Journal of Studies in Medical Sciences; Vol. 1, No. 1, 2025
ISSN 2810-9899
Published by July Press

Application of Bioinformatics in Medical Scientific Research

Shaode Feng[1]

[1] Shandong University of Traditional Chinese Medicine, Jinan, China

Correspondence: Shaode Feng, Shandong University of Traditional Chinese Medicine, Jinan, Shandong, China.

Received: August 7, 2025 Accepted: September 15, 2025 Online Published: October 5, 2025

doi:10.20849/jsms.v1i1.1529 URL: https://doi.org/10.20849/jsms.v1i1.1529

Abstract

Bioinformatics is a scientific discipline that studies biological information processing (collection, management, and analytical application) to extract new biological knowledge. It bridges biological data and medical scientific research. Bioinformatics plays an indispensable role in genomics research and offers possibilities for solving complex problems in biological systems. This article introduces the basic concepts of bioinformatics, distinguishes between homology and similarity, and provides an overview of the applications of bioinformatics in novel gene discovery, protein identification, and the analysis of biochip results.

Keywords: information science, biology, specialty, medical, bioinformatics, homology, similarity, genomics, proteomics, biochips

1. Introduction

The successful implementation of the Human Genome Project has ushered life science into the information age. The development of genomics, proteomics, and biochip technologies has led to a linear, rapid increase in data related to life sciences. Comprehensively and correctly interpreting this data offers the potential to elucidate the essence of life. Bioinformatics serves as the bridge connecting biological data with medical scientific research. Applying bioinformatics research methods to analyze biological data, propose genes or gene clusters related to disease occurrence and development, and subsequently validate them through experimentation is an efficient research pathway.

Homology and similarity are two fundamental concepts in bioinformatics and are often confused. Homology refers to descent from a common ancestor, sharing an evolutionary origin. It arises from divergent events during biological evolution; for example, homologous sequences for sequence fragments, or homologous proteins for proteins. Similarity refers to common structural or functional features formed by convergent evolution from different ancestral origins. Therefore, similarity simply compares the degree of sameness between the two.

Homology and similarity are distinct concepts; there is no direct equivalence between them. Similar entities are not necessarily homologous because, during evolution, genes or sequences from different sources may converge due to different mutations. Homologous entities generally exhibit similarity, but homologous entities are not necessarily more similar than non-homologous ones.

Homology is divided into orthology and paralogy. Orthology refers to genes originating from a common ancestor in different species, where the genes or proteins perform the same function, such as insulin in mammals. Paralogy refers to genes or proteins within the same organism that perform different but related functions, whose genes share a common origin, such as α-globin, β-globin, and myoglobin, all originating from globin. Orthologs allow for the study of relationships across species, while paralogs produced by gene duplication may help study fundamental evolutionary mechanisms, as duplicated genes can generate separate evolutionary pathways and evolve new characteristics through mutation and adaptation. The concepts of orthology and paralogy are sometimes used loosely and are subject to debate. Some literature also suggests that orthologs can have different functions, and paralogs are not necessarily confined to the same species. (Jensen, 2001; Gogarten & Olendzenski, 1999)

2. Basic Concepts of Bioinformatics

Bioinformatics is the science that studies biological information processing (collection, management, and analytical application) to extract new biological knowledge. Biological information includes various types of data, such as nucleotide and protein sequences, and data on protein secondary and tertiary structures. Primary

data consist of nucleic acid and protein sequences and three-dimensional structural data obtained from experiments; databases constructed from these are called primary databases. Secondary data derived from the analysis of primary data include secondary structures, hydrophobic sites, domains, proteins translated from nucleic acid sequences, and predicted secondary and tertiary structures. Innovative algorithms and software are the foundation for the sustained development of bioinformatics, while high-throughput biological research methods and platform technologies are key techniques for validating bioinformatics research results. Therefore, bioinformatics is a new interdisciplinary field involving biology, mathematics, information science, and other disciplines. Bioinformatics emerged alongside genome research and has expanded with the development of genomics and related research.

