International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

An Overview of Bioinformatics and Its Application: A Ray of Hope Towards Microbial Biotechnology

Rajrupa Ghosh[1], Rahul Deb Bera[2], Subhra Chandra Chandra[3]

[1,2,3]Assistant Professor, Department of Allied and Health Science, Institute of Management Study

Abstract

A new era in the analysis of biological data has been sparked by the revolutionary development in processing speed and memory capacity. With the sequencing of hundreds of microbial and numerous eukaryotic genomes—including a clearer draft of the human genome—better control over microbes is anticipated. The objectives are very high and include the creation of sensible medications and antimicrobial agents, the creation of novel, improved bacterial strains for bioremediation and pollution control, the creation of more effective vaccines that are simple to administer, the creation of protein biomarkers for a variety of bacterial diseases, and a deeper comprehension of the interactions between hosts and bacteria to prevent bacterial infections. Bioinformatics research currently being conducted on the basis of Genomics, Proteomics, cell visualization and application to the development of drugs and antimicrobial agents. The main contributions of bioinformatics research include automated genome sequencing, automated development of integrated genomics and proteomics databases, automated genome comparisons to determine the function of the genome, automated metabolic pathway derivation, automated gene expression analysis to determine regulatory pathways, automated statistical technique development, data mining and clustering techniques to determine protein-protein and protein-DNA interactions, modeling of 3D protein structure and 3D docking between proteins and biochemicals for rational drug design, and analysis of differences between pathogenic and non-pathogenic strains to identify candidate genes for vaccines and anti-microbial agents.

Keywords Microbial genome, human genome, genomics, proteomics, data mining, 3D docking, bioinformatics.

Background

Over the past ten years, significant advancements in computer technology and memory capacity have enabled the modeling of grand challenge issues like extensive internet-based genomic sequencing and massively integrated database management. Researchers worldwide now have access to a tremendous number of genomic and proteomic data thanks to this much increased computational capacity along with the large-scale downsizing of biochemical procedures like PCR, BAC, gel electrophoresis, and microarray chips. Wet-lab experiments cannot produce the same level of new tools and discoveries as the surge in genome and proteome research that has resulted from the availability of data (1-3).

The expectation of mankind to be able to regulate has increased due to the availability of genetic and proteomics data as well as enhanced bioinformatics and biochemical techniques by tampering with the

already-existing microorganisms. Increased agricultural quality and quantity, improved disease diagnosis through the use of protein biomarkers, disease prevention through the use of affordable vaccines and sensible drug design, and the development of methods that enable us to see and comprehend the intricate microbial machinery at the systemic level are just a few of the many benefits(4, 5). Since the first complete microbial genome of Haemophilus influenzae was sequenced in 1995, hundreds of microbial genomes have been sequenced and archived for public research in GenBank ftp://ftp.ncbi.nih.gov/genbank thanks to the coordinated efforts of national laboratories, academic universities, non-profit organizations like TIGR, multiple drug development companies like Celera, federal health agencies like the NIH and DOE in the USA, EMBL and EBI in Europe, and the DNA databank of Japan. The goal of comprehending host-pathogen interaction has increased with the sequencing of the human genome and other pertinent eukaryotic genomes. This will help in the development of more effective vaccines and sensible medications that target the aberrations at the gene and pathway levels that cause pathogenesis (6-8).

Without the availability of bioinformatics techniques, the enormous volume of data produced by genome sequencing programs would be unmanageable and unable to be understood because of a shortage of highly skilled personnel and the prohibitive expense of continuing such an endeavor. Over the past ten years, bioinformatics has quietly taken over the job of affordable data analysis. This has accelerated the rate at which new discoveries, medications, vaccines, and antimicrobial agents are being developed (9-11).

Furthermore, our knowledge of the genome's structure and the process of microorganism reformation has improved thanks to bioinformatics analysis. Understanding the cellular processes in order to treat and regulate microbial cells as factories, as well as analyzing the systemic level behavior of these processes, will be made easier and faster by bioinformatics analysis. Bioinformatics approaches have been created during the past ten years to discover and evaluate different components of cells, including relationships, metabolic and regulatory processes, and gene and protein activity. The next ten years will be dedicated to integrating bioinformatics, wet lab, and cell simulation tools to better understand cellular mechanisms and manipulation. These methods have recently been applied by researchers to the synthesis of recombinant proteins. This decade's semi-automated analysis of cellular function at the systemic level is expected to accelerate this capability (12, 13)

Literature review

Bioinformatics has been applied to microbial biotechnology in a variety of ways over the past ten years: developing genomic and proteomics databases; inferring phenotypes (higher level functions) from genotypes (gene level functions)(14-16); computationally analyzing wet-lab data; genome sequencing; identifying protein coding segments and using genome comparison to determine gene function. Four main studies have been conducted to better understand higher level functions: (i) automating the reconstruction and comparison of metabolic pathways (17-19); (ii) studying the interactions between proteins and DNA to comprehend regulatory pathways (20-22); (iii) modeling the 2D and 3D structures of proteins (23, 24); and (iv) modeling the docking of 3D models of proteins with drugs (25, 26). Knowing the three-dimensional structure of proteins is essential to comprehending the interactions between proteins. A solid understanding of the binding sites in signaling pathways will come from protein-protein and protein-DNA interactions; knowledge of the interactions between proteins and chemical substances has already aided in the development of medication design (27, 28).

In bioinformatics, three methods have been employed: (i) computational search and alignment methods to compare a newly sequenced genome to the set of known genes in order to annotate the structure and

function of genes (29, 30); (ii) mathematical modeling methods such as data mining, statistical analysis, neural networks, genetic algorithms, and graph matching methods to find common patterns, features, and high level functions (31-36); and (iii) an integrated method that combines search methods with mathematical modeling (37-41).

Genome Sequencing

The development of automated sequencing techniques that combine 2D gel electrophoresis, PCR or BAC-based amplification, and automated nucleotide reading has been the primary contribution of bioinformatics to genome sequencing. Other developments include contig assembly, which joins the sequences of smaller fragments (contigs) to form a complete genome sequence, and promoter and protein-coding region prediction. Limited size genome fragments are obtained by amplification techniques based on BAC (Bacterial Artificial Chromosome) or PCR (Polymerase Chain Reaction) (42, 43). Nucleotide reading mistakes, repetitions (extremely small and very similar fragments that fit in two or more portions of a genome), and chimeras (two separate parts of the genome or artifacts caused by contamination that unite end to end providing an artifactual fragment) are among the problems with the available fragment sequences (44-50).

The nucleotide reading error problem is resolved by creating several copies of the fragments, aligning the fragments, and using majority voting at the same nucleotide sites. To prove repeats and chimeras, several experimental copies are required. Before the genomic fragments are finally assembled, chimeras and repetitions are eliminated. A greedy technique is used to combine the fragments based on maximal overlap. The fragments are modeled as a mathematical weighted network, where nodes are fragments and weights of edges reflect the number of overlapping nucleotides. The majority of nodes with the highest (or lowest) scores collapse first in a greedy algorithm. The pieces with greater nucleotide sequence overlap are linked first to form contigs (21, 42-44).

Identification and recognition of genes automatically

Finding the genomes' ORFs (open reading frames), or areas that code for proteins, comes next once the contigs have been linked. There are three methods for identifying ORFs: Three methods are used to identify genes: (1) utilizing Hidden Markov Model (HMM) based approaches like GLIMMER (51-53) and GeneMark; (2) exploring known gene databases like GenBank ftp://ftp.ncbi.nih.gov/genbank ; and (3) using decision tree-based algorithms to identify the start and end codons of the coding regions (54, 55). Multiple probabilistic state machines, each able to recognize an ORF, are developed using HMM-based approaches (56). Every machine uses a state transition with greatest probability to predict the next nucleotide character. It then compares the predicted character to the current character in the sequence. HMM-based methods create several probabilistic state machines, each of which is able to recognize an ORF. Every machine uses a state transition with greatest probability to predict the next nucleotide character, then compares the anticipated character to the current character in the sequence. The likelihood of a state transition is determined through statistical training with known sample sequences. HMM-based programs like GLIMMER have produced results with 95% to 97% accuracy when it comes to microbial genomes (57).

Finding the function of genes: searching and aligning

Annotating the genes with the correct structure and function comes next after determining the ORFs (open

reading frames). Pair-wise gene alignment and well-known sequence search methods have been used to determine the gene's function. The four most widely used algorithms for functional gene annotation are BLAST and its variants, Smith-Waterman alignment and its variations using dynamic programming, indexing-based scheme FASTA and its variations, and BLOCKS, which employs multiple sequence alignment of conserved domains to identify motifs, or patterns that characterize proteins (58-62).

