Bioinformatics approaches for Biomedical Research

Kasandra Aguilar-Cázarez\textsuperscript{a1}, Ernesto Andrade-Collantes\textsuperscript{b2}, Marisol Verdugo-Meza\textsuperscript{b3}, Claudia María de-la-Rocha-Morales\textsuperscript{b4}, Cruz Fernando López-Carrera\textsuperscript{c5}, Paúl Alexis López-Durán\textsuperscript{d6}\textsuperscript{*}

\textsuperscript{a}Centro de Investigación de Estudios Avanzados, Ciudad de México, México.  
\textsuperscript{b}Universidad Autónoma de Sinaloa, Facultad de Ciencias Químico Biológicas, Sinaloa, México.  
\textsuperscript{c}Instituto Politécnico Nacional, Escuela Nacional de Ciencias Biológicas, Ciudad de México, México.  
\textsuperscript{d}Universidad Anáhuac México, Campus Norte, Facultad de Ciencias de la Salud, Estado de México, México.

ID ORCID:  
\textsuperscript{1}https://orcid.org/0000-0002-5740-7296, \textsuperscript{2}https://orcid.org/0000-0002-7924-0132, \textsuperscript{3}https://orcid.org/0000-0002-8300-049X,  
\textsuperscript{4}https://orcid.org/0000-0002-4505-5393, \textsuperscript{5}https://orcid.org/0000-0003-2643-7713, \textsuperscript{6}https://orcid.org/0000-0002-2577-6599.

https://doi.org/10.36105/psrua.2022v2n3.04

ABSTRACT

An enormous amount of data is generated and compiled in several databases every year. Along with this, comes a demand for the analysis and interpretation of the entirety of this biological information. Taking care of this task, bioinformatics promises breakthroughs in research and development in complex biomedical areas. In just a few years since its beginning, bioinformatics has led to great progress and demonstrated its potential. It has created an opportunity to solve arising medical and molecular issues faster and more efficiently, as compared to the traditional approach. The present review aims to present some of the main applications of bioinformatics in the field of biomedicine, such as comparative genomics, biomarker identification, computer-aided drug design, vaccine design, and personalized medicine. In addition, we also cover some of its steadily reduced limitations.

Key words: bioinformatics; biomedicine; comparative genomics; biomarkers; computer-aided drug design; vaccine design; personalized medicine.

RESUMEN

Cada año se genera una enorme cantidad de datos recopilados en bases de datos. Junto con esto surge una demanda de análisis e interpretación de la totalidad de la información biológica; esta importante tarea es atendida por la bioinformática, la cual promete grandes avances en la investigación y desarrollo de áreas biomédicas complejas. En tan sólo unos años desde su inicio ha llevado a un gran progreso y ha demostrado su potencial, creando una oportunidad para resolver los problemas médicos y moleculares de manera más rápida y eficiente en comparación con el enfoque tradicional. El presente artículo pretende discutir algunas de las principales aplicaciones de la bioinformática en el campo de la biomedicina, tales como, genómica comparativa, identificación de biomarcadores, diseño de fármacos, diseño de vacunas y medicina personalizada. De manera adicional, se enlistan algunas de sus principales limitaciones, las cuales están en constante disminución.

Palabras clave: bioinformática; biomedicina; genómica comparativa; biomarcadores; diseño computarizado de fármacos; diseño de vacunas; medicina personalizada.

* Corresponding author: Paúl Alexis López-Durán. Universidad Anáhuac México, Campus Norte, Facultad de Ciencias de la Salud. Address: Av. Universidad Anáhuac núm. 46, Lomas Anáhuac, C.P. 52786, Huixquilucan, Estado de México, México. Tel.: +52 55 5627 0210. Email: paullopezdu@anahuac.mx

Received: 19 November 2021.  
Accepted: 14 February 2022.
INTRODUCTION

With the increasing development of next-generation sequencing (NGS), large-scale high-throughput molecular profiling studies and omics sciences, we now have a better understanding of the complex nature of numerous illnesses.\(^1\) Each year, large amounts of data are generated and compiled in freely accessible databases\(^2\), bringing along a high demand for their analysis and interpretation. This is managed by evolving bioinformatics.

