**文章来源**

[**http://en.wikipedia.org/wiki/Bioinformatics**](http://en.wikipedia.org/wiki/Bioinformatics)

Bioinformatics

**Bioinformatics** [/ˌbaɪ.oʊˌɪnfərˈmætɪks/](https://en.wikipedia.org/wiki/Help:IPA/English) ([About this sound](https://en.wikipedia.org/wiki/File:En-us-bioinformatics.ogg) [listen](https://upload.wikimedia.org/wikipedia/commons/7/76/En-us-bioinformatics.ogg)) is an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) field that develops methods and [software tools](https://en.wikipedia.org/wiki/Software_tool) for understanding [biological](https://en.wikipedia.org/wiki/Biology) data. As an [interdisciplinary](https://en.wikipedia.org/wiki/Interdisciplinary) field of science, bioinformatics combines [computer science](https://en.wikipedia.org/wiki/Computer_science), [statistics](https://en.wikipedia.org/wiki/Statistics), [mathematics](https://en.wikipedia.org/wiki/Mathematics), and [engineering](https://en.wikipedia.org/wiki/Engineering) to analyze and interpret [biological](https://en.wikipedia.org/wiki/Biology) data. Bioinformatics has been used for [*in silico*](https://en.wikipedia.org/wiki/In_silico) analyses of [biological](https://en.wikipedia.org/wiki/Biological) queries using [mathematical](https://en.wikipedia.org/wiki/Mathematical) and statistical techniques.

Bioinformatics is both an [umbrella term](https://en.wikipedia.org/wiki/Umbrella_term) for the body of biological studies that use [computer programming](https://en.wikipedia.org/wiki/Computer_programming) as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the field of [genomics](https://en.wikipedia.org/wiki/Genomics). Common uses of bioinformatics include the identification of candidate [genes](https://en.wikipedia.org/wiki/Gene) and single [nucleotide](https://en.wikipedia.org/wiki/Nucleotide) polymorphisms ([SNPs](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism)). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organisational principles within [nucleic acid](https://en.wikipedia.org/wiki/Nucleic_acid) and [protein](https://en.wikipedia.org/wiki/Protein) sequences, called [proteomics](https://en.wikipedia.org/wiki/Proteomics).[[1]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-1)

## Contents

* [1Introduction](https://en.wikipedia.org/wiki/Bioinformatics#Introduction)
  + [1.1History](https://en.wikipedia.org/wiki/Bioinformatics#History)
    - [1.1.1Sequences](https://en.wikipedia.org/wiki/Bioinformatics#Sequences)
  + [1.2Goals](https://en.wikipedia.org/wiki/Bioinformatics#Goals)
  + [1.3Relation to other fields](https://en.wikipedia.org/wiki/Bioinformatics#Relation_to_other_fields)
* [2Sequence analysis](https://en.wikipedia.org/wiki/Bioinformatics#Sequence_analysis)
  + [2.1DNA sequencing](https://en.wikipedia.org/wiki/Bioinformatics#DNA_sequencing)
  + [2.2Sequence assembly](https://en.wikipedia.org/wiki/Bioinformatics#Sequence_assembly)
  + [2.3Genome annotation](https://en.wikipedia.org/wiki/Bioinformatics#Genome_annotation)
  + [2.4Computational evolutionary biology](https://en.wikipedia.org/wiki/Bioinformatics#Computational_evolutionary_biology)
  + [2.5Comparative genomics](https://en.wikipedia.org/wiki/Bioinformatics#Comparative_genomics)
  + [2.6Pan genomics](https://en.wikipedia.org/wiki/Bioinformatics#Pan_genomics)
  + [2.7Genetics of disease](https://en.wikipedia.org/wiki/Bioinformatics#Genetics_of_disease)
  + [2.8Analysis of mutations in cancer](https://en.wikipedia.org/wiki/Bioinformatics#Analysis_of_mutations_in_cancer)
* [3Gene and protein expression](https://en.wikipedia.org/wiki/Bioinformatics#Gene_and_protein_expression)
  + [3.1Analysis of gene expression](https://en.wikipedia.org/wiki/Bioinformatics#Analysis_of_gene_expression)
  + [3.2Analysis of protein expression](https://en.wikipedia.org/wiki/Bioinformatics#Analysis_of_protein_expression)
  + [3.3Analysis of regulation](https://en.wikipedia.org/wiki/Bioinformatics#Analysis_of_regulation)
* [4Analysis of cellular organization](https://en.wikipedia.org/wiki/Bioinformatics#Analysis_of_cellular_organization)
  + [4.1Microscopy and image analysis](https://en.wikipedia.org/wiki/Bioinformatics#Microscopy_and_image_analysis)
  + [4.2Protein localization](https://en.wikipedia.org/wiki/Bioinformatics#Protein_localization)
  + [4.3Nuclear organisation of chromatin](https://en.wikipedia.org/wiki/Bioinformatics#Nuclear_organisation_of_chromatin)
* [5Structural bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics#Structural_bioinformatics)
* [6Network and systems biology](https://en.wikipedia.org/wiki/Bioinformatics#Network_and_systems_biology)
  + [6.1Molecular interaction networks](https://en.wikipedia.org/wiki/Bioinformatics#Molecular_interaction_networks)
* [7Others](https://en.wikipedia.org/wiki/Bioinformatics#Others)
  + [7.1Literature analysis](https://en.wikipedia.org/wiki/Bioinformatics#Literature_analysis)
  + [7.2High-throughput image analysis](https://en.wikipedia.org/wiki/Bioinformatics#High-throughput_image_analysis)
  + [7.3High-throughput single cell data analysis](https://en.wikipedia.org/wiki/Bioinformatics#High-throughput_single_cell_data_analysis)
  + [7.4Biodiversity informatics](https://en.wikipedia.org/wiki/Bioinformatics#Biodiversity_informatics)
* [8Databases](https://en.wikipedia.org/wiki/Bioinformatics#Databases)
* [9Software and tools](https://en.wikipedia.org/wiki/Bioinformatics#Software_and_tools)
  + [9.1Open-source bioinformatics software](https://en.wikipedia.org/wiki/Bioinformatics#Open-source_bioinformatics_software)
  + [9.2Web services in bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics#Web_services_in_bioinformatics)
  + [9.3Bioinformatics workflow management systems](https://en.wikipedia.org/wiki/Bioinformatics#Bioinformatics_workflow_management_systems)
* [10Education platforms](https://en.wikipedia.org/wiki/Bioinformatics#Education_platforms)
* [11Conferences](https://en.wikipedia.org/wiki/Bioinformatics#Conferences)
* [12See also](https://en.wikipedia.org/wiki/Bioinformatics#See_also)
* [13References](https://en.wikipedia.org/wiki/Bioinformatics#References)
* [14Further reading](https://en.wikipedia.org/wiki/Bioinformatics#Further_reading)
* [15External links](https://en.wikipedia.org/wiki/Bioinformatics#External_links)