In order to find the largest matching nonrandom segment, BLAST search expands numerous probable seed points (greater than four nucleotides) that match. It does this by using scoring matrices like BLOSUM or PAM (63). When amino acids share biological or biophysical characteristics, their matching values in scoring matrices are positive; when they don't, they have negative match values. The frequency patterns of the amino acids present in conserved domains of protein families have been statistically compared to create substitution matrices like BLOSUM (BLOcks SUbstitution Matrix). A nucleotide matrix that penalizes non-matching places is used to score nucleotide sequences. The BLAST method is quick in its current implementation and has a temporal complexity that is almost linear. The BLAST algorithm, however, sacrifices some accuracy in order to increase computing speed by indexing the sequences in the database using the most likely combinations of nucleotide seeds (64-66).

In order to increase the BLAST algorithm's execution speed, accuracy, and reliance on specified scoring matrices, numerous heuristic changes have been made. There are two main enhancements: (i) using numerous matching iterations to build a position-specific scoring matrix that can be utilized in place of a predetermined biochemistry matrix, and using two or more hits inside a matching region before extending the high scoring segment. A well-liked BLAST implementation that makes advantage of both of these enhancements is called PSI-BLAST (Position Specific Iterative BLAST) (67-69). Utilizing a position-specific matrix enhances the search for weakly similar sequences in evolutionary distant species, while using two hits increases the segment extension's execution efficiency (70, 71). By determining the multiple sequence alignment of the best matching segments and examining the frequency of amino acid substitutions in the matching segments, a position-specific matrix is constructed.

For pairwise gene alignment, dynamic algorithms like Smith-Waterman and other indexing systems perform better. Using dynamic algorithms, gene pairs are aligned incrementally by maximizing the sum of the scores for matching the current amino acid characters (or nucleotide characters) with the best alignment of the previous subsequences. While nucleotide sequences use a nucleotide matrix for scoring that penalizes non-matching places, amino-acid sequence mismatches are punished using scoring matrices like BLOSUM or PAM (65, 72). To demonstrate the insertion and deletion of nucleotides, a gap is added (or amino-acids). Gaps are entered by users as parameters and are not included in a substitution matrix. A scoring penalty is also incurred when there is a gap (73).

A gap also results in a deduction from score. Global and local dynamic programming protein (or gene) alignments are the two main varieties. The amino-acid (or nucleotide) characters are arranged in global alignment to maximize the final score. Local alignment, on the other hand, identifies the segment with the highest score and disregards the segments with lower ratings. Large-scale changes in amino acid composition are best handled by local alignment when comparing amino acid sequences from evolutionary distant taxa. When only a little quantity of random mutation is present, global alignment performs effectively (64, 73-75).

Utilizing methods of multiple sequence alignment, one can determine conserved portions and an evolutionary tree by comparing several homologous genes (genes with similar sequences). Pair-wise alignment between two homologs and the concept of distance between two nucleotide or amino acid

sequences are combined in this technique. The concept of distance can be obtained in two ways: first, as the evolutionary distance between two microorganisms indicated by an evolutionary tree; second, as the edit distance, which is the number of mismatches obtained following pairwise alignment of two sequences. The method, which is implemented as a greedy algorithm, is based on a progressive pair-wise comparison to create intermediate alignments between nearest neighbor's homologs with the least distance (76, 77).
The assigning of user-defined equal weight to indels (gaps), which lessens the significance of a particular amino acid or set of amino acid characters, is a primary cause of error in the aforementioned sequence comparison techniques. Repeat characters are another small issue with the sequences; they cannot be combined with other amino-acid characters and only indicate the structural or functional separation of the component components within a gene. Multiple sequence comparison methods, like BLOCK, are useful for deriving motifs and have been used to find conserved subsequences in gene sequences that are strikingly similar. A protein's motif, which is a collection of distinct subsequences that define it, has been shown to be highly helpful in identifying genes that share the same functions (78-81).
Once the sequences are aligned, the conserved subsequences of the functionally equivalent genes from other organisms are found, leading to the derivation of motif sequences. The fundamental building block of a protein's function, the protein domain is linked to a single, distinct pattern of folding (beta sheets, helices, or their variants) at the structural level. To determine whether areas of several homologous genes are individually homologous to one another, the researchers employed multiple sequence alignment and HMM. These areas are most likely domains (82-84).
Numerous domain-related databases, including PRODOM, Pfam, and SMART, are available at the moment. A database of numerous protein domain or conserved protein region alignments is called Pfam. The alignments show a conserved evolutionary structure that affects the function of the protein. Profile not visible Even in cases of limited homology, Markov models (profile HMMs) constructed from the Pfam alignments can be used to automatically identify a novel protein as a member of an established protein family. Pfam is currently automatically generated using PRODOM database cluster analysis. The methods based on sequence search assume that the best sequence is adequate to annotate the function. In general, this presumption is accurate. However, in many instances, the function cannot be determined by best sequence match because: (1) the function is restricted to a particular region of the protein, such as the hydrophobic region; (2) the function depends on the presence of a certain pattern of amino acids; (3) 3D structure conformation in protein containing multiple domains (85, 86). Occasionally, a little nucleotide mutation will change the matching amino acids, changing the protein's three-dimensional structure. The inability of best match methods to pinpoint every potential role for a multi-domain protein is another drawback. A protein can be multifunctional and have several domains. Since there is no direct relationship between a protein's number of domains and its functionality, the issue is more complicated (87-93).

Docking of 3D structure

Depending on how it interacts with other proteins, a protein can exist in one or many low free energy conformational forms. A protein's stable conformational state exposes certain areas to interactions between proteins and DNA. Protein function can also be inferred by comparing the three-dimensional (3D) structures of known and unknown proteins, since function depends on exposed active regions. Nevertheless, there are few 3D structures obtained using NMR spectroscopy and X-ray crystallography. A different method of matching genes is therefore required. The relationship between gene sequence and three-dimensional structure is generally close. In these situations, function annotation only requires

sequence matching. Even if amino acid sequences do not match, various sequences frequently map to the same three-dimensional structure. In these situations, it is necessary to find corresponding 3D structures and 2D structures, such as alpha helix and beta sheet patterns, to confirm the function of the recently sequenced protein (94-99).

Sequence homology based prediction and ab initio (or de novo) method are the two main methods used to model the three-dimensional structure of a protein. Sequence homology is a technique that predicts the overall 3D structure by using sequence alignment to find the best matching 3D structure for various components, such as side chains, loop portions, and conserved portions from the database. Based solely on the sequence, the ab initio technique predicts the structure using the energy minimization concept. Modern developments in ab initio techniques incorporate biophysical and biological characteristics, such as beta sheet folding and hydrophobic area information, to improve accuracy (100-106).

When two molecules (a receptor and a ligand) attach to one another, their 3D structures are compared to find the best matches by simulating interaction surfaces and minimizing free energy at the domain level. This process is known as docking. Finding the optimal match to suit two surfaces without having too much intersection is necessary while solving the docking issue, which involves modeling surfaces using spheres (or grids). Binding locations and other biochemical data are frequently supplied. In docking, there are three main issues: Docking techniques suffer from three basic drawbacks: (i) conformation may vary during docking for multidomain proteins; (ii) significant computational cost causes large-scale modeling to be very slow; and (iii) over prediction causes a high proportion of false positives (107-110).

Genome comparison (Pairwise)

Pairwise genome comparisons are a logical next step following the identification of gene-functions. The features of paralogous genes—duplicate genes with identical sequences but different functions—can be found by performing a pairwise genome comparison of one genome against the other. Genome comparisons between genomes in pairs have yielded a plethora of information, including orthologous genes that are functionally equivalent but diverged in two genomes due to speciation, various types of gene-groups, which are adjacent genes that are compelled to occur in close proximity because they are involved in some common higher level function, lateral gene-transfer, which is the transfer of genes from a distantly related microorganism, gene-fusion/gene-fission, gene-group duplication, gene-duplication, and difference analysis, which is used to identify genes unique to a group of genomes like pathogens, as well as conserved genes (111-114).