"Homology" and "similarity or analogy" are the two most basic concepts in bioinformatics and are often confused. Homology refers to descent from a common ancestor, sharing an evolutionary origin. It arises from divergent events during biological evolution; for example, homologous sequences for sequence fragments, or homologous proteins for proteins. Similarity refers to common structural or functional features formed by convergent evolution from different ancestral origins. Therefore, similarity simply compares the degree of sameness between the two.

Homology and similarity are distinct concepts; there is no direct equivalence between them. Similar entities are not necessarily homologous because, during evolution, genes or sequences from different sources may converge due to different mutations. Homologous entities generally exhibit similarity, but homologous entities are not necessarily more similar than non-homologous ones.

Homology is divided into orthology and paralogy. Orthology refers to genes originating from a common ancestor in different species, where the genes or proteins perform the same function, such as insulin in mammals. Paralogy refers to genes or proteins within the same organism that perform different but related functions, whose genes share a common origin, such as α-globin, β-globin, and myoglobin, all originating from globin. Orthologs allow for the study of relationships across species, while paralogs produced by gene duplication may help study fundamental evolutionary mechanisms, as duplicated genes can generate separate evolutionary pathways and evolve new characteristics through mutation and adaptation . The concepts of orthology and paralogy are sometimes used loosely and are subject to debate. Some literature also suggests that orthologs can have different functions, and paralogs are not necessarily confined to the same species. (Jensen, 2001; Gogarten & Olendzenski, 1999)

3. Main Medical Applications of Bioinformatics Research

Bioinformatics research involves two levels: first, the collection, organization, and servicing of massive data, i.e., managing the data well; second, discovering new patterns from it, i.e., using the data well. According to the central dogma of genetics, biological data involve DNA, RNA, and protein data, as well as data generated by related technologies. The following mainly covers applications currently related to basic medical research.

3.1 Discovery of Novel Genes

The occurrence and development of diseases are related to changes in specific genes. Identifying disease-related genes is a direction scientists are actively exploring. Methods such as mRNA differential display, suppression subtractive hybridization, cDNA microarray, and serial analysis of gene expression (SAGE) yield numerous differentially expressed sequence tags (ESTs), which need further identification. Sequence similarity searches are typically performed using BLAST or FASTA tools (Table 1). Our laboratory has established a high-throughput EST analysis platform based on the free Linux operating system and low-cost PCs. Using the Phred/Phrap/Consed software suite and self-compiled Perl programs, along with common nucleic acid sequence databases, we have automated the entire analysis process for large batches of differential gene fragments, from sequencing chromatograms to nucleic acid sequence conversion, sequence assembly, and sequence alignment. Using this platform, we successfully analyzed over 300 differentially expressed sequences in colorectal tumors, achieving good results. Similarly, an analysis platform for proteomics research results can be established based on this principle.

Furthermore, there are existing software tools for finding novel genes conforming to gene structures within the whole genome, such as Genefinder and GeneScan.

Through the above automated analysis, we can obtain information valuable for further validation: 1) Is it a novel gene? 2) Does it have homology or similarity to known genes? 3) Does it belong to a known multigene family? 4) What function might the protein encoded by this gene perform? The results obtained from bioinformatics are then validated through further experiments.

Table 1. A set of BLAST programs and their applications

Program	Database	Query Content	Application
Blastn	Nucleotide	Nucleotide sequence vs. Nucleotide database	May find matching sequences with distantly related connections
Blastp	Protein	Amino acid sequence vs. Protein database	Suitable for finding matches with high scores
Blastx	Protein	Nucleotide sequence (6-frame translation) vs. Protein database	Suitable for analysis of new DNA sequences and EST sequences
tblastn	Nucleotide	Protein sequence vs. Nucleotide database (6-frame translation)	Suitable for finding new and unannotated genomic sequences
tblastx	Nucleotide	Nucleotide sequence (6-frame translation) vs. Nucleotide database (6-frame translation)	Suitable for EST sequence comparisons