Bioinformatics is defined as an interdisciplinary science that applies computational tools and analysis to the capture and interpretation of biological data, which brings together computer science, mathematics, physics, and biology.\(^3\) It requires research, development, and application of informatics tools and approaches to acquire, store, visualize, and interpret biological and medical data.\(^4\) It allows to test hypotheses through *in silico* methods to have a better knowledge before proceeding with expensive studies. In addition, bioinformatics provides results that are more accurate when assembling reliable interpretations. Therefore, it promises breakthroughs in research and development in complex biomedical systems as well as public health, drug design, comparative genomics, and personalized medicine, among others.\(^5\)

From the first use of molecular sequences for evolutionary studies by Zuckerkandl and Pauling (1965)\(^4\) to the massive up-to-date sequencing programs, bioinformatics has evolved just as much as its matter of study.\(^6\) It has branched out into bioinformatics tools applicable in genomics, proteomics, drug design, and simulations of molecular dynamics. Combined, they provide a more complete understanding of how diseases work.\(^7\) As an emerging field, it is essential for managing data in modern biology and medicine; for instance, it is a major protagonist in the development of all vaccines against COVID-19.\(^8\) (Fig. 1).

In the present review, the main points on topics of novel interest in the area of applied bioinformatics in biomedicine are discussed.

---

**Figure 1.** Applications of bioinformatics in biomedicine. The general workflow starting from the patient is summarized for each bioinformatics application: Comparative genomics, biomarker identification, Computer-aided drug design, and vaccine design. SNPs: single nucleotide polymorphisms; CNV: copy number variation; DEGs: differentially expressed genes; PPI: protein-protein interaction; QSAR: quantitative structure–activity relationship.
Comparative genomics

Since the sequencing of the human genome, scientists have developed a special interest in the field of biological research called comparative genomics (CG). CG can be defined as the comparison of biological information derived from whole genome sequences. It aims to determine similarities and differences between genomes to provide information about the biology of the respective organisms, describing the structure and identifying coding and non-coding regions of the genome. The comparison of DNA sequences can be applied in several areas. This information reveals the molecular basis of individuality, uncovers novel regulatory mechanisms, predicts the metabolic capabilities of an organism, and enables the prediction of gene function, among others. Assuming that you have a complete genome, the steps in CG start with genome annotation; that is, gene finding and function assignment. The process of gene search predicts the section of the genome that contains genes while the function assignment seeks to predict the function of the coded proteins. This process is performed by automated software algorithms, called pipelines, using resources from databases and biological data. Annotation allows for clustering genes in homolog or ortholog families. Diverse applications can be inferred with comparative genomics; in definition, unique signatures can be detected by forensic microbiology. For instance, virulence and antibiotic resistance genes can be identified as targets for genetic manipulation in public health. Sequence comparison can be used to identify specific types of parasites for treatment or diagnosis. Nowadays, the comparison of SARS-COV-2 distribution, spread, and evolution uses CG for multiple purposes. It can also be used in the advanced molecular characterization of infectious agents in clinical environments. This could improve both the identification and genetic characterization including resistance profiles, promoting outbreak investigations and molecular surveillance. For instance, the comparison of metagenomes could help to understand the dynamics of a given microbiota, disease versus healthy states (e.g., cancer tumor versus normal tissue), different diets, and different geographical locations.

Biomarker identification

Biomarkers can be defined as “an indicator of normal biological processes, pathogenic processes or pharmacological response to a therapeutic intervention”, which may have a monitoring, diagnostic, or predictive value. They can be single nucleotide polymorphisms (SNPs), structural variants, circulating DNAs, methylated DNAs, mRNAs, microRNAs (miRNAs), proteins, and metabolites.

In genomics, there are two approaches, the search for unique genetic variants that can act as predictors of disease and the composite predictor in the form of a genetic risk score. On their behalf, the RNA and protein profiling strategies have an important biological impact. Despite current practical limitations, transcriptomic assays together with proteomics can provide essential biological information. This kind of analysis is performed thanks to high-throughput technologies, including RNA sequencing and microarrays. Bioinformatics analysis is a valuable tool with great potential, especially when there is a discrepancy between genomic alterations and gene expression. Statistical tools are used to differentiate signal and noise in the output of gene expression. Gene expression meta-analysis is increasingly used in many fields to improve the reproducibility of a study and discover new robust biomarkers. The necessary data are frequently obtained from the gene expression omnibus (GEO) database. Additionally, metabolomics, the profiling of metabolites in biofluids, cells and tissues, is routinely applied as a tool for biomarker discovery. A metabolome-based strategy for identifying candidates with biological activity would consist of statistically analyze a list of metabolites generated from databases. Metabolic pathways and bioinformatics analysis of interaction networks can be used to reduce data complexity. The most important step in determining the activity is to use an appropriate screening strategy. This includes gene expression, protein expression, and protein activity, the modulation of desired cellular phenotypes.