Genomes are treated as an ordered collection of genes and a pair of genomes is modeled as a bipartite graph where each node in one set is connected to homologous nodes (similar genes via pair-wise gene-alignment) in the second set in order to obtain orthologs and sets of gene-groups. The best matched homologs are deduced to be orthologs. A window of neighboring genes is created in both genomes and slid until the next gene in the first genome has no homologous gene in the corresponding neighborhood window in the second genome. This process is used to identify homologous gene-groups. First, two neighboring genes in one genome that are homologous to two neighboring genes in the other genome are identified. Following the identification of a non-matching gene, the matching genes are gathered as one-gene-group (111, 115-117).

This extensive comparative analysis has revealed that: (i) a significant portion of these gene groups are cotranscribed or co-regulated; (ii) there are different kinds of gene groups in a genome; (iii) homologous genes in a gene group do not always have the same order in two microorganisms; (iv) gene groups are

frequently duplicated; (v) all genes in ordered gene groups are embedded in the same pathway, whereas unordered gene groups occur at the intersections of adjacent pathways; (vi) larger genomes share more gene-groups despite not being too closely related evolutionary (vii) Genes involved in cell surface interaction, nutrition transport, and sensor proteins are frequently duplicated, and (viii) horizontal gene transfer and gene fusion are also frequent methods of genome reorganization. Duplication is justified by the necessity for adaptation to various environmental factors and by the utilization of comparable mechanisms by several sensors and transport proteins. Identification of candidates for vaccine production and the creation of antimicrobial medications is greatly aided by the knowledge of genes conserved in pathogens, genes inserted or deleted from pathways analogous to genes in plasmids, and genes peculiar to pathogens. Pairwise genome comparison studies have revealed an intriguing finding: genome restructuring is brought about by a mix of gene fusion, duplication, and insertion/deletion of domains. Nevertheless, because of their computational complexity and the scarcity of domain-level functional data regarding different genes from the wet-labs, domain-level comparative analysis techniques are still in their experimental stages (111, 118, 119).

Reestablishment of Metabolic processes

A new area of bioinformatics research has been sparked by the identification of gene functionality: the automated reconstruction and comparison of pathways in newly sequenced species. The reconstruction of pathways has been the subject of numerous initiatives and strategies. The three main methods are categorized as follows: (i) a global network of enzyme-catalyzed reactions; (ii) a network of gene-groups connected by enzyme-catalyzed reactions embedded in the gene-groups; and (iii) a global modeling of chemical reactions in microbial cells. Using known biochemical pathways and enzymes, the first method matches the product and substrate of chemical reactions catalyzed by enzymes to build a network of reactions. It also uses BLAST based search or pairwise genome comparison of evolutionary close genomes to identify the enzyme function of new genes in a newly sequenced genome. This is a very effective strategy. Nevertheless, it has numerous shortcomings: (i) it cannot determine the precise location in pathways for homologous genes; (ii) it ignores genes that appear in the same pathway because of co-transcription and gene grouping; and (iii) it ignores the response rate (120, 121).

An integrated method for rebuilding metabolic pathways has been developed using the knowledge of gene-groups. There are four steps in this method: Gene-groups that share a promoter can be identified by (i) using ortholog analysis to identify the enzymes and their functions in a newly sequenced genome; (ii) by analyzing the promoter region of the genes; (iii) by deriving the gene groups by pairwise comparison of the newly sequenced genome with multiple genomes; and (iv) by connecting the network of gene-groups using biochemical knowledge of existing pathways and enzymes. With the exception of the leading gene, cotranscribed gene-groups (potential operons) typically have intergenic distances (the distance between the start codons of the next gene and the stop codons of the previous gene) of fewer than 75 nucleotides. The majority of these potential co-transcribed gene groupings are found by computationally comparing the intergenic distance. Nonetheless, the identification of pathways cannot be achieved solely by using the knowledge of co-transcribed gene groups, as (i) these groups may contain genes that are absent due to conservative cutoff threshold estimates; (ii) gene insertion/deletion resulting from genome restructuring may separate multiple adjacent co-transcribed gene-groups within a single pathway; and (iii) some regulating genes that are in close proximity and regulate pathways may not be detected. By combining genes from different pairwise genome comparisons with the recently sequenced genome that belong to the

same gene group, these three issues are diminished. Combining the data from pairwise genome comparison analysis and promoter-based analysis yields the overall gene-groups. Since the gene-groups in a pathway are dispersed across the genome, enzyme databases are used to match the biochemical product and substrates in the reactions that the enzymes contained in the gene-groups catalyze. This allows the gene-groups to connect with one another. This system incorporates several regulatory genes associated with a pathway, decreases the ambiguity of homologous genes, and increases computational efficiency. However, because reaction rate is not included in this system, cell level behavior cannot be modeled (122, 123).

Modeling global biochemical reactions involving products, byproducts, and the impact of cofactors on reaction rate forms the basis of the third strategy. Based on the study state flux distribution in a metabolic network required for target product synthesis, the model represents the network of metabolic reactions as a collection of vector processes known as extreme pathways. This method models the entire network of routes as a matrix, with columns representing particular reactions and rows representing extreme pathways. This method works well for simulating a microbial cell's general metabolic activity. The gene functions that are now available from wet laboratories limit the metabolic pathway approaches used today. An additional concern is that merely identifying metabolic pathways is insufficient without knowledge of reaction rates and the impact of stress response on response times. Although modern methodologies have been developed to simulate the rate of reaction of metabolic pathways, the whole picture remains unverified, mostly because wet lab gene functions are not available (124, 125).

Comparing automated pathways and phenotypic similarities

In order to comprehend the impact of gene insertion and deletion in diverse microbes and to comprehend evolution at the pathway level, the researchers' next line of inquiry is to compare similar pathways. The genes in a route are aligned as follows so that two pathways can be compared. If every protein in the first pathway (or the gene-group within a pathway) has a homologous gene in the other pathway (or the gene-group within the pathway), then the two pathways are entirely matched. When a homologous gene is added or removed, there is a mismatch, and when the matching homologous genes have a low similarity score, there is a gap. Many routes between *H. pylori* and *yeast* have been found to be compared based on this modeling. More significantly, a quantification mechanism for comparing two paths has been discovered (126, 127).

Development of regulatory pathways and mechanisms

Rebuilding metabolic pathways, identifying signaling pathways, and conducting promoter analysis to find transcription factors for protein-DNA interactions have all been advances in the field of genomics and proteomics research. Studying protein-DNA interactions can be done in four main ways: micro-array analysis of gene expressions in response to various cell stressors; (ii) statistical analysis of the promoter regions of orthologous genes (functionally equivalent genes in different organisms identified as best homologs); (iii) worldwide analysis of dimer frequency patterns in the intergenic region of a genome, which is the promoter region between adjacent protein coding regions; and (iv) atomic bond level biochemical modeling to comprehend how a protein will bind to nucleotides. Of the four ways, only the microarray analysis methodology is based on experimental data; the other two rely on sequence analysis and mathematical modeling (128, 129).

Using a two-step process, micro array analysis evaluates how a change in stimuli affects the relative changes in gene expressions for stressed (or stimulated) cells and changes in cellular expression patterns, such as differentiation, cellular cycle, tissue remodeling, sporulation, etc. The process involves mapping all the genes in a genome that has been etched onto a thin glass plate, hybridizing the genes of a healthy cell with etched genes to derive the regular gene expression under equilibrium conditions, and (ii) hybridizing the affected cells with etched genes to determine the gene expression of affected cells under equilibrium conditions. The information about the damaged genes is obtained through a comparative analysis of gene expressions under normal and stimulated (or stressed) conditions. Assuming that auto regulation occurs within a gene-group and that there is no cyclic self-regulation, the observed changes in gene expression can be attributed to the interaction between transcription factors and proteins. The gene-expression data is subjected to either (i) cluster analysis, which identifies significant gene-expression patterns, or (ii) data mining techniques, which are statistical methods that establish correlations between expressed genes and various stress conditions. Using pair-wise genome comparison databases (see http://www.cs.kent.edu/~arvind/intellibio/orthos.html) or the knowledge of cluster of orthologs (COGS), a group of genes in a super family archived at NCBI at NIH that has been derived by multiple genome comparisons, the second method of statistical promoter analysis first identifies the orthologous genes from evolutionary close microorganisms with active pathways. The second stage involves identifying and comparing the upstream regions of two orthologous genes to find statistically conserved patterns. The transcription factors, or regions of promoters involved in enhancing or repressing the gene-expression of the associated gene, for protein-DNA interaction in the promoters of orthologous genes would likewise be very similar if functionally equivalent genes in the very similar pathways of evolutionary close organisms have similar regulation mechanisms. Many transcription factors have been found as a result of this investigation. Plotting the frequency of occurrence of the dimers in the intergenic region over the entire genome by statistical analysis is the third method. It's possible that the more common non-random dimers interact with DNA through interactions with proteins (130-133).

