Silakari
Department of Pharmaceutical Science and Drug Research, Punjabi University, Patiala, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
A. K. Upadhyay et al. (eds.), Bioinformatics for agriculture: High-throughput approaches, https://doi.org/10.1007/978-981-33-4791-5_1
where all this data can be stored and retrieved when needed by anyone throughout the world. This aspect is taken care of by computational science.

At times the line which can distinguish biology from computational science is very blurry, hence many people use Bioinformatics and computational biology terms invariably [3]. In our view Bioinformatics and computational biology are interdisciplinary fields which involve researchers having expertise in different fields; for example, computer science, molecular biology, genetics, mathematics, statistics, physics, etc. Aim of these two fields can be defined as follows:

a) Bioinformatics: It is concerned with collecting and storing the biological information. Anything related to biological databases are included in this field.

b) Computational biology: It is concerned with the development of computational programs/algorithms and various statistical models needed to understand the complex biological data.

We can further understand these two fields according to the definition given by NIH.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, archive, analyze, or visualize such data [1].

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems [1].

So why has Bioinformatics become the new buzz word in the market?

Human genome project was a 13-year-long international, collaborative program which aims to determine all the genes that make up the human genome [4]. Once it was completed almost half of the genes identified have no known function. Bioinformatics has played a key role in identifying, establishing their functions and further understanding their role in diagnosing, preventing, and treating diseases. Bioinformatics can guide us to uncover hidden information in our DNA and moreover it can help companies to save money and time [5]. In addition to this a lot of data is being generated due to advances in biotechnology techniques which need to be managed. The old rule of supply and demand is also favoring the rise of bioinformatics. There are not many people who are adequately trained in both fields, i.e. biology and computer science. So, we need a biologist who can work with computers or vice versa to solve the modern day biotechnology problems. We can conclude that there are four main reasons which have played a significant role in rise of bioinformatics:

a) First, is the never-ending collection of DNA and protein sequences which we are generating at a pace faster than ever, this data needs to be managed so that it can be used to solve the problems which earlier were impossible to solve.

b) Second, to generate a meaningful conclusion from the raw data (DNA and protein sequences) computer algorithms, programs as well as number crunching power of computers are needed.
c) Third, is the availability of high-powered computers to biologist which till World War II were not available.
d) Fourth, is the idea that macromolecules are the source of a variety of information becomes the core of scientific discoveries and is being considered as the connecting link between computational biology and molecular biology which leads to acceptance of computer science by biologists.

We have now understood why there is a buzz for Bioinformatics in the market; but another question remains unanswered, what the aims of Bioinformatics are? Or what we can achieve with the use of Bioinformatics?

a) First and foremost is the organization of existing data so that researchers can easily access the information, they must also be able to submit new entries as and when they are generated, e.g. DDBJ for DNA related information.
b) All the information which is stored in various databases are essentially useless until they can be analyzed. So, another goal of bioinformatics is to develop tool and resources to analyze data [6].
c) Third aim is to successfully implement these tools to analyze data and draw logical conclusion [7].

Some Basic Biological Science

In order to understand, implement, and improve the existing bioinformatics tools and techniques for smooth analysis of raw data, one must first be well acquainted with basics of biological science especially molecular biology.

Scale and Time

Biology can be defined as the science of life and all living organism present on this earth. An organism is a living entity composed of a single cell, e.g. bacteria, or multiple cells, e.g. animals, plants, etc. Multicellular organisms are visible to human eyes but single cell organism cannot be visualized by naked human eye owing to their very small size, ranging between 1 μm to 100 nm (some viruses) [8]. All the life forms are in essence made up of small molecules having a size range of about 1 nm. Because it is difficult to visualize anything at this small scale so scientists have to design novel visualization techniques to see these molecules. These techniques have generated huge amount of data due to which scientists were able to understand the complex nature of various human life processes.

Life on earth began long back almost 4 billion years ago, not very long after earth comes in to existence. Since then, all life forms have seen evolution happening during the course of their life. If one has to compress earth’s evolution in terms of time, let us say in a month, then we can say that origin of life happened around first 3–4 days but majority of the life forms came into existence only after the 27th day
A lot of multicellular organism appeared in the last few days, i.e. land plants and land animals came in to existence on the 28th day, mammals came in to existence on 29th day, and on the last day all the birds and flowering plants appeared on earth. Modern day’s humans or homo sapiens came in to existence in the last 10 mins of the last day. The process of gradual change in the genetic composition of a living organism and becoming more complex to adapt to new changing environment is known as evolution [10]. The process of evolution often ends with the development of a new species. While studying evolution, it is important to realize that the tree of evolution is composed of various branches or leaves each depicting a species which has evolved from the previous one. So, in order to completely understand evolution of a particular specie one must be able to compare related species.

