Shree H. N. Shukla Institute of Pharmaceutical Education and Research, Rajkot

B. Pharm
Semester-2

STUDY MATERIAL

Subject Name: computer application in pharmacy
Subject Code: BP204TP

CHAPTER 4 : Bioinformatics:

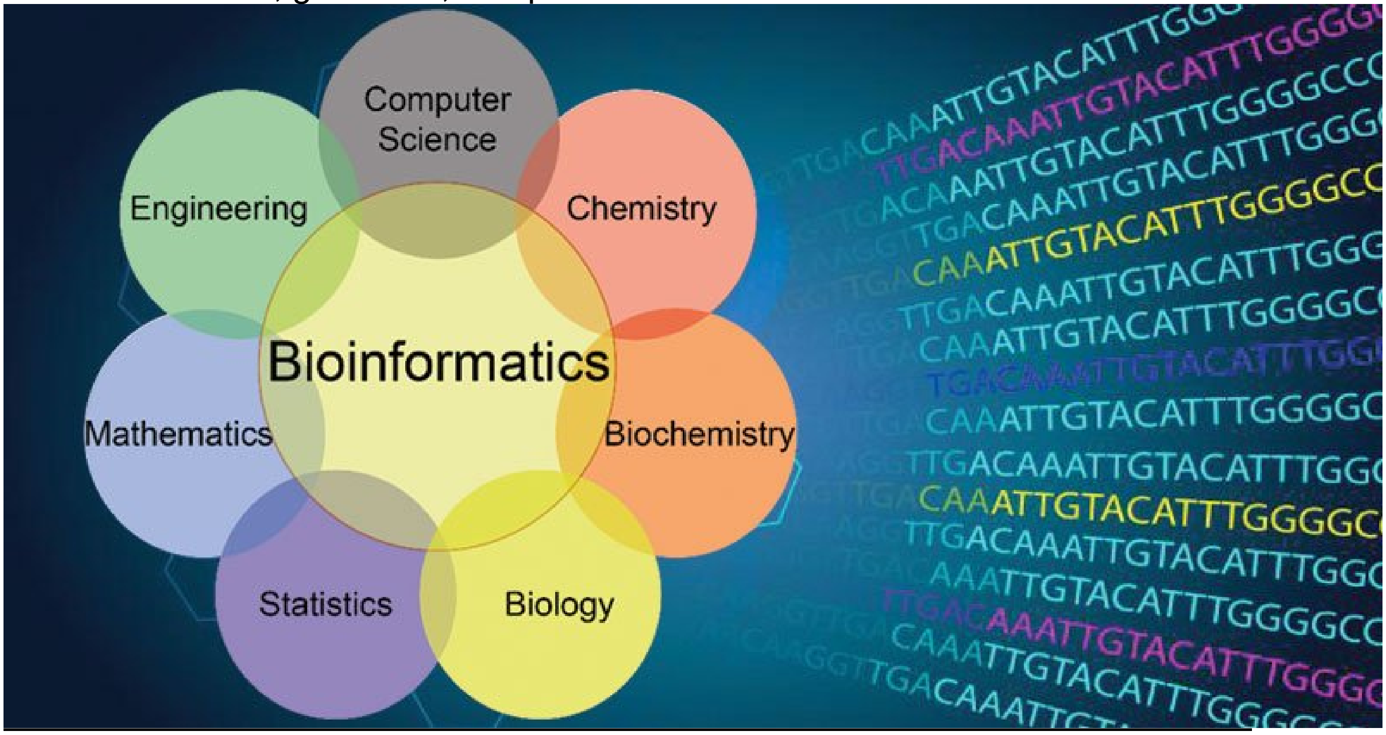
Introduction, Objective of Bioinformatics, Bioinformatics Databases, Concept of Bioinformatics, Impact of Bioinformatics in Vaccine Discovery