Hydrogen bonds in amino-acid base interactions, Van der Wall forces at contacts, and water-mediated bonds at varying degrees of closeness between two molecules are all taken into account in the biochemical approach to studying protein-DNA interactions at the atomic bond level. Based on the bond analysis and the actual statistical results, it has been determined that the complex and biased interactions between different amino acids and DNA—that is, the preferences of different amino acids for different types of bases—play a significant role in binding, Van der Wall forces provide stabilization, and protein-DNA interactions are fundamental to binding. Some amino acids that prefer guanine include arginine, histidine, serine and lysine (134-136).

As of yet, no researcher has made an attempt to combine the biochemical technique with the other four ways in a hybrid manner. A more complete picture will be obtained through an integrated approach. A multi-transcription factor co-regulated gene may have weaker transcription factors individually and/or correlations with other transcription factors, which presents another challenging issue. A two-step procedure can be used to determine the weak transcription factor: The steps are as follows: (i) use one of the earlier methods to find the strong associated transcription factor; (ii) search for patterns in the vicinity of the strong pattern (137, 138).

Determining the signaling pathway by analyzing the connectivity in protein-protein interactions has proven to be an arduous task. Two methods have evolved in the last two years: (1) combining entropy-based modeling and microarray analysis to identify gene clustering of genes implicated in the same

regulatory pathway, and (2) using random algorithms to maximize transition probability. The first method groups the protein groups with more mutual information above a threshold by computing the mutual information of each gene pair. The mutual information is based on an entropy-based methodology and is obtained by adding up all of the frequency patterns that gene pairs occur in. Gene expressions are separated into discrete histograms, and the mutual information between each gene pair is calculated to determine entropy. Increased mutual information indicates a direct genetic link. Genes associated with the same pathway have been shown statistically to cluster together. Many signaling pathways in the yeast-based system have been found using this cluster analysis. The analysis is a versatile method that works with both eukaryotic and prokaryotic systems (139-142).

The transient temporal behavior of numerous genes participating in the regulation and auto-regulatory mechanisms of operons, a co-transcribed gene-group inside a pathway involved in a common functionality, cannot be explained, not even by determining connection. Since the data corresponds to an equilibrium state of processes, hybridization-based microarray analysis is unable to capture the modeling of transitory activity of genes. It is necessary to research the general organization and behavior, including transitory behavior and stress responses, in order to comprehend harmful bacterial strains and malfunctioning cells (143-145).

Reexamining Microbial Evolution

In order to correlate and categorize the genomes into different families and to investigate evolution, bioinformatics experts have extensively analyzed many genomes. Numerous researchers have established that point-based mutations leading to specialization and genome restructuring based on gene duplications, gene insertions, gene deletions, gene fusion/fission, horizontal gene transfer, and domain-level restructuring constitute the basis of overall evolution. The three types of evolutionary study approaches are as follows: (1) the use of multiple sequence alignment of 16SrRNA to construct a traditional evolutionary tree using point-based mutation approaches (146-148); (2) The investigation of genome restructuring through gene-level inversion and transposition; and (3) the study based on whole genome comparisons using gene identities of orthologous genes across multiple microbial genomes.

The 16SrRNA method leverages multiple sequence alignment and the 16SrRNA database to create an evolutionary tree. It is based on the idea that conserved genes undergo point mutations because of their slow rate of mutation. This method was deemed quantitatively sound prior to the sequencing of microbial genomes, and it allowed for the identification of three separate domains bacteria, archaea, and eukaryotes using the 16SrRNA database. Because the Archea domain is hyperthermophilic, its 16S rRNA differs somewhat from that of bacteria (148-150).

Scientists have been attempting to construct the evolutionary tree by examining other highly conserved genes since 1998, when numerous microbial genomes became available. According to the findings, there is no obvious differentiation between bacteria and archaea on the evolutionary tree, which fluctuates greatly depending on the selection of conserved genes. The conventional evolutionary trees based on point mutations in 16S RNA have been called into question by this observation as well as the understanding of genome restructuring brought about by domain level and gene level restructuring, such as horizontal gene transfer (82, 150, 151).

According to the second method, the genomic distance between the two organisms is determined by rearranging the genome due to gene shuffling. It is inversion and transposition that cause gene shuffling. As a departure from the conventional gene-order in two genomes, the distance measure serves as the

foundation for this system. The cumulative score for the genome is calculated by adding the breakaway distances of each orthologous gene using this scheme. Between two genomes, this score serves as a measure. Until recently, the use of pairwise comparison rendered the construction of large-scale evolutionary trees unfeasible. However, advancements in parallel algorithms have made the creation of such trees feasible. Does this plan since duplications are mapped to a single gene and insertions and deletions are not included in the assumption, horizontal gene transfer is not included. Duplication, insertion, and deletion of genes and gene domains have been demonstrated to be important aspects of evolution. Particularly duplicated genes are essential to several sensing and transportation networks, including ABC transporters, and should not be disregarded (152-157).

To find the cumulative similarity of two genomes, the third method compares the overall gene content of functionally identical genes. To account for variations in genome size, the data has been standardized. The underlying presumption of this technique is that slow mutation rates only aid in high multiple sequence alignment and that conserved genes are rare and do not form a consensus. Comparisons of entire genomes can counteract the inaccuracy caused through comparison of a single conserved gene. The findings demonstrate that there is no discernible difference in the overall amino acid composition of microorganisms between archaea and bacteria to warrant the classification of archaea as a distinct domain. Furthermore, it is impossible to separate the makeup of other hyperthermophilic bacteria from that of archaea (158-161).

To categorize the genomes, no proteome level method has yet been proposed. A future version of this strategy might rely on pathway alignment and comparative analysis across several genomes. As all three factors are directly involved in the pathway variations, under this scheme, the distance between two genomes could be described by combining, once the pathways are aligned, the cumulative number of gene insertions and deletions in the pathways, gene duplication in the same pathway, and gene shuffling. Nevertheless, more research is needed to determine the precise process integrating these three pathway evolution components (162-165).

Conclusion

Bioinformatics, although still in its infancy, has contributed to fundamental microbiology and biotechnology by developing tools, algorithms, and discoveries that improve the abstract model of microbial cell functioning. The automation of microbial genome sequencing, the creation of integrated databases via the Internet, and the study of genomes to comprehend gene and genome function have been the main contributions of bioinformatics. When comparing genes and genomes, the BLAST-based database search and the Smith-Waterman-based gene-pair alignment technique, along with its variants, are widely utilized. These methods have become the foundation for determining the functionality of genes and genomes. Comparative genome analysis has been extremely successful in identifying conserved function within a genome family, identifying particular genes within a group of genomes, modeling 3D protein structures, and docking biological substances and receptors. These achievements directly influence the creation of vaccines, antimicrobial agents, and sensible medicine formulation. Reconstructing metabolic pathways has become almost entirely automated by combining the knowledge of orthologs and gene functions, gene grouping based on the integration of pairwise genome comparison, co-transcribed gene groups, and graph-based matching of substrates and products catalyzed by enzymes. Nowadays, the focus is on identifying regulatory pathways, identifying interactions between proteins, DNA, and RNA, simulating metabolic reactions to examine the impact of reaction rates, and analyzing experimental data

from microarray data to investigate the relationship between gene expressions and stress conditions. The knowledge gained from wet laboratories and the computational tools and algorithms at hand are vital to the majority of bioinformatics procedures. Regretfully, due to the large number of unknowns in genomics and proteomics, neither resource can effectively handle the massive amount of data required for interpretation. There are many gaps in the overall picture of gene functions in many recently sequenced genomes since the wet lab data only provides a small selection of gene functions.

The transcription factors for regulatory pathways, metabolic pathways, variants in metabolic pathways, and novel findings to identify candidate genes for vaccines and rational medication design can all be modeled mathematically. Nevertheless, there are a lot of false positives and false negatives in the modeling findings. Wet-lab tests are required to confirm and validate these findings. Complete verification, however, is becoming unfeasible due to a lack of resources, specialists, coordination issues, and dynamic bioinformatics databases brought about by fresh research and findings.