Cell

Cell can be defined as the smallest structural and functional unit of an organism. Some organisms are single cellular, i.e. they have only one cell, whereas some other are multicellular having billions of cells, e.g. plants and animals.

Broadly cells can be divided in to two types: Prokaryotic and eukaryotic cells. The major difference between prokaryotes and eukaryotes is the absence of nucleus in the former and presence in the later [11]. Living organism are also divided in to prokaryotes and eukaryotes based on the presence of prokaryotic cell and eukaryotic cells. The earliest form of life on earth mainly composed of prokaryotes which include single cell organism like bacteria and archaea. Eukaryotes mainly composed of higher animals like plants and animals and certain unicellular organism like yeast. E. coli, a bacterium, is the most extensively studied prokaryote having a very simple life process; on the other hand, eukaryotic cells consist of more complex structures commonly known as organelles [11]. The nucleus of the eukaryotic cells contains the genetic material DNA which is coiled or compressed in to chromatin or chromosomes. When a cell is not participating or undergoing cell division the proteins and DNA are aggregated to form chromatin which is scattered all over the nucleus [12]. When cell starts dividing these chromatin molecules further gets packed in to a structure commonly known as chromosomes. It has two arms the P-arm (shorter arm) and Q arm (longer arm) joined to each other with the help of centromere.

DNA and Chromosome

DNA stands for deoxyribonucleic acid; it stores all the genetic information of a cell. Some organism has RNA as the genetic material like some virus (Coronavirus). DNA and RNA are made up of nucleotides. Nucleotides itself are composed of bases (adenine, thymine, guanine, and cytosine), a molecule of pentose sugar and phosphoric acid [13]. Nucleosides are composed of a molecule of pentose sugar and a nitrogenous base as shown in Fig. 1. The sugar molecule present in DNA molecule is
de-oxy ribose sugar, while ribose sugar is present in an RNA molecule. Bases can be divided into two groups: purine and pyrimidine—adenine (A) and guanine (G) are purines with two fused rings, while thymidine (T) and cytosine (C) having single ring belong to the category of pyrimidine [13].

Thymidine (T) is replaced by a new base Uracil (U) in a RNA molecule. DNA is composed of two complementary strands which run in opposite direction, i.e. one is running in 5' to 3' direction and another one in 3' to 5' direction. Pentose sugar and phosphate group act as backbone of the entire DNA/RNA molecule.

The famous double helix of DNA is made up of two complementary strands having hydrogen bonds between them (A and T have two hydrogen bonds, while G and C have three hydrogen bonds). This bonding is named as base pairing. Generally, RNA is present in nature as single stranded molecule but occasionally it pairs with DNA, the base pairing rule for this situation is A-U, T-A, G-C, and C-G [14].

The pentose sugar present in the DNA/RNA molecule has five carbon atoms numbered as 1'-5'. DNA or RNA molecule generally referred to as running in a 5' to 3' direction or 5' end/3' end, this naming convention is also based on pentose sugar numbering system. The DNA sequence is always read in 5' to 3' direction.

5'-ATTACGGTACCGT-3'
3'-TAATGCCATGGCA-5'

DNA is a very complex molecule and it is present inside the nucleus in a highly compact form. It binds with histones to form nucleosomes that appear as beads on a DNA string. This nucleosome gets compressed by coiling again and again to form supercoiled chromatin fibers. This coiled structure further folds to form loops and on
further coiling chromosomes are formed. The entire length of a DNA molecule inside a single cell is approximately 2 m but due to this supercoiling phenomenon, the entire DNA molecule can fit inside a nucleus of diameter approximately 5 μm [15]

Central Dogma

The central dogma in biology is described as the mechanism through which the encoded information in DNA is passed on to messenger RNA (mRNA), and then further used to direct the synthesis of proteins. The former process is known as transcription (DNA→mRNA), while the latter is known as translation (mRNA→proteins). Proteins are made up of small chains of amino acids joined together by peptide bond [16]. There are 20 different types of standard amino acids. The synthesis of proteins is based on the mRNA sequence which is tightly regulated by a universal genetic code as depicted in Fig. 2. A genetic code is a three-letter code made up of nucleotides, is specific for each amino acid, and is referred to as triplet. Since 64 unique amino acids can be coded by three nucleotides the presence of redundancies is inevitable [18]. One codon can code for more than one amino acid.