## Introduction

Bioinformatics has become an important part of many areas of biology. In experimental [molecular biology](https://en.wikipedia.org/wiki/Molecular_biology), bioinformatics techniques such as [image](https://en.wikipedia.org/wiki/Image_processing) and [signal processing](https://en.wikipedia.org/wiki/Signal_processing) allow extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed [mutations](https://en.wikipedia.org/wiki/Mutation). It plays a role in the [text mining](https://en.wikipedia.org/wiki/Text_mining) of biological literature and the development of biological and gene [ontologies](https://en.wikipedia.org/wiki/Ontology_(information_science)) to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA,[[2]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-:0-2) RNA,[[2]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-:0-2)[[3]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-3) proteins[[4]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-4) as well as biomolecular interactions.[[5]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-5)[[6]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-6)[[7]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-7)

### History[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=2" \o "Edit section: History)]

Historically, the term *bioinformatics* did not mean what it means today. [Paulien Hogeweg](https://en.wikipedia.org/wiki/Paulien_Hogeweg" \o "Paulien Hogeweg) and [Ben Hesper](https://en.wikipedia.org/w/index.php?title=Ben_Hesper&action=edit&redlink=1) coined it in 1970 to refer to the study of information processes in biotic systems.[[8]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-Hogeweg2011-8)[[9]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-9)[[10]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-10) This definition placed bioinformatics as a field parallel to [biophysics](https://en.wikipedia.org/wiki/Biophysics) (the study of physical processes in biological systems) or [biochemistry](https://en.wikipedia.org/wiki/Biochemistry) (the study of chemical processes in biological systems).[[8]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-Hogeweg2011-8)

#### Sequences [[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=3)]

[https://upload.wikimedia.org/wikipedia/commons/thumb/d/d8/Image_DNA_sequence_-_png.png/220px-Image_DNA_sequence_-_png.png](https://en.wikipedia.org/wiki/File:Image_DNA_sequence_-_png.png)

Sequences of genetic material are frequently used in bioinformatics and are easier to manage using computers than manually.