Microbial wet lab investigations will become more goal-focused as a result of improved cell visualization tools, abstract genomics models based on current bioinformatics analysis, and their integration with existing biochemical knowledge. Wet-lab technique development and bioinformatics advancement must continue to be targeted, interdependent, and complementary to one another for both current and future biotechnology advancement. The application of strategies in an integrated manner to manipulate microbial cells at the systemic level will become increasingly important in the future.

Declarations

Author's contribution

There are three authors who contributed the contents of the manuscript. The review is based upon the published current research in the area of microbial bioinformatics. The lead and co-responding author Prof. Rajrupa Ghosh contributed the main idea and concept of the content. Besides the data and information of Bioinformatics, Microbiology and Biotechnology were collected from different published journals and prescribed by this corresponding author.

The second and third author Prof. Rahul Deb Bera and Prof. Subhra Chandra Chandra contributed to collect the information from different published articles.

Conflict of Interest

The authors have no conflicts of interest to declare. There is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

No any data taken from any other journals or web sources.

Acknowledgement

The author was only able to recognize a limited number of pertinent papers that directly contributed to the field of microbial bioinformatics because of the review's narrow scope. The author thanks all the microbial

bioinformatics researchers who have contributed, either directly or indirectly, to the field, particularly those who have worked on developing algorithms and eukaryotic research using models based on yeast. Additionally, the author thanks the anonymous referees for their insightful comments provided during the review process.

Reference

1. Meng X, Wang M, Luo M, Sun L, Yan Q, Liu Y. Systematic evaluation of multiple NGS platforms for structural variants detection. Journal of Biological Chemistry. 2023;299(12).
2. Tanzo JT, Li VL, Wiggenhorn AL, Moya-Garzon MD, Wei W, Lyu X, et al. CYP4F2 is a human-specific determinant of circulating N-acyl amino acid levels. Journal of Biological Chemistry. 2023;299(6).
3. Cao X, Wang L, Selby CP, Lindsey-Boltz LA, Sancar A. Analysis of mammalian circadian clock protein complexes over a circadian cycle. Journal of Biological Chemistry. 2023;299(3).
4. Amiri-Dashatan N, Koushki M, Abbaszadeh HA, Rostami-Nejad M, Rezaei-Tavirani M. Proteomics Applications in Health: Biomarker and Drug Discovery and Food Industry. Iranian journal of pharmaceutical research : IJPR. 2018 Fall;17(4):1523-36. PubMed PMID: 30568709. Pubmed Central PMCID: PMC6269565. Epub 2018/12/21. eng.
5. Planque C, Kulasingam V, Smith CR, Reckamp K, Goodglick L, Diamandis EP. Identification of five candidate lung cancer biomarkers by proteomics analysis of conditioned media of four lung cancer cell lines. Molecular & cellular proteomics. 2009;8(12):2746-58.
6. Sallam RM. Proteomics in cancer biomarkers discovery: challenges and applications. Disease markers. 2015;2015.
7. Calderón-González KG, Hernández-Monge J, Herrera-Aguirre ME, Luna-Arias JP. Bioinformatics tools for proteomics data interpretation. Modern Proteomics–Sample Preparation, Analysis and Practical Applications. 2016:281-341.
8. Palagi PM, Hernandez P, Walther D, Appel RD. Proteome informatics I: bioinformatics tools for processing experimental data. Proteomics. 2006;6(20):5435-44.
9. Deutsch EW, Lam H, Aebersold R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. Physiological genomics. 2008;33(1):18-25.
10. Blueggel M, Chamrad D, Meyer HE. Bioinformatics in proteomics. Current pharmaceutical biotechnology. 2004;5(1):79-88.
11. Englbrecht CC, Facius A. Bioinformatics challenges in proteomics. Combinatorial chemistry & high throughput screening. 2005;8(8):705-15.
12. Xu M, Xu C, Chen M, Xiao Z, Wang Y, Xu Y, et al. Comparative analysis of commonly used bioinformatics software based on omics. Gene Reports. 2023 2023/09/01/;32:101800.
13. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Frontiers in Genetics. 2017 2017-March-06;8. English.
14. Slavkin HC. From Phenotype to Genotype: Enter Genomics and Transformation of Primary Health Care around the World. Journal of dental research. 2014 Jul;93(7 Suppl):3S-6S. PubMed PMID: 24799423. Pubmed Central PMCID: PMC4293721. Epub 2014/05/07. eng.
15. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype–phenotype associations. Bioinformatics. 2016;32(20):3207-9.

16. Du J, Wang C, Wang L, Mao S, Zhu B, Li Z, et al. Automatic block-wise genotype-phenotype association detection based on hidden Markov model. BMC Bioinformatics. 2023 2023/04/07;24(1):138.
17. Sulheim S, Fossheim FA, Wentzel A, Almaas E. Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters. BMC Bioinformatics. 2021 2021/02/23;22(1):81.
18. Qi Q, Li J, Cheng J. Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods. BMC proceedings. 2014;8(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer):S5. PubMed PMID: 25374614. Pubmed Central PMCID: PMC4202177. Epub 2014/11/07. eng.
19. de Crécy-Lagard V. Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. Computational and Structural Biotechnology Journal. 2014 2014/06/01/;10(16):41-50.
20. Heidorn T, Camsund D, Huang H-H, Lindberg P, Oliveira P, Stensjö K, et al. Chapter Twenty-Four - Synthetic Biology in Cyanobacteria: Engineering and Analyzing Novel Functions. In: Voigt C, editor. Methods in Enzymology. 497: Academic Press; 2011. p. 539-79.
21. Ahmad Mir R, Mansoor Shafi S, Zargar SM. Chapter 4 - Analysis of genomes—II∗∗Tracing evolution of eukaryotes by understanding genomes, identifying the functional regions of genomes—DNA footprinting, gel retardation assay or electrophoretic mobility shift assay, and chromatin immunoprecipitation. In: Ahmad Mir R, Mansoor Shafi S, Zargar SM, editors. Principles of Genomics and Proteomics: Elsevier; 2023. p. 65-87.
22. Mukherjee S, Nithin C. Chapter 11 - Advanced computational tools for quantitative analysis of protein–nucleic acid interfaces. In: Tripathi T, Dubey VK, editors. Advances in Protein Molecular and Structural Biology Methods: Academic Press; 2022. p. 163-80.
23. Gao W, Mahajan SP, Sulam J, Gray JJ. Deep Learning in Protein Structural Modeling and Design. Patterns. 2020 2020/12/11/;1(9):100142.
24. Rost B. Protein structure prediction in 1D, 2D, and 3D. Encyclopedia of Computational Chemistry. 1998:2242-55.
25. Schmidt T, Bergner A, Schwede T. Modelling three-dimensional protein structures for applications in drug design. Drug discovery today. 2014;19(7):890-7.
26. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O, et al. G protein-coupled receptors: in silico drug discovery in 3D. Proceedings of the National Academy of Sciences. 2004;101(31):11304-9.
27. Sinha R, Vidyarthi AS. A molecular docking study of anticancer drug paclitaxel and its analogues. 2011.
28. Wang J-F, Chou K-C. Insights from modeling the 3D structure of New Delhi metallo-β-lactamse and its binding interactions with antibiotic drugs. PLoS One. 2011;6(4):e18414.
29. Claverie J-M. Computational methods for the identification of genes in vertebrate genomic sequences. Human molecular genetics. 1997;6(10):1735-44.
30. Koonin E, Galperin MY. Sequence—evolution—function: computational approaches in comparative genomics. 2002.
31. Ye N. Data mining: theories, algorithms, and examples: CRC press; 2013.
32. Kantardzic M. Data mining: concepts, models, methods, and algorithms: John Wiley & Sons; 2011.