| U       | C       | A       | G       |
|---------|---------|---------|---------|
| UUU     | UCU     | UAU     | UGU     |
| UUC     | UCC     | UAC     | UGC     |
| UUA     | UCA     | UAA     | UGA     |
| UUG     | UCG     | UAG     | UGG     |
| C       | CCU     | CAU     | CGU     |
| CUC     | CCC     | CAC     | CGC     |
| CUA     | CCA     | CAA     | CGA     |
| CUG     | CGG     | CAG     | CCC     |
| A       | ACU     | AAC     | AGU     |
| AUC     | ACC     | AAG     | AGC     |
| AUA     | ACA     | AAA     | AGA     |
| AUG     | ACA     | Lysine  | Arginine|
|         | ACG     | AAG     | AAG     |
| G       | GCC     | GAA     | GGU     |
| GUU     | GCA     | GAG     | GGC     |
| GUC     | GCG     | Glutamic acid | Glycine |
| GUA     | GGC     | Glutamic acid | GGA     |
| GUG     | GCG     | Glutamic acid | GGG     |

Fig. 2 Standard genetic code
All the redundant codons have first two identical nucleotides differing only in the last nucleotide. The start codon for any protein molecule is AUG which codes for methionine, while there are three different stop codons, viz, CAA, CAG, UGA.

The process of protein synthesis in prokaryotes is different from the process in eukaryotes. In prokaryotes, after the unwinding of DNA, one strand is used as a template to generate mRNA which is simultaneously used to synthesize proteins with the help of tRNA [17]. In eukaryotes, first half takes place inside the nucleus where DNA resides and mRNA is synthesized. Synthesized mRNA is referred to as pre-mRNA. The second half takes places in the cytoplasm so pre-mRNA needs to undergo certain modification like a poly A tails is added to protect it from the cytoplasmic enzymes and certain parts are removed referred to as splicing [18]. After entering the cytoplasm the process of translation begins and with the help of tRNA protein molecules are synthesized.

Till here we have talked about the basic biological science which one must know before diving in to the ocean of bioinformatics as this knowledge helps one to understand when, what, and how to apply bioinformatics to get desired results. From here onwards we will be discussing various bioinformatics tools, databases, online resources which are frequently used by a modern day bioinformatician.

### Databases

Due to the advent of new technologies, huge amount of raw sequence data is being generated nowadays and as the volume of this data grows, modern day bioinformatician needs to develop more sophisticated computational tools to manage this data. Hence management of this huge volume of information becomes imperative, so there is need to constantly improve and develop new, advanced computer databases. A Biological database is no different from a conventional computer database except it stores complex biological data, e.g. DNA/protein sequences [19]. Like all databases it is organized, managed by computational algorithms which help in constant submission, retrieval, and updating of the database. In simple terms a database can be defined as an organized collection of raw data, generally stored and accessed with the help of a computer to generate meaningful results. Example: DDBJ (DNA Data Bank of Japan) which collects DNA sequences, PDB (Protein Data Bank) which stores structures of proteins.

### Types of Biological Databases

Biological databases can be broadly divided in two categories: On the basis of source and on the basis of nature of data. On the basis of source there are two types of databases: primary databases and secondary databases [20].

On the basis of nature of data, databases can be broadly divided in to 5 sub categories: Sequence database, structure database, signal transduction pathway database, gene expression database, and metabolic pathway database as shown in Fig. 3.

**Primary Databases**

Primary databases contain experimentally generated raw data such as DNA and protein sequences. Scientist working all over the world directly submits their experimental data after which an accession number is given to each entry so that it can be easily retrieved wherever necessary [21].

Example: DDBJ and Genbank for genome sequences, Swiss-Prot and PIR for protein sequences, Protein Data Bank for three-dimensional protein structures.