Introduction

Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. Data intensive, large-scale biological problems are addressed from a computational point of view. The most common problems are modeling biological processes at the molecular level and making inferences from collected data. A bioinformatics solution usually involves the following steps: Collect statistics from biological data. Build a computational model. Solve a computational modeling problem. Test and evaluate a computational algorithm. This chapter gives a brief introduction to bioinformatics by first providing an introduction to biological terminology and then discussing some classical bioinformatics problems organized by the types of data sources. Sequence analysis is the analysis of DNA and protein sequences for clues regarding function and includes subproblems such as identification of homologs, multiple sequence alignment, searching sequence patterns, and evolutionary analyses. Protein structures are three-dimensional data and the associated problems are structure prediction (secondary and tertiary), analysis of protein structures for clues regarding function, and structural alignment. Gene expression data is usually represented as matrices and analysis of microarray data mostly involves statistics analysis, classification, and clustering approaches. Biological networks such as gene regulatory networks, metabolic pathways, and protein-protein interaction networks are usually modeled as graphs and graph theoretic approaches are used to solve associated problems such as construction and analysis of large-scale networks.

- With a large number of prokaryotic and eukaryotic genomes completely sequenced and more forthcoming, access to the genomic information and synthesizing it for the discovery of new knowledge have become central themes of modern biological research.
- Mining the genomic information requires the use of sophisticated computational tools.
- It therefore becomes imperative for the new generation of biologists to initiate and familiarize with a field of study that is concerned with the careful storage, organization and indexing of information in order to tackle the new challenges in the genomic era.
- Information science has been applied to biology to produce a field is called bioinformatics.

- It is concerned with the state of- the-art computational tools available to solve biological research problems.
- The term bioinformatics was coined by Paulien Hogeweg and Ben Hesper to describe “the study of informatic processes in biotic systems” and it found early use when the first biological sequence data began to be shared.
- **Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.**
- The development of bioinformatics as a field is the result of advances in both molecular biology and computer science over the past 30–40 years.
- As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyze and interpret biological data.
- The key areas of bioinformatics include biological databases, sequence alignment, gene and promoter prediction, molecular phylogenetics, structural bioinformatics, genomics, and proteomics.



Bioinformatics vs Computational Biology

- Bioinformatics differs from a related field known as computational biology.
- Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered computational molecular biology.
- However, computational biology encompasses all biological areas that involve computation.

- Bioinformatics as the development and application of computational tools in managing all kinds of biological data, whereas computational biology is more confined to the theoretical development of algorithms used for bioinformatics.

Applications of Bioinformatics

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. The main uses of bioinformatics include:

- Bioinformatics plays a vital role in the areas of structural genomics, functional genomics, and nutritional genomics.
- It covers emerging scientific research and the exploration of proteomes from the overall level of intracellular protein composition (protein profiles), protein structure, protein-protein interaction, and unique activity patterns (e.g. post-translational modifications).
- Bioinformatics is used for transcriptome analysis where mRNA expression levels can be determined.
- Bioinformatics is used to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically.
- Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.
- Bioinformatics is playing an increasingly important role in almost all aspects of drug discovery and drug development.
- Bioinformatics tools are very effective in prediction, analysis and interpretation of clinical and preclinical findings.

Applications in Other Fields

Its major applications include in the following fields:

Molecular medicine

- The human genome will have profound effects on the fields of biomedical research and clinical medicine.
- The completion of the human genome and the use of bioinformatic tools means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised medicine

- Clinical medicine will become more personalised with the development of the field of pharmacogenomics.
- This is the study of how an individual’s genetic inheritence affects the body’s response to drugs.
- Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment.
- In the future, doctors will be able to analyse a patient’s genetic profile and prescribe the best available drug therapy and dosage from the beginning.

Preventative medicine

- With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality.

Gene therapy

- In the not too distant future with the use of bioinformatics tool, the potential for using genes themselves to treat disease may become a reality.
- Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a person's genes.