33. Zaki MJ, Meira W. Data mining and analysis: fundamental concepts and algorithms: Cambridge University Press; 2014.
34. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowledge and information systems. 2008;14:1-37.
35. Mughal MJH. Data mining: Web data mining techniques, tools and algorithms: An overview. International Journal of Advanced Computer Science and Applications. 2018;9(6).
36. Dong G, Bailey J. Contrast data mining: concepts, algorithms, and applications: CRC Press; 2012.
37. Vakil V, Trappe W. Drug combinations: mathematical modeling and networking methods. Pharmaceutics. 2019;11(5):208.
38. de Castro Vivas R, Sant'Anna AMO, Esquerre KPO, Freires FGM. Integrated method combining analytical and mathematical models for the evaluation and optimization of sustainable supply chains: A Brazilian case study. Computers & Industrial Engineering. 2020;139:105670.
39. Pimentel C, Alvelos F. Integrated urban freight logistics combining passenger and freight flows–mathematical model proposal. Transportation research procedia. 2018;30:80-9.
40. Akgün İ, Erdal H. Solving an ammunition distribution network design problem using multi-objective mathematical modeling, combined AHP-TOPSIS, and GIS. Computers & Industrial Engineering. 2019;129:512-28.
41. Barquero B, Bosch M, Romo A. Mathematical modelling in teacher education: dealing with institutional constraints. ZDM. 2018;50:31-43.
42. Köser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. 2012.
43. Brosch R, Gordon SV, Billault A, Garnier T, Eiglmeier K, Soravito C, et al. Use of a Mycobacterium tuberculosis H37Rv bacterial artificial chromosome library for genome mapping, sequencing, and comparative genomics. Infection and immunity. 1998;66(5):2221-9.
44. Kwong JC, McCallum N, Sintchenko V, Howden BP. Whole genome sequencing in clinical and public health microbiology. Pathology. 2015;47(3):199-210.
45. Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in genetics. 2015;6:348.
46. Christensen H. Introduction to bioinformatics in microbiology: Springer; 2018.
47. Logares R, Haverkamp TH, Kumar S, Lanzén A, Nederbragt AJ, Quince C, et al. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. Journal of microbiological methods. 2012;91(1):106-13.
48. Chen K, Pachter L. Bioinformatics for whole-genome shotgun sequencing of microbial communities. PLoS computational biology. 2005;1(2):e24.
49. Bansal AK. Bioinformatics in microbial biotechnology–a mini review. Microbial Cell Factories. 2005;4:1-11.
50. Dunne W, Westblade L, Ford B. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. European journal of clinical microbiology & infectious diseases. 2012;31(8):1719-26.
51. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic acids research. 1999 Dec 1;27(23):4636-41. PubMed PMID: 10556321. Pubmed Central PMCID: PMC148753. Epub 1999/11/11. eng.

52. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleic acids research. 1998 Jan 15;26(2):544-8. PubMed PMID: 9421513. Pubmed Central PMCID: PMC147303. Epub 1998/02/28. eng.
53. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007 Mar 15;23(6):673-9. PubMed PMID: 17237039. Pubmed Central PMCID: PMC2387122. Epub 2007/01/24. eng.
54. A Decision Tree System for Finding Genes in DNA. Journal of Computational Biology. 1998;5(4):667-80. PubMed PMID: 10072083.
55. Che D, Hockenbury C, Marmelstein R, Rasheed K. Classification of genomic islands using decision trees and their ensemble algorithms. BMC Genomics. 2010 2010/11/02;11(2):S1.
56. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Current genomics. 2009 Sep;10(6):402-15. PubMed PMID: 20190955. Pubmed Central PMCID: PMC2766791. Epub 2010/03/02. eng.
57. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics. 2006 2006/03/16;7(1):142.
58. Donkor E, Dayie N, Adiku T. Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). Journal of Bioinformatics and Sequence Analysis. 2014 04/30;6:1-6.
59. Pertsemlidis A, Fondon JW, 3rd. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome biology. 2001;2(10):REVIEWS2002. PubMed PMID: 11597340. Pubmed Central PMCID: PMC138974. Epub 2001/10/13. eng.
60. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research. 2004 Jul 1;32(Web Server issue):W20-5. PubMed PMID: 15215342. Pubmed Central PMCID: PMC441573. Epub 2004/06/25. eng.
61. Pearson WR. Finding Protein and Nucleotide Similarities with FASTA. Current protocols in bioinformatics. 2016 Mar 24;53:3 9 1-3 9 25. PubMed PMID: 27010337. Pubmed Central PMCID: PMC5072362. Epub 2016/03/25. eng.
62. Pietrokovski S, Henikoff JG, Henikoff S. The Blocks Database—A System for Protein Classification. Nucleic acids research. 1996;24(1):197-200.
63. Polyanovsky V, Lifanov A, Esipova N, Tumanyan V. The ranging of amino acids substitution matrices of various types in accordance with the alignment accuracy criterion. BMC Bioinformatics. 2020 2020/09/14;21(11):294.
64. Trivedi R, Nagarajaram HA. Substitution scoring matrices for proteins - An overview. Protein science : a publication of the Protein Society. 2020 Nov;29(11):2150-63. PubMed PMID: 32954566. Pubmed Central PMCID: PMC7586916. Epub 2020/09/22. eng.
65. Mount D. Comparison of the PAM and BLOSUM amino acid substitution matrices. CSH protocols. 2008 06/01;2008:pdb.ip59.
66. Brick K, Pizzi E. A novel series of compositionally biased substitution matrices for comparing Plasmodium proteins. BMC Bioinformatics. 2008 2008/05/16;9(1):236.
67. Bhagwat M, Aravind L. PSI-BLAST Tutorial. In: Bergman NH, editor. Comparative Genomics. Totowa, NJ: Humana Press; 2008. p. 177-86.
68. Li Y, Chia N, Lauria M, Bundschuh R. A performance enhanced PSI-BLAST based on hybrid alignment. Bioinformatics. 2011;27(1):31-7.

69. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends in biochemical sciences. 1998;23(11):444-7.
70. Bhagwat M, Aravind L. PSI-BLAST tutorial. Methods in molecular biology (Clifton, NJ). 2007;395:177-86. PubMed PMID: 17993673. Pubmed Central PMCID: PMC4781153. Epub 2007/11/13. eng.
71. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402.
72. Mount DW. Comparison of the PAM and BLOSUM Amino Acid Substitution Matrices. Cold Spring Harbor Protocols. 2008 June 1, 2008;2008(6):pdb.ip59.
73. Bandyopadhyay S, Mitra R. A Parallel Pairwise Local Sequence Alignment Algorithm. IEEE Transactions on NanoBioscience. 2009;8(2):139-46.
74. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic acids research. 2010;38(suppl_2):W7-W13.
75. Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics. 2017 2017/10/01/;109(5):419-31.
76. Wei L, Liu Y, Dubchak I, Shon J, Park J. Comparative genomics approaches to study organism similarities and differences. Journal of Biomedical Informatics. 2002 2002/04/01/;35(2):142-50.
77. Shatnawi M. Chapter 6 - Review of Recent Protein-Protein Interaction Techniques. In: Tran QN, Arabnia H, editors. Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology. Boston: Morgan Kaufmann; 2015. p. 99-121.
78. Chatzou M, Magis C, Chang J-M, Kemena C, Bussotti G, Erb I, et al. Multiple sequence alignment modeling: methods and applications. Briefings in Bioinformatics. 2016;17(6):1009-23.
79. Thompson JD, Linard B, Lecompte O, Poch O. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLOS ONE. 2011;6(3):e18093.
80. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792-7. PubMed PMID: 15034147. Pubmed Central PMCID: PMC390337. Epub 2004/03/23. eng.
81. Ibrahim MK, Yusof UK, Eisa TAE, Nasser M. Enhanced Genetic Method for Optimizing Multiple Sequence Alignment. Mathematics. 2023;11(22):4578. PubMed PMID: doi:10.3390/math11224578.
82. Zhan Q, Wang N, Jin S, Tan R, Jiang Q, Wang Y. ProbPFP: a multiple sequence alignment algorithm combining hidden Markov model optimized by particle swarm optimization with partition function. BMC Bioinformatics. 2019 2019/11/25;20(18):573.
83. Mulia S, Mishra D, Jena T. Profile HMM based Multiple Sequence Alignment for DNA Sequences. Procedia Engineering. 2012 2012/01/01/;38:1783-7.
84. Pei J, Grishin NV. MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic acids research. 2006;34(16):4364-74.
85. Papaleo E, Saladino G, Lambrughi M, Lindorff-Larsen K, Gervasio FL, Nussinov R. The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. Chemical Reviews. 2016 2016/06/08;116(11):6391-423.