**Secondary Databases**

Secondary databases store information which is derived from the primary data. They generally store information derived from various resources like scientific literature and other databases [21]. They are highly curated and one can find information related to conserved sequences, active site residue, and signature sequences in these databases. Examples: SCOP, CATH, PROSITE.
Special Databases

Apart from all these databases there are certain special databases which cater to a specialized research interest, e.g. OMIM inherited diseases database (Online Mendelian Inheritance in Man), Gene expression omnibus—Microarray database, Array expression database—Microarray database, Whole genome database—ENSEMBL.

Sequence Alignment

What is a sequence alignment? What is the purpose of aligning two sequences? How to do sequence alignment? These all are the question which comes to mind as soon as someone talks about sequence alignment. In simple terms sequence alignment is a technique/process which tells us that how much similarity is present between two or more sequences (the sequence can be of DNA, RNA, or proteins), is there any evolutionary relationship between the sequences? [23]. In an alignment of two sequences (protein or nucleotide), certain parts of the sequence will be exactly matching, we call them as “match,” while some other parts will not match, we call them as “mismatch” [22]. Let us understand sequence alignment with the help of an example:

Consider two sequences \texttt{abcdef} and \texttt{abdgf}, we have to align them. Write second sequence below the first one

\begin{verbatim}
abcdef
abdgf
\end{verbatim}

Now move sequence in order to generate a match between two sequences.

\begin{verbatim}
abcdef
   |
   |
abdgf
\end{verbatim}

The characters that are matching are marked with vertical lines. In order to maximize the alignment, we inserted a gap between \texttt{b} and \texttt{d} in the lower sequence.
In this alignment e and g do not match. So, the goal of aligning two sequences is

- To maximize base to base matches.
- If required insert gaps in either of the sequence so that the overall alignment can be made.
- The order of bases in each sequence must remain preserved.
- Gap to gap match is not considered.

Now how to evaluate if the alignment generated by the above method is a good alignment? For this we need some sort of scoring scheme. A scoring scheme consists of score for each possible replacement of bases (positive score for a match and negative score for a mismatch) and penalties are added when gaps are encountered. An overall alignment score is generated by adding all the substitution score and gap penalties. The higher the score, better the alignment. In the above example if we assign the following scoring pattern: Match = +1, Mismatch = −1, and Gap = 0, then we will get final score as 3 (1 + 1 + 0 + 1 − 1 + 1).

Let us consider another example to calculate which the best alignment is.

| Query seq: | ATGGCG |
| Seq1:      | ATGAG |

Now we can align Seq1 in two below mentioned ways:

| Query:          | ATGGCG |
|-----------------|--------|
| Alignment 1: ATG_AG | Score +1 + 1 + 0 − 1 + 1 = 3 |
| Alignment 2: A_TGAG | Score +1 + 0 − 1 + 1 − 1 + 1 = 1 |

As discussed earlier the higher the score, the better is the alignment. Hence, we consider alignment 1 as the best alignment.

What Is the Need of Sequence Alignment?

If two protein or nucleotide sequences are aligned, then the degree of similarity between two sequences can tell us how closely related they are; in other words do they have a common ancestor or not? Since the secondary, tertiary, and quaternary structures of proteins are dependent on the sequence of protein so if two proteins share similarity in the sequence, then we can say, they share similarity in the structure and function. Hence the purpose of doing sequence alignment is to
determine the functionality of unknown protein or to find conserved region within a nucleotide sequence [23].

**Alignment Methods**

The sequences which are very similar or very short can be aligned manually. However, in most cases the alignment is done between very lengthy and immensely variable sequences which are very difficult to align manually. Hence a number of algorithms and computer programs have been designed to make our work easier. With the help of these algorithms very complex, highly variable, and immensely lengthy sequences can be aligned in a matter of few minutes. These tools also help us to interpret the results by strategically presenting the patterns in the results which are otherwise difficult to show algorithmically [24]. Computational methodologies used for sequence alignment can be divided into two types: pair wise sequence alignment and multiple sequence alignment. The major difference between these two is the former is used only when an alignment needs to be generated between two sequences, whereas the latter is used to identify similar region (which may indicate functional, structural, and evolutionary connection) among three or more sequences [25]. Further, pair wise sequence alignment is divided into local and global alignment (Fig. 4). In global alignment an end to end alignment of two sequences is created to check the similarity between two sequences, while in local alignment the algorithm looks for one or more small areas that show similarity within the two sequences. A number of computational programs have been used for aligning sequences. These include dynamic programming which is slow, and probabilistic method which is used for large-scale database search but it does not guarantee to give the best results. For generating global alignment Needleman–Wunsch algorithm [26] is used, whereas Smith–Waterman algorithm [27] is used to generate local alignment.