Genes are composed of various components, such as exons, introns, promoters, and enhancers. A full understanding of these components helps in further predicting the encoded protein and studying its function. In-depth analysis of promoters helps understand the regulatory mechanisms of gene expression. Information on protein molecular weight, isoelectric point, secondary structure, tertiary structure, quaternary structure, transmembrane segments of membrane proteins, enzyme active sites, and protein-protein interactions can be obtained from sequence information using bioinformatics tools. Of course, various prediction methods are based on existing biological data and known biological knowledge. However, different analysis programs built on different models and algorithms have their specific applicable scopes and corresponding limitations. Therefore, it is best to use several different analysis programs for the same biological question and synthesize the results obtained by various methods. Due to the accelerated pace of knowledge updates, such as webpage refreshes, replacements, and the continuous establishment of new databases, we must constantly learn to search for bioinformatics methods suitable for our own specialty and practical conditions.

3.2 Protein Identification From Experimental Data

As a core technology in proteomics research, two-dimensional electrophoresis (2-DE) separates proteins based on their molecular weight and isoelectric point. Under strictly controlled experimental conditions, the protein spots on a 2-DE gel are sufficient for identifying some proteins. SWISS-2DPAGE provides many standard gel images and tools for detecting shifts of proteins on gels. By comparing with known protein maps of cells or tissues, some marker proteins can be identified. However, in practice, due to the inherent variability of protein samples, the non-repeatability of sample preparation, and the inability of 2-DE to separate all proteins in a sample, it is difficult to identify proteins using this method alone.

Mass spectrometry can obtain highly accurate peptide mass fingerprints or amino acid sequence tags, which, combined with a series of databases and analysis tools, can be used to identify proteins. Functional prediction of a protein starts with its sequence. If the protein is known, its function can be determined through sequence databases and major literature. If the protein is unknown, clues about its function need to be obtained by analyzing related proteins. BLAST analysis yields a list of closely related proteins, some of which are highly annotated. If there is a full-length sequence match, similar function can be predicted. If only short fragment sequences match, it indicates that these fragments are conserved, possibly hinting at some functions of the protein. The realization of these analytical functions can also be achieved by integrating tools into a platform, enabling rapid high-throughput analysis, similar to EST analysis.

Proteomics will ultimately generate data far exceeding that stored in DNA sequence databases. A complete

catalog of human and model organism proteins and the annotation of protein functions will drive the rapid development of proteomics. The goal of proteomics is a comprehensive understanding of proteins. However, because the expression, modification, and interaction with other proteins of proteins differ in each cell type, proteomic analysis is much more complex than genomic analysis. The exploration of the proteome will always depend on new data processing resources.

3.3 Processing and Analysis of Gene Chip Data

Bioinformatics analysis of biochip data has become an important field for studying the molecular mechanisms of clinical diseases, disease diagnosis, and drug screening through gene expression (mRNA and protein) research. Gene chips use the hybridization of cDNA from biological samples to probes on the chip to obtain high-density hybridization array images. Image analysis software is used to extract fluorescence signal data such as the absorbance value and absorbance ratio of each hybridization spot, converting them into a gene expression matrix for quantitative analysis. A gene expression matrix is a matrix used to describe gene expression data from gene chips, typically with rows representing genes and columns representing different samples (e.g., different tissue specimens, treatments, or developmental stages). Therefore, each cell represents the expression level of a gene in a specific sample. Establishing the gene expression matrix is the foundation for gene chip data analysis. First, image processing and analysis software automatically identifies the original hybridization signal spots, extracts the signal intensity (e.g., absorbance) of each hybridization spot, and subtracts the background signal level to obtain the net signal value of the hybridization spot. Due to systematic errors in chip experiments, such as sample differences, fluorescent labeling efficiency, and detection rates, the raw signals need to be normalized before data analysis (Leung, & Cavalleri, 2003). Common normalization methods involve adjusting normalization coefficients so that the average ratio is 1, such as the housekeeping gene method, total intensity normalization method, and density-dependent method (Wang, & Zhang, 2003), among others.