86. Wriggers W, Chakravarty S, Jennings PA. Control of protein functional dynamics by peptide linkers. Peptide Science: Original Research on Biomolecules. 2005;80(6):736-46.
87. Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. Nucleic acids research. 1998;26(1):323-6.
88. Sharma AK, Becker JW, Ottesen EA, Bryant JA, Duhamel S, Karl DM, et al. Distinct dissolved organic matter sources induce rapid transcriptional responses in coexisting populations of P rochlorococcus, P elagibacter and the OM60 clade. Environmental Microbiology. 2014;16(9):2815-30.
89. Berón CM, Curatti L, Salerno GL. New Strategy for Identification of Novel cry-Type Genes from Bacillus thuringiensis Strains. Applied and Environmental Microbiology. 2005;71(2):761-5.
90. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. Nucleic acids research. 2012;40(D1):D302-D5.
91. Ulrich LE, Zhulin IB. MiST: a microbial signal transduction database. Nucleic acids research. 2007;35(suppl_1):D386-D90.
92. Uchiyama I. MBGD: microbial genome database for comparative analysis. Nucleic acids research. 2003;31(1):58-62.
93. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research. 2005;33(suppl_2):W244-W8.
94. Nithin C, Ghosh P, Bujnicki JM. Bioinformatics tools and benchmarks for computational docking and 3D structure prediction of RNA-protein complexes. Genes. 2018;9(9):432.
95. Chen Y-C. Beware of docking! Trends in pharmacological sciences. 2015;36(2):78-95.
96. Tovchigrechko A, Wells CA, Vakser IA. Docking of protein models. Protein Science. 2002;11(8):1888-96.
97. Muchtaridi M, Dermawan D, Yusuf M. Molecular docking, 3D structure-based pharmacophore modeling, and ADME prediction of alpha mangostin and its derivatives against estrogen receptor alpha. Journal of Young Pharmacists. 2018;10(3):252.
98. Humblet C, Dunbar Jr JB. 3D Database searching and docking strategles. Annual Reports in Medicinal Chemistry. 1993;28:275-84.
99. Morris GM, Lim-Wilby M. Molecular docking. Molecular modeling of proteins. 2008:365-82.
100. Nayeem A, Sitkoff D, Krystek Jr S. A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. Protein Science. 2006;15(4):808-24.
101. Wiltgen M. Algorithms for structure comparison and analysis: Homology modelling of proteins: Elsevier Cambridge; 2018.
102. Xiang Z. Advances in homology protein structure modeling. Current Protein and Peptide Science. 2006;7(3):217-27.
103. Hasani HJ, Barakat K. Homology modeling: an overview of fundamentals and tools. Int Rev Model Simul. 2017;10(2):1-14.
104. Munsamy G, Soliman ME. Homology modeling in drug discovery-an update on the last decade. Letters in Drug Design & Discovery. 2017;14(9):1099-111.
105. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nature protocols. 2009;4(1):1-13.

106. Dorn M, e Silva MB, Buriol LS, Lamb LC. Three-dimensional protein structure prediction: Methods and computational strategies. Computational biology and chemistry. 2014;53:251-76.
107. Guedes IA, de Magalhães CS, Dardenne LE. Receptor–ligand molecular docking. Biophysical reviews. 2014;6:75-87.
108. Pujadas G, Vaque M, Ardevol A, Blade C, Salvado M, Blay M, et al. Protein-ligand docking: A review of recent advances and future perspectives. Current Pharmaceutical Analysis. 2008;4(1):1-19.
109. Fukunishi Y, Mikami Y, Nakamura H. Similarities among receptor pockets and among compounds: analysis and application to in silico ligand screening. Journal of Molecular Graphics and Modelling. 2005;24(1):34-45.
110. Sousa SF, Fernandes PA, Ramos MJ. Protein–ligand docking: current status and future challenges. Proteins: Structure, Function, and Bioinformatics. 2006;65(1):15-26.
111. Dunn CW, Zapata F, Munro C, Siebert S, Hejnol A. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proceedings of the National Academy of Sciences. 2018;115(3):E409-E17.
112. Richter M, Rosselló-Móra R, Oliver Glöckner F, Peplies J. JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. Bioinformatics. 2016;32(6):929-31.
113. Torreno O, Trelles O. Breaking the computational barriers of pairwise genome comparison. BMC bioinformatics. 2015;16:1-13.
114. Mueller C, Dalkilic MM, Lumsdaine A. High-performance direct pairwise comparison of large genomic sequences. IEEE Transactions on Parallel and Distributed Systems. 2006;17(8):764-72.
115. Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC bioinformatics. 2011;12(1):1-11.
116. Margulies EH, Chen CW, Green ED. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. Trends in Genetics. 2006;22(4):187-93.
117. Elnitski L, Riemer C, Petrykowska H, Florea L, Schwartz S, Miller W, et al. PipTools: a computational toolkit to annotate and analyze pairwise comparisons of genomic sequences. Genomics. 2002;80(6):681-90.
118. Bao Y, Chetvernin V, Tatusova T. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. Archives of virology. 2014;159:3293-304.
119. Bansal AK, Bork P, Stuckey PJ. Automated pair-wise comparisons of microbial genomes. Math Model Sci Comput. 1998;9:1-23.
120. Salaverry LS, Lombardo T, Cabral-Lorenzo MC, Gil-Folgar ML, Rey-Roldán EB, Kornblihtt LI, et al. Metabolic plasticity in blast crisis-chronic myeloid leukaemia cells under hypoxia reduces the cytotoxic potency of drugs targeting mitochondria. Discover Oncology. 2022 Jul 8;13(1):60. PubMed PMID: 35802257. Pubmed Central PMCID: PMC9270554. Epub 2022/07/09. eng.
121. Sadat MA, Jeon J, Mir AA, Choi J, Choi J, Lee Y-H. Regulation of Cellular Diacylglycerol through Lipid Phosphate Phosphatases Is Required for Pathogenesis of the Rice Blast Fungus, Magnaporthe oryzae. PLOS ONE. 2014;9(6):e100726.
122. Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. Microbial genome analysis: the COG approach. Briefings in bioinformatics. 2019;20(4):1063-70.

123. Bansal AK, Meyer TE. Evolutionary analysis by whole-genome comparisons. Journal of bacteriology. 2002;184(8):2260-72.
124. Wu Y, Kameshwar AKS, Zhang B, Chen F, Qin W, Meng M, et al. Genome and transcriptome analysis of rock-dissolving Pseudomonas sp. NLX-4 strain. Bioresources and Bioprocessing. 2022 2022/06/01;9(1):63.
125. Peng Y, Gao X, Li R, Cao G. Transcriptome sequencing and De Novo analysis of Youngia japonica using the illumina platform. PLoS One. 2014;9(3):e90636. PubMed PMID: 24595283. Pubmed Central PMCID: PMC3942458. Epub 2014/03/07. eng.
126. Braun IR, Lawrence-Dill CJ. Automated Methods Enable Direct Computation on Phenotypic Descriptions for Novel Candidate Gene Prediction. Frontiers in plant science. 2019;10:1629. PubMed PMID: 31998331. Pubmed Central PMCID: PMC6965352. Epub 2020/01/31. eng.
127. Jessica AC, Merrill EA, Fatmagül B, Isabella L, Alison EP, Roy DW. Phenotypic similarity is a measure of functional redundancy within homologous gene families. bioRxiv. 2022:2022.07.25.501402.
128. Santos F, Spinler JK, Saulnier DMA, Molenaar D, Teusink B, de Vos WM, et al. Functional identification in Lactobacillus reuteri of a PocR-like transcription factor regulating glycerol utilization and vitamin B12 synthesis. Microbial Cell Factories. 2011 2011/07/21;10(1):55.
129. Sinaeda A, Aymeric N, Cédric J, Alain B, Pierre T, Sébastien R. AURTHO: autoregulation as facilitator of cis-acting element discovery of orthologous transcription factors. bioRxiv. 2022:2022.04.06.487287.
130. Moran BM, Payne C, Langdon Q, Powell DL, Brandvain Y, Schumer M. The genomic consequences of hybridization. eLife. 2021 2021/08/04;10:e69016.
131. Berthenet E, Bénéjat L, Ménard A, Varon C, Lacomme S, Gontier E, et al. Whole-Genome Sequencing and Bioinformatics as Pertinent Tools to Support Helicobacteracae Taxonomy, Based on Three Strains Suspected to Belong to Novel Helicobacter Species. Frontiers in Microbiology. 2019 2019-December-06;10. English.
132. Anderssen S, Naômé A, Jadot C, Brans A, Tocquin P, Rigali S. AURTHO: Autoregulation of transcription factors as facilitator of cis-acting element discovery. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2022 2022/07/01/;1865(5):194847.
133. Gauthier GM, Sullivan TD, Gallardo SS, Brandhorst TT, Vanden Wymelenberg AJ, Cuomo CA, et al. SREB, a GATA Transcription Factor That Directs Disparate Fates in Blastomyces dermatitidis Including Morphogenesis and Siderophore Biosynthesis. PLOS Pathogens. 2010;6(4):e1000846.
134. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic acids research. 2001 Jul 1;29(13):2860-74. PubMed PMID: 11433033. Pubmed Central PMCID: PMC55782. Epub 2001/07/04. eng.
135. Corona RI, Guo JT. Statistical analysis of structural determinants for protein-DNA-binding specificity. Proteins. 2016 Aug;84(8):1147-61. PubMed PMID: 27147539. Pubmed Central PMCID: PMC4945413. Epub 2016/05/06. eng.
136. Lin M, Guo J-t. New insights into protein–DNA binding specificity from hydrogen bond based comparative study. Nucleic acids research. 2019;47(21):11103-13.