**Fig. 4** Schematic representation of global and local alignment
**Online Resources to Perform Global and Local Alignment**

BLAST (Basic Local Alignment Search Tool) and FASTA are the two most commonly used online tools to find local alignment. BLAST tool is developed by NCBI (National Centre for Biotechnology Information) situated in Bethesda, Maryland, USA [28], whereas FASTA was developed by David J. Lipman and William R. Pearson in 1985. FASTA is the first algorithm which is used to find similar sequences from databases. The algorithm finds optimal local alignments by examining the sequence for small matches referred to as “words.” Initially, the scores of segments in which there are multiple word hits are calculated (“init1”). Later the scores of several segments may be summed to generate an “initn” score. An optimized alignment that includes gaps is shown in the output as “opt.” The sensitivity and speed of the search are inversely related and controlled by the “k-tup” variable that specifies the size of a “word” [27].

BLAST is a sequence comparison algorithm which is optimized for increased speed to produce best local alignment for a query sequence. The search starts by searching for word of length “W” that scores at least “T” when compared to the query sequence utilizing a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of “S.” The “T” parameter dictates the speed and sensitivity of the search.

The input required for BLAST is FASTA or Genbank format. Any sequence can be converted to a FASTA format just by adding a “>” symbol in front of it. A variety of BLAST tool are developed, e.g. blastp or protein blast (search protein databases using a protein query), blastn or nucleotide blast (search nucleotide databases using a nucleotide query), blastx (search protein databases using a translated nucleotide query), tblastn (search translated nucleotide databases using a protein query), tblastx (search translated nucleotide databases using a translated nucleotide query). Some other formats for BLAST are also developed like Smart-BLAST, Primer-BLAST, IgBLAST, CDART, MOLE-BLAST, MEGA-BLAST.

The output can be retrieved in various formats, e.g. HTML, plain text, and XML formatting. The default output from NCBI page is HTML. The results are presented in a number of ways like graphical output which shows the hits obtained, a tabular format shows various hits encountered and related data such as E value, percentage query coverage, percentage similarity. The tabular output is the most convenient to understand. Table 1 is showing a list of various online resources for local alignment.

| Name of site                                    | Web address                                                   | References                          |
|-------------------------------------------------|--------------------------------------------------------------|-------------------------------------|
| FASTA program suite                             | https://fasta.bioch.virginia.edu/                           | Pearson and Miller [27]             |
|                                                 | fasta_web2/fasta_down.shtml                                  |                                     |
| SIM—Local similarity program for finding alternative alignments | https://web.expasy.org/sim/sim_notes.html                  | Huang et al. [29] and Huang and Miller [30] |
| BLAST 2 sequence alignment (BLASTN, BLASTX)     | https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins       | Altschul et al. [28]               |
**Substitution Matrices**

In Bioinformatics, substitution matrices are defined as the rate at which one amino acid is replaced by another amino acid in a protein sequence during the course of evolution. In other words, it gives the probability of one amino acid being replaced by another one during evolution [31].

**What Is the Need of Creating Substitution Matrices?**

Let us understand this with the help of an example. Suppose we have two sequences:

| Seq 1       | ATGACTTGA |
|-------------|-----------|
| Seq 2       | ATATGA    |

If we follow the scoring pattern of
- Match→1
- Mismatch→0

the two possible alignments with alignment scores are shown below:

| Alignment 1                | Alignment 2               |
|----------------------------|---------------------------|
| ATGACTTGA                  | ATGACTTGA                 |
| ATA _ _TG_ A Score 5       | AT _ A_T_ GA Score 6      |

So, which one is the best alignment?

The best alignment is the one that best represents the matches among different characters and has maximum alignment score. Hence in this case we can say alignment 2 is the best alignment.

All amino acids are different in chemical nature except the one which belong to the same category like all the charged amino acids have same chemical properties. So, during the course of evolution the probability of substitution of one charged amino acid with an uncharged amino acid is far less than the probability of substitution of a charged amino acid with another charged amino acid. If the former situation occurs, then the simple scoring scheme of 1 for a match and 0 for a mismatch is not enough. So, in order to identify what the probability of substitution of one amino acid is by another amino acid substitution matrices were developed.