The protein/gene-protein/gene interaction networking, hub gene identification, gene enrichment, and functional gene annotation analyses are powerful tools for the identify potential diagnosis and treatment biomarkers in diseases such as cancer, bipolar disorder, depression, diabetes, arteriosclerosis, and others. For instance, Wan et al. in 2020 applied bioinformatics analysis to identify eight candidate genes that could be a potential prognostic marker of thyroid carcinoma based on expression analysis profiles from the GEO database. Liu et al. in 2019 identified several hub genes and key pathways associated with bipolar disorder (BD) based on a gene co-expression network analysis, which might have important clinical implications for BD treatment and diagnosis.

Computer-aided drug design

The traditional methodology for novel drug discovery is a costly, extremely risky, and time-consuming process. Indeed, 75% of the total cost of drug development corresponds to lead molecules in clinical trials that never enter the consumer clinical market. The steps of traditional drug discovery can be summarized as follows: target identification, target validation, lead identification, lead optimization, preclinical stage, and clinical trial stage.

To address the challenges of traditional drug discovery, new strategies such as computer-aided drug design (CADD) have been developed. CADD comprises approaches such as ab
initio design, toxicity profile, quantitative structure–activity relationship (QSAR), docking, molecular dynamics (MD), quantitative free-energy calculations, and homology modeling.\textsuperscript{38}

These approaches can be classified in two categories: structure-based drug design (SBDD) and ligand-based drug design (LBDD). The first depends on the availability of the 3D target structure to screen promising ligand molecules by calculating the interaction energies between the target and compound. The knowledge of the binding site in the target protein structure is used to identify and evaluate ligands based on their interactions with the residues present in the active site.\textsuperscript{39} LBDD is used when the drug target structure is unavailable. The knowledge of the molecules interacting with the drug target is applied to develop pharmacophore and 3D-QSAR models to understand the characteristics a molecule must possess to bind to the target.\textsuperscript{40}

CADD has emerged as a crucial ally in the design of therapeutic agents against COVID-19 since the beginning of the outbreak. For instance, Elfiky (2020) used homology modeling, MD, and molecular docking to report the relevance of sofosbuvir, ribavirin, galidesivir, remdesivir, and others as candidate drugs for clinical trials.\textsuperscript{41}

**Vaccine design**

Vaccine design is supported by bioinformatics, which identifies and predicts the essential components of a vaccine (target antigen, B and T epitopes, linker, and adjuvant). These could be coupled with molecular modeling, MD simulations, and protein–protein docking, among others, to help scientists predict the adequate properties of vaccine construction in a shorter time by making conscientious decisions. Vaccine design allows the analysis of pathogenic organisms that cannot be cultivated \textit{in vitro} and improves the allergen filter process for a better selection.\textsuperscript{3,42}

Reverse vaccinology (RV) and Structural vaccinology (SV), two bioinformatics procedures, result in fundamental steps to achieve an optimum vaccine. RV defines the process of antigen discovery for further vaccine development, starting from genome information.\textsuperscript{43} It involves the identification of novel antigens and the design of B- and T-cell epitopes through the study of the genome of an organism.\textsuperscript{44} It is also useful to find genes encoding proteins that could reveal adequate epitopes.\textsuperscript{45-47} RV employs software to determine antigenic and physicochemical characteristics associated with antigenic epitopes.\textsuperscript{48} On the other hand, SV uses 3D structural information to design novel and/or improved peptide-based vaccine antigens.\textsuperscript{49} It assesses the 3D molecular structure of the epitope in the antigen–antibody complex, creates MD simulations to predict and model the epitope and its interactions, tests reengineered antigen into immunoinformatics platforms (i.e., epitome), and finally assesses vaccine candidates (VC) for efficacy and safety \textit{in vivo}.\textsuperscript{50}

Diverse platforms exist with the purpose of exploring epitopes and simulating antigen-protein docking, present antibody structures, and antigen-antibody reactions, among others. Many of these pipelines have been created or refocused before the advent of the pandemic.\textsuperscript{54,51}

Recently, the procedures mentioned above have been used extensively in the face of the emergence of the COVID-19 pandemic. Due to the current context, whole genome sequences have been available since December 2019 and, therefore, were perfect candidates for vaccine design. In this sense, Grifoni et al. (2020) made predictions for effective epitopes that can facilitate a vaccine design against the SARS-COV-2 virus, especially supported by the platform IEDB and its incorporated tools as ViPR\textsuperscript{52} designed a multi-epitope peptide vaccine against COVID-19 using immunoinformatics. They used IEDB for T-cell epitope prediction and RaptorX\textsuperscript{53} for 3D structural modeling.