Drug development

- At present all drugs on the market target only about 500 proteins.
- With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed.
- These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial genome applications

- The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications.
- For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction.
- By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

Waste cleanup

- *Deinococcus radiodurans* is known as the world's toughest bacteria and it is the most radiation resistant organism known.
- Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

Climate change Studies

- Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change.
- Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels.
- One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

Alternative energy sources

- Scientists are studying the genome of the microbe *Chlorobium tepidum* which has an unusual capacity for generating energy from light

Biotechnology

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.
- These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes
- Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine.
- The substance is employed as a source of protein in animal nutrition.
- Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal.
- *Lactococcus lactis* is one of the most important micro-organisms involved in the dairy industry.
- Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of *lactis* to serve as a vehicle for delivering drugs.

Antibiotic resistance

- Scientists have been examining the genome of *Enterococcus faecalis*-a leading cause of bacterial infection among hospital patients.
- They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium’s transformation from a harmless gut bacteria to a menacing invader.
- The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

Forensic analysis of microbes

- Scientists used their genomic tools to help distinguish between the strain of *Bacillus anthracis* that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

The reality of bioweapon creation

- Scientists have recently built the virus poliomyelitis using entirely artificial means.
- They did this using genomic data available on the Internet and materials from a mail-order chemical supply.

- The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons.
- The researchers also hope their work will discourage officials from ever relaxing programs of immunisation.
- This project has been met with very mixed feelings.

Evolutionary studies

- The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

Crop improvement

- Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed.
- These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops.
- At present the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.

Insect resistance

- Genes from *Bacillus thuringiensis* that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes.
- This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

Improve nutritional quality

- Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients.
- This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively.
- Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

Development of Drought resistance varieties

- Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities.
- These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base.
- Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

Veterinary Science

- Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

Comparative Studies

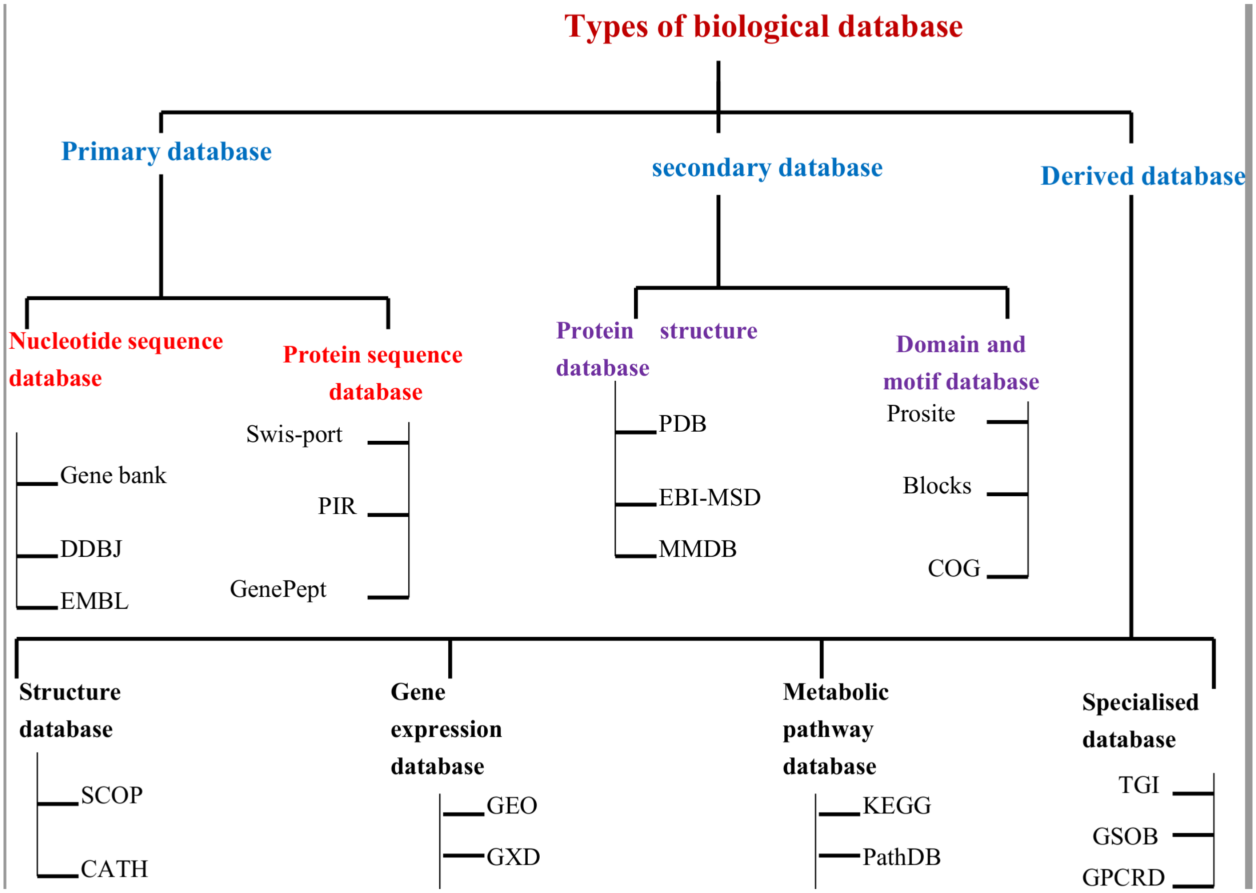
- Analysing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution.
- Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organisms.