**Types of Substitution Matrices**

There are two basic types of substitution matrices:
1. Point Accepted Mutation Matrices (PAM): It was developed by Margaret Dayhoff in the 1972s. The basis of PAM construction is to find the probability of mutation of one amino acid by another amino acid during the course of evolution. PAM1 matrix gives us Mutation probability of two amino acid for the evolutionary distance of 1 PAM (i.e., one Accepted Point Mutation per 100 amino acids) [32]. PAM1: A PAM unit is a time period over which 1% of amino acids in a sequence is expected to undergo accepted mutations some of which may occur in the same position. Relation between PAM and BLOSSUM matrix is shown in Fig. 5

2. Blocks Substitution Matrices (BLOSUM): The matrices were created by merging (clustering) all sequences that were more similar than a given percentage into one single sequence and then comparing those sequences (that were all more divergent than the given percentage value) only; thus, reducing the contribution of closely related sequences. The percentage used was appended to the name, giving BLOSUM80, for example, where sequences that were more than 80% identical were clustered. BLOSUM r: the matrix built from blocks with less than r% of similarity—E.g., BLOSUM62 is the matrix built using sequences with less than 62% similarity (sequences with ≥62% identity were clustered) [33].

**Differences Between PAM and BLOSUM Matrices**

PAM matrices are generally used to score alignments between closely related sequences as compared to BLOSUM which is used to score alignments between evolutionary divergent sequences. PAM matrices are based on global alignment as compared to local alignment on which BLOSUM matrices are based. Figure 5 depicts the relationship between PAM and BLOSUM matrices. Higher PAM matrices like PAM 250 serve the same purpose as BLOSUM 30 which essentially means that these matrices are inversely related. Higher PAM matrices are equivalent to lower BLOSUM matrices [35].
Multiple Sequence Alignment

Due to the advances in modern microbiological and analytical techniques, it is now evident that DNA sequences of various organisms are related and show some level of similarity in their genomes. Some of the widely divergent species have been reported to have conserved sequences of similar genes, often times showing exactly similar functionality, and at some other times they mutate or rearrange themselves to show a completely different function. Hence a lot of genes are present in conserved form in many organisms. A sequence alignment of these genes can reveal the parts which have undergone mutation.

Multiple sequence alignment (MSA) has the potential to reveal the functional and structural similarity of proteins and nucleic acid sequences [34]. The output of MSA reveals homology and the evolutionary relationship between biological sequences. MSA can also be used to identify conserved sequences of protein domains, secondary and tertiary structures of proteins and in some case single amino acid/nucleotides. Computationally, MSA encounters several challenges, first is to find a good alignment for more than two sequences which includes matches, mismatches, and indels and the second is to consider the variation amount in all the sequences for which an alignment needs to be created. A list of few tools used in MSA is shown in Table 2.

A variety of alignment methods comes under the umbrella of MSA to get maximum score without sacrificing on the correctness of alignments.

1. Progressive global alignment: This method was developed by Da-Fei Feng and Doolittle in 1987. This method starts with the two most similar sequences and performing pairwise alignment on them and then it progresses to next one till it reaches the most distantly related. Hence it is known as progressive alignment as it builds an alignment between two most alike sequences and then progressively building the alignment by adding more sequences [36].

2. Iterative method: This method has the same working algorithm as progressive method but after making its first alignment of group of sequences it realigns the results to achieve a more accurate result [35].

3. Alignment which takes into account the locally conserved patterns found in the same order in the sequences.

4. Statistical and probabilistic methods: These methods can assign probability of occurrence to all feasible combination including matches, mismatches, and indels in order to find out the most appropriate MSA.

5. Consensus method: These methods work to give best MSA from multiple different alignments of same group of sequences, e.g. M-COFFEE [37].
Phylogenetic Tree

In 1866, Haeckel has coined the term Phylogeny which essentially refers to the evolutionary development of any organism (plant and animal species) or the origin and evolution of a group of plants or animals [49]. The main benefit of phylogenetic tree is it helps in understanding the developmental history (origin of a particular species), assist in understanding epidemiology of infectious diseases. Phylogenetic analysis gives a phylogenetic tree which depicts the relationship among a group of sequences or different species in hierarchical fashion [50].