**Personalized medicine**

Personalized precision medicine is a new strategy where different methods, diagnoses, and therapies are applied to the peculiarities each patient presents. It also serves as prevention of emerging diseases, considering the patient’s genome, lifestyle, and environment. This may improve side effects related to medications and generate successful therapies. Thanks to the advances in bioinformatics in recent years, changes have been generated in medicine and treatments for patients. An earlier detection of diseases has been achieved, along with a better targeted therapy. In addition, genome sequencing and data analysis have played a key role in the development of efficient personalized medicine.\textsuperscript{22}

The main tools of precision medicine are multi-omics techniques and the data obtained from these, such as DNA sequences, transcriptomes, proteomes, metabolomes, epigenomes, and microbiomes. There are different bioinformatic tools for the data processing and analysis of genomic sequences, while the software to be used depends on the origin of the genetic information.\textsuperscript{54}

The new sequencing technologies and their rapid development gave rise to the so-called new P4 medicine (predictive, preventive, personalized, and participatory), with novel approaches, opportunities, and challenges.\textsuperscript{55} Firstly, because of prediction and a deeper knowledge of biology and diseases using genomic technologies, individuals prone to certain diseases can be identified even before they show symptoms. Secondly, prevention depends on early detection. People can be
guided to improve their lifestyle, due to the risk of contracting certain diseases, avoid health problems, and choose the best treatment for their disease. Thirdly, personalization is related to the knowledge of the patient’s genome and lifestyle. By understanding the biology of the disease in a better way, an adequate therapy can be prescribed, resulting in fewer unwanted side effects.

Finally, the participatory approach uses the patient’s information. Platforms and tools are created for the discussion of diseases and allow for general knowledge.55,56

Advances in bioinformatics in precision medicine improve efficiency in the way people are characterized and the way personalized medicines are made and tested to demonstrate their usefulness. With the big data obtained, there is a better understanding of the diseases of certain groups of patients, a better diagnosis, as well as a better development of specific drugs. It is possible to reduce the adverse effects of aggressive treatments and generate more information on rare and complex diseases. The holistic approach of personalized precision medicine will help to gain more information to improve people’s quality of life.56

**Limitations of bioinformatics**

Bioinformatics has proven to be a useful tool in scientific research fields. It provides an exceptional and outstanding platform and opportunity for scientists to integrate omics science57, bioinformatics tools and data, clinical research, disease-specific biomarkers58, dynamic networks, and precision medicine.59 It creates an opportunity to solve arising issues faster and more efficiently.

Nevertheless, technological biases and limitations may not have been sufficiently considered in the development of bioinformatics tools. Herein, we mention and further explain what we consider to be the main limitations bioinformatics face and how technology must move forward in order to bridge these gaps.

1. **Amounts of data**

Contemporary sequencing technologies are capable of generating as much as 600–1000 GB per run. Further analysis of raw sequences will increase the amount of data by a factor of ten to twenty in each analysis. This general development of sequencing techniques has led to an increase in the depth of sequencing, generating a critical problem of data storage.60 The rapid development of sequencing techniques has brought sequencing-based research to bear in all areas of life sciences. However, as the technology grew faster than its computational counterpart did, unforeseen amounts of sequencing data were generated, posing a real challenge for data analysis and bioinformatics in general.

2. **Dependence on reference databases**

Most bioinformatics analysis pipelines depend on sequence comparison against reference databases.61,62 This can be problematic considering the potential incompleteness of databases. The increased error rate of the current emerging long-read technologies can have a negative impact on biological interpretations. For example, errors in protein-coding regions can affect the accuracy of protein predictions.63 Recent studies apply correction methods (hybrid error correction) to show that the original error rate of 19% is the limit for perfect correction; beyond that, long reads are too error-prone to be corrected by these methods. Therefore, updated databases with better quality data and innovative bioinformatics tools are necessary to support a stable and effective analysis.64

3. **Genetic information privacy**

Genetic testing companies that offer direct-to-consumer (DTC) services are in constant growth, raising new privacy and ethical concerns. These companies collect data from increasing numbers of people, some of whom may have shared their data without knowing the consequences for themselves and their blood relatives. Privacy breaches can have serious social implications and adversely affect genomic-driven research by decreasing data collection and data sharing.65,66 Then, it is essential to ensure privacy as both a fundamental right for individuals and a strategy to support responsible data sharing.