Biological databases

• Biological data are complex, exception-ridden, vast and incomplete. Therefore several databases has been created and interpreted to ensure unambiguous results. A collection of biological data arranged in computer readable form that enhances the speed of search and retrieval and convenient to use is called biological database. A good database must have updated information.

Importance of biological database

• A range of information like biological sequences, structures, binding sites, metabolic interactions, molecular action, functional relationships, protein families, motifs and homologous can be retrieved by using biological databases. The main purpose of a biological database is to store and manage biological data and information in computer readable forms.



Primary database vs. secondary database

- A primary database contains only sequence or structural information.

- The database derived from the analysis or treatment of primary data are secondary database. It is very important for interfering protein function.

Examples of some primary biological database

GeneBank

One of the fastest growing repositories of known nucleotide sequences, GeneBank (Genetic Sequence Databank), has a flat file structure. It is an ASCII text file, readable by both humans and computers. Besides sequence data, GeneBank files contain information such as accession numbers and gene names, phylogenetic classification and references to published literature.

This database has been developed and maintained at the NCBI, Bethesda, MD, USA, as a part of International Sequence Database Collaboration (INSDC).

It is an open access sequence database.

It coordinates with individual laboratories and other sequence databases like EMBL and DDBJ.

It is an annotated collection of all nucleotide sequences that are available to the public.

The nucleotide database was divided into three databases at NCBI:

CoreNucleotide database, Expressed Sequence Tag (EST) and Genome Survey Sequence (GSS).

CoreNucleotide database has most of the nucleotide sequences used. It also encloses all nucleotide records that are not in the EST and GSS databases.

Submission of sequences to GeneBank can be done using BankIt, Sequin and tbl2asn tools.

EMBL(European Molecular Biology Laboratory)

- A comprehensive database of DNA and RNA sequences, EMBL nucleotide sequence database is collected from scientific literature, patient offices and is directly submitted by researchers. EMBL has been prepared in collaboration with GeneBank (USA) and the DNA Database of Japan (DDBJ).
- It is established in 1980.
- It is maintained by EBI (European Bioinformatics Institute)

Swiss-Port

This is a curated protein sequence database that offers a high level of integration with other databases and also has a very low level of redundancy. Swiss-Port strives to provide protein sequences with a high level of annotation (for instance, the description of protein function, domain structure and post translational modifications, etc.).

It is established in 1986 and maintained collaboratively , since 1987, by the department of Medical Biochemistry of the University of Geneva and the EMBL data Library.

TrEMBL is a computer–annotated supplement of Swiss-Port that contains all translations of EMBL nucleotide sequence entries, which is not yet integrated in Swiss-Port.

Currently Swiss-Port have 0.5 and TrEMBL have 7.6 milliom sequences.