Phylogenetic trees can be divided in to two broad categories: rooted and unrooted tree. A rooted tree refers to the one which is having a root or a common ancestor or from where all the sequences or species originated; on the other hand, it is very difficult to understand what the hierarchical pattern is in an unrooted tree of species or sequences as shown in Fig. 6.
In a phylogenetic tree every end point is referred to as a node which represents a sequence or specie. So in Figure “a” there are 5 nodes while it is not possible to determine number of nodes as it is not rooted. The point from where a species evolved into a different one is known as branch point. The branch length in a phylogenetic tree represents how closely one specie is related to its ancestor.

Rooted tree can be of following types:

- Cladogram: Branch length has no meaning.
- Phylogram: Branch length represents evolutionary change.
- Ultra-metric: Branch length represents time and length from root to the leaves are the same.

**How to Construct a Phylogenetic Tree?**

There are various methods to generate a phylogenetic tree. Few of them are listed below:

a) UPGMA (Unweighted pair group method with arithmetic mean) [50].
b) Neighbor joining [51].
c) Neighbor relation [52].
d) Maximum likelihood approach [52].
e) Transformed distance method [52].

**UPGMA Method**

Consider 6 sequences for which a phylogenetic tree needs to be constructed.

A: ATCGTGTTGTACTG
B: CCGGAAGACTAG
C: AACGTGCTACTG
D: ATGGTGAAAGTG
E: CCGGAAAACCTTG
F: TGGCCCTGTATC

---

*Fig. 6* A rooted tree and an unrooted tree
Step 1: Create a distance matrix simply by checking how much two sequences differ from each other. Like Sequence A and B have 9 differences A and C have 2 difference. The whole matrix is filled following the same process.

|     | A   | B   | C   | D   | E   | F   |
|-----|-----|-----|-----|-----|-----|-----|
| A   |   9 |  2  |  4  |  9  | 10  |    |
| B   |  9  |   6 |  6  |  2  | 10  |    |
| C   |  5  |   9 |  9  |  7  | 10  |    |
| D   |  6  |   10|    |  10 |    |    |
| E   |  10 |    |    |  10 |    |    |

Step 2: Identify the sequences with fewest difference between them, in this case A–C and B–E are the two pairs which have minimum distance between them.

Step 3: Draw the grouping in the tree; since A–C and B–E have minimum difference between them so they are closely related and hence will be grouped as one.

```
    A
   / \  
  C   B
  / \ 
 E  D
```

Now, the distance matrix will look like this

|     | A/C | B   | D   | E   | F   |
|-----|-----|-----|-----|-----|-----|
| A/C |  9  | 4.5 |  9  | 10  |    |
| B   |  6  |  2  |  6  | 10  |    |
| D   |  6  |  2  |    | 10  |    |
| E   |  10 |    |    |    |    |
| F   |    |    |    |    |    |

The values in the column B and D will be calculated by taking the average of values for A–B, C–B and A–D, C–D, respectively, from the original table.

Step 4: Complete the table with B-E grouped together following the same process.

|     | A/C | B/E | D   | F   |
|-----|-----|-----|-----|-----|
| A/C |  9  | 4.5 |  6  | 10  |
| B/E |  6  |  6  |    | 10  |
| D   |    |    |    | 10  |
| F   |    |    |    |    |
Step 5: Repeat steps 2 to 4 until the complete tree is made. The final phylogenetic tree will look like this:

References

1. Huerta M, Downing G, Haseltine F, Seto B, Liu Y. NIH working definition of bioinformatics and computational biology: US National Institute of Health; 2000.
2. Fenstermacher D. Introduction to bioinformatics. J Am Soc Inf Sci Technol. 2005;56(5):440–6.
3. Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 2001;40(04):346–58.
4. Collins FS, Fink L. The human genome project. Alcohol Health Res World. 1995;19(3):190.
5. Gill SK, Christopher AF, Gupta V, Bansal P. Emerging role of bioinformatics tools and software in evolution of clinical research. Perspect Clin Res. 2016;7(3):115.
6. Mehmood MA, Sehar U, Ahmad N. Use of bioinformatics tools in different spheres of life sciences. J Data Min Genomics Proteomics. 2014;5(2):1.
7. Bettinotti MP, Olsen A, Stroncek D. The use of bioinformatics to identify the genomic structure of the gene that encodes neutrophil antigen NB1, CD177. Clin Immunol. 2002;102(2):138–44.
8. Bell G, Mooers AO. Size and complexity among multicellular organisms. Biol J Linn Soc. 1997;60(3):345–63.
9. Glaessner MF. The first three billion years of life on Earth. J Geogr (Chigaku Zasshi). 1966;75(6):307–15.
10. Ambrose SH. Paleolithic technology and human evolution. Science. 2001;291(5509):1748–53.
11. Vellai T, Vida G. The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. Proc R Soc Lond Ser B: Biol Sci. 1999;266(1428):1571–7.
12. Hartman H. The origin of the eukaryotic cell. Specul Sci Technol. 1984;7(2):77–81.
13. Bansal M. DNA structure: revisiting the Watson–Crick. Curr Sci. 2003;85(11):557.
14. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature. 2003;423(6936):145–50.
15. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, Segal E, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009;458(7236):362–6.
16. Li GW, Xie XS. Central dogma at the single-molecule level in living cells. Nature. 2011;475(7356):308–15.
17. Moldave K. Eukaryotic protein synthesis. Annu Rev Biochem. 1985;54(1):1109–49.
18. Copley SD, Smith E, Morowitz HJ. A mechanism for the association of amino acids with their codons and the origin of the genetic code. Proc Natl Acad Sci. 2005;102(12):4442–7.
19. Birney E, Clamp M. Biological database design and implementation. Brief Bioinform. 2004;5(1):31–8.
20. Zou D, Ma L, Yu J, Zhang Z. Biological databases for human research. Genomics Proteomics Bioinformatics. 2015;13(1):55–63.
21. Chang J, Zhu X. Bioinformatics databases: intellectual property protection strategy. J Intellect Prop Rights. 2010;15(6):447–54.
22. Vingron M, Waterman MS. Sequence alignment and penalty choice: review of concepts, case studies and implications. J Mol Biol. 1994;235(1):1–12.
23. Barton GJ. Sequence alignment for molecular replacement. Acta Crystallographica Sect D: Biol Crystallogr. 2008;64(1):25–32.
24. Notredame C. Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol. 2007;3(8):e123.
25. Lambert C, Campenhout JM, DeBolle X, Depiereux E. Review of common sequence alignment methods: clues to enhance reliability. Curr Genomics. 2003;4(2):131–46.
26. Likic V. The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne; 2008. p. 1–46.
27. Pearson WR, Miller W. Dynamic programming algorithms for biological sequence comparison. In: Methods in enzynology, vol. 210: Academic Press; 1992. p. 575–601.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
29. Huang X, Hardison RC, Miller W. A space-efficient algorithm for local similarities. Bioinformatics. 1990;6(4):373–81.
30. Huang X, Miller W. A time-efficient, linear-space local similarity algorithm. Adv Appl Math. 1991;12(3):337–57.
31. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol. 1991;219(3):555–65.
32. Dayhoff MO. A model of evolutionary change in proteins. In: Atlas of protein sequence and structure, vol. 5; 1972. p. 89–99.
33. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci. 1992;89(22):10915–9.
34. Corpet F. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. 1988;16(22):10881–90.
35. Mount DW. Using iterative methods for global multiple sequence alignment. Cold Spring Harb Protoc. 2009;2009(7), pdb-top44
36. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 1987;25(4):351–60.
37. Collingridge PW, Kelly S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics. 2012;13(1):117.
38. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–80.
39. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. In: Methods in enzynology, vol. 266: Academic Press; 1996. p. 383–402.
40. Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput Chem. 1999;23(3–4):341–64.
41. Morgenstern B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci. 1996;93(22):12098–103.
42. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol. 1996;264(4):823–38.
43. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
44. Grasso C, Lee C. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 2004;20(10):1546–56.
45. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. J Comput Biol. 1998;5(2):211–21.
46. Schuler GD, Altschul SF, Lipman DJ. A workbench for multiple alignment construction and analysis. Proteins: Struct Funct Bioinf. 1991;9(3):180–90.
47. Liu JS, Neuwald AF, Lawrence CE. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. J Am Stat Assoc. 1995;90(432):1156–70.
48. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol. 1994;235(5):1501–31.
49. Dayrat B. The roots of phylogeny: how did Haeckel build his trees? Syst Biol. 2003;52(4):515–27.
50. Borriss R, Rueckert C, Blom J, Bezuidt O, Reva O, Klenk HP. Whole genome sequence comparisons in taxonomy. In: Methods in microbiology, vol. 38: Academic Press; 2011. p. 409–36.
51. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
52. Saitou N, Imanishi T. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree; 1989.