137. He H, Yang M, Li S, Zhang G, Ding Z, Zhang L, et al. Mechanisms and biotechnological applications of transcription factors. Synthetic and systems biotechnology. 2023 Dec;8(4):565-77. PubMed PMID: 37691767. Pubmed Central PMCID: PMC10482752. Epub 2023/09/11. eng.
138. Titsias MK, Honkela A, Lawrence ND, Rattray M. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. BMC Systems Biology. 2012 2012/05/30;6(1):53.
139. Ali S, Duan J, Charles TC, Glick BR. A bioinformatics approach to the determination of genes involved in endophytic behavior in Burkholderia spp. Journal of Theoretical Biology. 2014 2014/02/21/;343:193-8.
140. Nashiry A, Sarmin Sumi S, Islam S, Quinn JMW, Moni MA. Bioinformatics and system biology approach to identify the influences of COVID-19 on cardiovascular and hypertensive comorbidities. Briefings in Bioinformatics. 2021;22(2):1387-401.
141. Yuan H, Jiang T, Zhang WD, Yang Z, Luo S, Wang X, et al. Multiomics and bioinformatics identify differentially expressed effectors in the brain of Toxoplasma gondii infected masked palm civet. Frontiers in cellular and infection microbiology. 2023;13:1267629. PubMed PMID: 37818043. Pubmed Central PMCID: PMC10561248. Epub 2023/10/11. eng.
142. Kindrachuk J, Ork B, Hart BJ, Mazur S, Holbrook MR, Frieman MB, et al. Antiviral Potential of ERK/MAPK and PI3K/AKT/mTOR Signaling Modulation for Middle East Respiratory Syndrome Coronavirus Infection as Identified by Temporal Kinome Analysis. Antimicrobial Agents and Chemotherapy. 2015;59(2):1088-99.
143. Li Z, Wang C, Zhang X, Xu X, Wang M, Dong L. Crosstalk between septic shock and venous thromboembolism: a bioinformatics and immunoassay analysis. Frontiers in cellular and infection microbiology. 2023;13:1235269. PubMed PMID: 38029239. Pubmed Central PMCID: PMC10666789. Epub 2023/11/29. eng.
144. Shi X, Wegener-Feldbrügge S, Huntley S, Hamann N, Hedderich R, Søgaard-Andersen L. Bioinformatics and Experimental Analysis of Proteins of Two-Component Systems in Myxococcus xanthus. Journal of Bacteriology. 2008;190(2):613-24.
145. Mahbobi R, Fallah F, Behmanesh A, Yadegar A, Hakemi-Vala M, Ehsanzadeh SJ, et al. Helicobacter pylori Infection Mediates Inflammation and Tumorigenesis-Associated Genes Through miR-155-5p: An Integrative Omics and Bioinformatics-Based Investigation. Current Microbiology. 2022 2022/05/13;79(7):192.
146. Clarridge JE, 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical microbiology reviews. 2004 Oct;17(4):840-62, table of contents. PubMed PMID: 15489351. Pubmed Central PMCID: PMC523561. Epub 2004/10/19. eng.
147. Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer K, et al. The all-species living tree PROJECT: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and applied microbiology. 2008 09/01;31:241-50.
148. Green AG, Swithers KS, Gogarten JF, Gogarten JP. Reconstruction of ancestral 16S rRNA reveals mutation bias in the evolution of optimal growth temperature in the Thermotogae phylum. Molecular biology and evolution. 2013 Nov;30(11):2463-74. PubMed PMID: 23966548. Epub 2013/08/24. eng.

149. Hassler HB, Probert B, Moore C, Lawson E, Jackson RW, Russell BT, et al. Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies. Microbiome. 2022 2022/07/08;10(1):104.
150. Budimir I, Giampieri E, Saccenti E, Suarez-Diez M, Tarozzi M, Dall'Olio D, et al. Intraspecies characterization of bacteria via evolutionary modeling of protein domains. Scientific Reports. 2022 2022/10/05;12(1):16595.
151. Schloss PD. The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. PLOS Computational Biology. 2010;6(7):e1000844.
152. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences of the United States of America. 2003 Sep 30;100(20):11484-9. PubMed PMID: 14500911. Pubmed Central PMCID: PMC208784. Epub 2003/09/23. eng.
153. Newman S, Hermetz KE, Weckselblatt B, Rudd MK. Next-generation sequencing of duplication CNVs reveals that most are tandem and some create fusion genes at breakpoints. American journal of human genetics. 2015 Feb 5;96(2):208-20. PubMed PMID: 25640679. Pubmed Central PMCID: PMC4320257. Epub 2015/02/03. eng.
154. Ebler J, Schönhuth A, Marschall T. Genotyping inversions and tandem duplications. Bioinformatics. 2017;33(24):4015-23.
155. Shao H, Ganesamoorthy D, Duarte T, Cao MD, Hoggart CJ, Coin LJM. npInv: accurate detection and genotyping of inversions using long read sub-alignment. BMC Bioinformatics. 2018 2018/07/13;19(1):261.
156. Suzuki T, Tsurusaki Y, Nakashima M, Miyake N, Saitsu H, Takeda S, et al. Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing. Journal of Human Genetics. 2014 2014/12/01;59(12):649-54.
157. Mantere T, Neveling K, Pebrel-Richard C, Benoist M, van der Zande G, Kater-Baats E, et al. Optical genome mapping enables constitutional chromosomal aberration detection. The American Journal of Human Genetics. 2021 2021/08/05/;108(8):1409-22.
158. Kellner S, Spang A, Offre P, Szöllősi GJ, Petitjean C, Williams TA. Genome size evolution in the Archaea. Emerging topics in life sciences. 2018 Dec 14;2(4):595-605. PubMed PMID: 33525826. Pubmed Central PMCID: PMC7289037. Epub 2018/12/14. eng.
159. Trouche B, Schauberger C, Bouderka F, Auguet J-C, Belser C, Poulain J, et al. Distribution and genomic variation of ammonia-oxidizing archaea in abyssal and hadal surface sediments. ISME Communications. 2023 2023/12/22;3(1):133.
160. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic acids research. 2008 Dec;36(21):6688-719. PubMed PMID: 18948295. Pubmed Central PMCID: PMC2588523. Epub 2008/10/25. eng.
161. Grogan DW. Archaea. In: Maloy S, Hughes K, editors. Brenner's Encyclopedia of Genetics (Second Edition). San Diego: Academic Press; 2013. p. 180-2.
162. Hjelm LN, Chin EL, Hegde MR, Coffee BW, Bean LJ. A simple method to confirm and size deletion, duplication, and insertion mutations detected by sequence analysis. The Journal of molecular diagnostics : JMD. 2010 Sep;12(5):607-10. PubMed PMID: 20639189. Pubmed Central PMCID: PMC2928424. Epub 2010/07/20. eng.

163. Ceyhan-Birsoy O, Pugh TJ, Bowser MJ, Hynes E, Frisella AL, Mahanta LM, et al. Next generation sequencing-based copy number analysis reveals low prevalence of deletions and duplications in 46 genes associated with genetic cardiomyopathies. Molecular genetics & genomic medicine. 2016 Mar;4(2):143-51. PubMed PMID: 27066507. Pubmed Central PMCID: PMC4799872. Epub 2016/04/12. eng.
164. García-Castaño A, Madariaga L, Azriel S, Pérez de Nanclares G, Martínez de LaPiscina I, Martínez R, et al. Identification of a novel large CASR deletion in a patient with familial hypocalciuric hypercalcemia. Endocrinology, diabetes & metabolism case reports. 2018;2018. PubMed PMID: 30530875. Pubmed Central PMCID: PMC6280130. Epub 2018/12/12. eng.
165. Oehler J, Morrow CA, Whitby MC. Gene duplication and deletion caused by over-replication at a fork barrier. Nature Communications. 2023 2023/11/25;14(1):7730.