The purpose of analyzing the gene expression matrix is to explore underlying biological information and patterns. There are two basic types of methods. The first is differential analysis. Currently, the ratio value (Cy3/Cy5 ratio) is often used. A ratio between 0.5 and 2 typically indicates no significant expression difference between samples. However, due to variation when analyzing multiple replicates of the same sample, significance tests such as the t-test or analysis of variance can be applied. The other type of method is cluster analysis. Cluster analysis groups genes or samples with similar or identical properties into the same cluster, while those with larger differences are grouped into another cluster. The principle of using cluster analysis for gene chip data is based on the hypothesis that genes with similar expression profiles have similar biological functions. Therefore, the function of unknown genes can be inferred from the functions of known genes with similar expression profiles. Two common strategies are supervised and unsupervised analysis (Brown, Grundy, Lin, *et al.*, 2000). Supervised analysis uses known information, such as functional classification or disease status parameters, to establish classification criteria for predicting the function of unknown genes. Supervised learning techniques include Support Vector Machines (SVM), logistic regression, neural networks, and Linear Discriminant Analysis (LDA). Unsupervised analysis, on the other hand, involves grouping expression profiles to find related genes or samples with correlations. Technically, correlation coefficients or Euclidean distance are commonly used as similarity measures to compare relationships between data points. Common methods include hierarchical clustering, self-organizing maps (SOM), and K-means clustering. Clustering can analyze functionally related or co-expressed genes under different conditions and has great application value in disease classification, diagnosis, efficacy, and prognosis assessment. However, because the underlying assumptions are not perfect, new methods for analyzing gene chip data need to be continuously explored.

4. Concluding Remarks

The above only briefly outlines the three most frequently encountered bioinformatics issues in current medical scientific experimental research. An important future direction for bioinformatics is to explore various functions by integrating multifaceted information and adopting a more holistic understanding of complex biological systems. For the realization of physiological functions, proteins need to function within interrelated networks. Therefore, we must consider not only information from modules like metabolic pathways and signal transduction systems but also the integrated information formed by linking these modules into networks. Regarding diseases, most diseases are determined by multiple genes, and the interaction of these genes with the environment forms the disease phenotype. Therefore, the relationship between an individual's genotype and disease phenotype is extremely complex. Deciphering this complex relationship requires the assistance of continuously improving bioinformatics tools. Differences in human susceptibility to diseases and responses to drugs are related to SNPs. The mapping of the human genome haplotype map is an important step towards the application of the Human Genome Project, but elucidating the relationship between SNP data and diseases relies on the application of

bioinformatics techniques. Nowadays, as a molecular biologist, it is almost impossible to be competent without possessing some basic bioinformatics skills; similarly, as a medical science researcher, it is impossible to enter the forefront of molecular medicine without mastering basic bioinformatics knowledge. Every laboratory technique, from simple cloning and PCR to gene expression analysis, requires computer-based data processing. Although creating algorithms and programming software may not be the primary task for medical science researchers, understanding the basic principles of DNA and protein analysis tools is necessary. Enhancing communication, exchange, and cooperation between medical professionals and experts from different disciplines can foster the development of tools for understanding complex problems, thereby promoting the advancement of medicine.

References

Attwood, T. K. (2000). Genomics: The Babel of Bioinformatics. *Science*, *290*(5491), 471-473. https://doi.org/10.1126/science.290.5491.471

Brown, M. P., Grundy, W. N., Lin, D., *et al.*. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA*, *97*(1), 262-267. https://doi.org/10.1073/pnas.97.1.262

Gogarten, J. P., & Olendzenski, L. (1999). Orthologs, paralogs and genome comparisons. *Curr Opin Genet Dev*, *9*, 630-636. https://doi.org/10.1016/S0959-437X(99)00029-5

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biology*, *2*(8), interactions1002.1-1002.3. https://doi.org/10.1186/gb-2001-2-8-interactions1002

Leung, Y. F., & Cavalleri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends in Genetics*, *19*(11), 649-659. https://doi.org/10.1016/j.tig.2003.09.015

Wang, Y.-Y., & Zhang, Y.-Y. (2003). Analysis and treatment of gene chip data. *Progress in Biochemistry and Biophysics*, *30*(2), 321-323.

Copyrights