Protein Information Resource(PIR)

• PIR is an integrated public bioinformatics resource to support genomic and proteomic research and scientific studies. Nowadays, PIR offers a wide variety of resources mainly oriented to assisting the propagation and consistency of protein annotations like PIRSF, ProClass and ProLINK.

Examples of Some Secondary Biological Database

Motif Databases

• Protein sequence motif is a set of conserved amino acid residues that are important for protein function and are located within a certain distance from one another. These motifs usually provide clues to the functions of otherwise uncharacterised proteins.

• The PROSITE database consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them.

• PRINT is a database for protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

Domain Database

• A protein domain is an independently folded, structurally compact unit that forms a steady three- dimensional structure and shows a certain level of evolutionary conservation. Typically , a conserved domain contains one or more motifs.

• ProDom is a protein domain database automatically generated from the Swiss-Port and TrEMBL sequence database.

• SMART is a highly reliable and sensitive tool for domain identification.

• COG is a database and a convenient tool for motif and domain identification. 3D Structure databases

3D Structure databases

• PDB (Protein Data bank) is the main primary database for 3D structures of biological macromolecules determined by X-ray, crystallography and NMR. It also accepts experimental data used to determine the structures and homology models.

• SCOP (Structural Classification of Protein database) classifies protein 3D structures in a hierarchical scheme of structure classes. All the protein structures in PDB are classified her, and the updated new structures are deposited in PDB.

- The CATH database (Class, Architecture, Topology, Homologous) contains a hierarchical classification of protein domain structure.

The Impact of Bioinformatics on Vaccine Design

Vaccines are the pharmaceutical products that offer the best cost-benefit ratio in the prevention or treatment of diseases. In that a vaccine is a pharmaceutical product, vaccine development and production are costly and it takes years for this to be accomplished. Several approaches have been applied to reduce the times and costs of vaccine development, mainly focusing on the selection of appropriate antigens or antigenic structures, carriers, and adjuvants. One of these approaches is the incorporation of bioinformatics methods and analyses into vaccine development. This chapter provides an overview of the application of bioinformatics strategies in vaccine design and development, supplying some successful examples of vaccines in which bioinformatics has furnished a cutting edge in their development. Reverse vaccinology, immunoinformatics, and structural vaccinology are described and addressed in the design and development of specific vaccines against infectious diseases caused by bacteria, viruses, and parasites. These include some emerging or re-emerging infectious diseases, as well as therapeutic vaccines to fight cancer, allergies, and substance abuse, which have been facilitated and improved by using bioinformatics tools or which are under development based on bioinformatics strategies.