4. **Experimental proof and lab work are still required**

Bioinformatics findings and research may provide evidence for the progress of potential biomarkers and help understand the role of certain genes and/or proteins in any given ailment. Still, these theoretical breakthroughs are still to be proven experimentally if they are used as therapeutic treatments.67-69 So, public and private investment for the generation and transfer of knowledge could support the development of high-quality translational and collaborative research. This, in turn, will prevent and help to face menacing situations, as the current COVID-19 pandemic, or older issues such as obesity and cancer.70

5. **Bioinformatics: not for everyone**

Every novel or inexperienced user of bioinformatics, from ecologists to population geneticists or cell biologists, deal with bioinformatics matters. As stated above, bioinformatics plays an increasingly central role in biological research, medicine, and other areas of human life. Then, we need sophisticated and user-friendly bioinformatics resources to accompany it. The
ability to access and study molecular sequence data should not be reserved for those with exceptional computer skills but be made available to all scientists, medical professionals, and the general population.\textsuperscript{71}

CONCLUSION

Bioinformatics is a relatively young field compared to other areas. It is rapidly growing and has a great impact in biomedicine. Despite its limitations, it has led to great progress over the years and has demonstrated its potential use for a better understanding of molecular mechanisms of diseases. This allows us to identify biomarkers, create 3D molecular models, screen drugs, predict vaccines, and achieve personalized medicine. In brief, the rapid development of bioinformatics tools, software, genomics, transcriptomics, proteomics, and metabolomics data has greatly facilitated biomedical research. It has also enhanced the understanding of the biological meaning of DNA/RNA/protein modifications, interactions within complex organism networks, and the discovery of new ways to accurately apply this knowledge, enabling new therapeutic measures. This helps us to better face diseases from an increasingly broader perspective, safer, more effective, and more specific.

REFERENCES


Acinetobacter baumannii strains in a tertiary level hospital.
https://doi.org/10.1371/journal.pone.0231829

https://doi.org/10.1111/cas.14298

https://doi.org/10.1038/s41598-020-61192-y

https://doi.org/10.1186/s40168-020-00826-9

https://doi.org/10.1067/mcp.2001.113989

https://doi.org/10.1093/rheumatology/keaa484

https://doi.org/10.1093/bib/bbx170

https://doi.org/10.2217/pgs-2018-0155

https://doi.org/10.1016/j.ctrv.2020.102019

https://doi.org/10.1093/nar/gkl212

https://doi.org/10.1093/bioinformatics/bty721

https://doi.org/10.1038/nrm.2016.25

https://doi.org/10.1038/s41580-019-01084-4

https://doi.org/10.1186/s12859-020-3524-8

http://www.spandidos-publications.com/10.3892/or.2019.7400

https://doi.org/10.7717/peerj.9120

https://doi.org/10.1016/j.neubiorev.2016.02.011

https://doi.org/10.1186/s13568-020-01009-3

https://doi.org/10.18632/aging.102200

https://doi.org/10.3389/fphys.2019.01081

https://doi.org/10.3762/bjoc.12.267

https://doi.org/10.1016/C2009-0-64537-2


44. Del Tordello E, Rappuoli R, Delany I. Reverse Vaccinology. https://doi.org/10.1016/B978-0-12-822312-3.00013-8


https://doi.org/10.1016/j.jbiotec.2017.08.012

https://doi.org/10.1080/07391102.2020.1783365

https://doi.org/10.1128/JVI.79.4.2559-2572.2005

https://doi.org/10.1038/s41587-018-0004-z

https://doi.org/10.1186/s13059-019-1885-y

https://doi.org/10.1097/gim.0b013e3182227589

https://doi.org/10.1159/000334718

https://doi.org/10.5853/jos.2020.05085

https://doi.org/10.1002/jcla.23768

https://doi.org/10.1016/j.sjbs.2021.07.068

https://doi.org/10.3390/jpm11060475

https://doi.org/10.3389/fgene.2013.00187