Reverse vaccinology

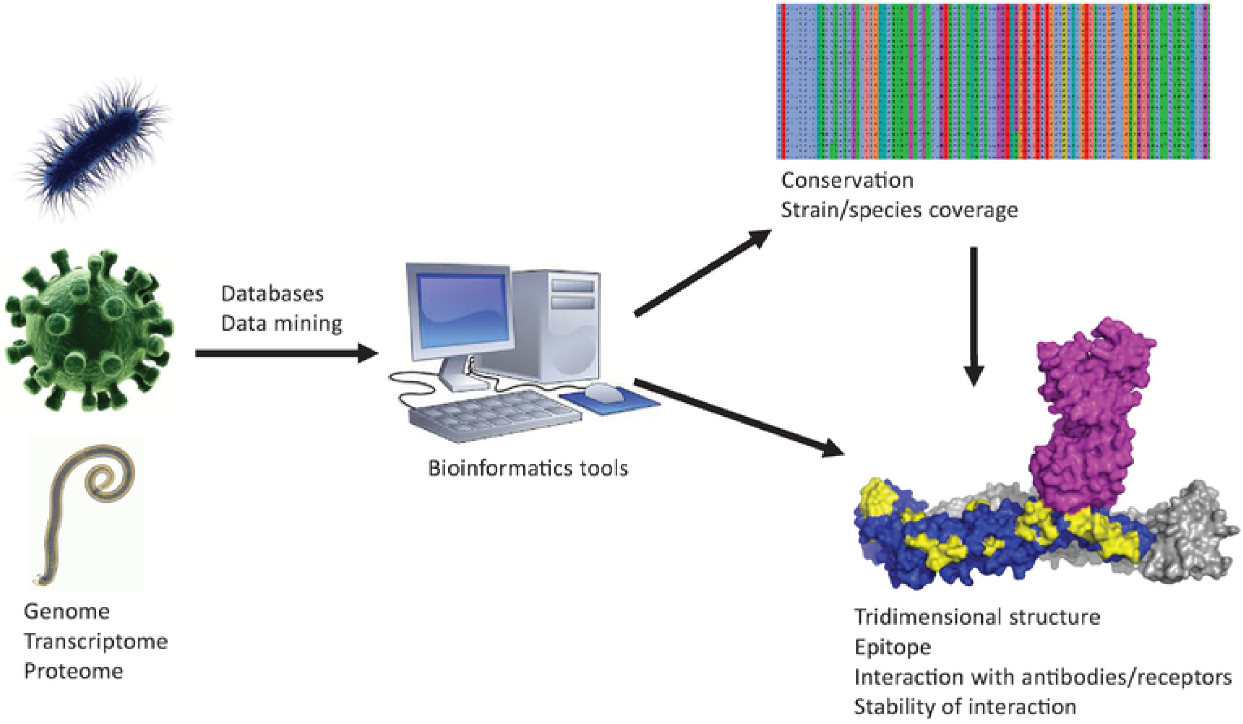
Reverse vaccinology is a methodology that uses bioinformatics tools for the identification of structures from bacteria, virus, parasites, cancer cells, or allergens that could induce an immune response capable of protecting against a specific disease [7].

This approach possesses many advantages over traditional vaccinology: it reduces time and cost in vaccine development; refines the number of proteins to be studied, facilitating the selection process; can identify antigens present in small amounts or expressed only at certain stages, which would hinder or prevent their purification; and allows for the study of noncultivable or risky microorganisms [3]

An important requirement for utilizing this methodology is the availability of genomic information of the pathogen under study and, in some instances, even the human or animal cell genome must be known (i.e., DNA vaccines and therapeutic vaccines). Once the genome sequence is obtained, it is possible to identify all likely proteins that could be expressed. For this purpose, several software systems and programs identify all open reading frames (ORFs) that constitute the sequences expressing the majority of proteins [8–10].

The next step in reverse vaccinology is to determine several antigenic and physicochemical properties that have been associated with good antigens. These characteristics must be analyzed for each protein in the proteome under study, employing different bioinformatics approaches to select the protein(s) with the best properties for testing through in vitro and in vivo assays, in order to demonstrate its safety and immunogenicity. With the best vaccine *candidates, different types of vaccines can be designed and developed, for example: subunit, recombinant, and nucleic acid vaccines [11].*

The first application of reverse vaccinology was to study Neisseria meningitidis to obtain a new subunit vaccine based on the genome study of this microorganism by means of bioinformatics tools [12]. Thereafter, this technology has been used to study pathogenic agents including eukaryotic organisms and those involved in diseases transmitted by vectors [13], to design and obtain not only vaccines for humans but also for animals [5]. The majority of new vaccines against infectious diseases that have been developed with this technology are currently found in preclinical or clinical trial. However, it is important to mention that in some instances, the vaccine candidate obtained by this technology could fail as a good vaccine antigen, because it is identified based solely on computational probabilistic studies, and there are other factors that could interfere when this antigen is administered in a complete organism. In addition, vaccine candidates identified by this technology are restricted to proteins or lipoproteins, in that they are encoded in the genome. By reverse vaccinology, it is impossible to identify carbohydrate or lipid antigenic molecules [3, 14].


Conservation
Strain/species coverage
Databases
Data mining
Bioinformatics tools
Genome
Transcriptome
Proteome
Tridimensional structure
Epitope
Interaction with antibodies/receptors
Stability of interaction