Computers became essential in molecular biology when [protein sequences](https://en.wikipedia.org/wiki/Protein_sequences) became available after [Frederick Sanger](https://en.wikipedia.org/wiki/Frederick_Sanger) determined the sequence of [insulin](https://en.wikipedia.org/wiki/Insulin) in the early 1950s. Comparing multiple sequences manually turned out to be impractical. A pioneer in the field was [Margaret Oakley Dayhoff](https://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff), who has been hailed by [David Lipman](https://en.wikipedia.org/wiki/David_Lipman), director of the [National Center for Biotechnology Information](https://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information), as the "mother and father of bioinformatics."[[11]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-11) Dayhoff compiled one of the first protein sequence databases, initially published as books[[12]](https://en.wikipedia.org/wiki/Bioinformatics" \l "cite_note-12) and pioneered methods of sequence alignment and molecular evolution.[[13]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-pmid17775169-13) Another early contributor to bioinformatics was [Elvin A. Kabat](https://en.wikipedia.org/wiki/Elvin_A._Kabat), who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991.[[14]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-14)

### Goals[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=4" \o "Edit section: Goals)]

To study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and [amino acid sequences](https://en.wikipedia.org/wiki/Amino_acid_sequence), [protein domains](https://en.wikipedia.org/wiki/Protein_domain), and [protein structures](https://en.wikipedia.org/wiki/Protein_structure).[[15]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-15) The actual process of analyzing and interpreting data is referred to as [computational biology](https://en.wikipedia.org/wiki/Computational_biology). Important sub-disciplines within bioinformatics and computational biology include:

* Development and implementation of computer programs that enable efficient access to, use and management of, various types of information
* Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a [gene](https://en.wikipedia.org/wiki/Gene) within a sequence, to predict protein structure and/or function, and to [cluster](https://en.wikipedia.org/wiki/Cluster_analysis) protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [data mining](https://en.wikipedia.org/wiki/Data_mining), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms, and [visualization](https://en.wikipedia.org/wiki/Biological_Data_Visualization). Major research efforts in the field include [sequence alignment](https://en.wikipedia.org/wiki/Sequence_alignment), [gene finding](https://en.wikipedia.org/wiki/Gene_finding), [genome assembly](https://en.wikipedia.org/wiki/Genome_assembly), [drug design](https://en.wikipedia.org/wiki/Drug_design), [drug discovery](https://en.wikipedia.org/wiki/Drug_discovery), [protein structure alignment](https://en.wikipedia.org/wiki/Protein_structural_alignment), [protein structure prediction](https://en.wikipedia.org/wiki/Protein_structure_prediction), prediction of [gene expression](https://en.wikipedia.org/wiki/Gene_expression) and [protein–protein interactions](https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interactions), [genome-wide association studies](https://en.wikipedia.org/wiki/Genome-wide_association_studies), the modeling of [evolution](https://en.wikipedia.org/wiki/Evolution) and [cell division/mitosis.](https://en.wikipedia.org/wiki/Cellular_model)

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing [DNA](https://en.wikipedia.org/wiki/DNA) and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

### Relation to other fields[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=5" \o "Edit section: Relation to other fields)]

Bioinformatics is a science field that is similar to but distinct from [biological computation](https://en.wikipedia.org/wiki/Biological_computation), while it is often considered synonymous to [computational biology](https://en.wikipedia.org/wiki/Computational_biology). Biological computation uses [bioengineering](https://en.wikipedia.org/wiki/Bioengineering) and [biology](https://en.wikipedia.org/wiki/Biology) to build biological [computers](https://en.wikipedia.org/wiki/Computer), whereas bioinformatics uses computation to better understand biology. Bioinformatics and computational biology involve the analysis of biological data, particularly DNA, RNA, and protein sequences. The field of bioinformatics experienced explosive growth starting in the mid-1990s, driven largely by the [Human Genome Project](https://en.wikipedia.org/wiki/Human_Genome_Project) and by rapid advances in DNA sequencing technology.

Analyzing biological data to produce meaningful information involves writing and running software programs that use [algorithms](https://en.wikipedia.org/wiki/Algorithm) from [graph theory](https://en.wikipedia.org/wiki/Graph_theory), [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence), [soft computing](https://en.wikipedia.org/wiki/Soft_computing), [data mining](https://en.wikipedia.org/wiki/Data_mining), [image processing](https://en.wikipedia.org/wiki/Image_processing), and [computer simulation](https://en.wikipedia.org/wiki/Computer_simulation). The algorithms in turn depend on theoretical foundations such as [discrete mathematics](https://en.wikipedia.org/wiki/Discrete_mathematics), [control theory](https://en.wikipedia.org/wiki/Control_theory), [system theory](https://en.wikipedia.org/wiki/System_theory), [information theory](https://en.wikipedia.org/wiki/Information_theory), and [statistics](https://en.wikipedia.org/wiki/Statistics).

## Sequence analysis [[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=6)]



The sequences of different genes or proteins may be aligned side-by-side to measure their similarity. This alignment compares protein sequences containing [WPP domains](https://en.wikipedia.org/wiki/WPP_domain).

Since the [Phage Φ-X174](https://en.wikipedia.org/wiki/Phi_X_174) was [sequenced](https://en.wikipedia.org/wiki/Sequencing) in 1977,[[16]](https://en.wikipedia.org/wiki/Bioinformatics" \l "cite_note-pmid870828-16) the [DNA sequences](https://en.wikipedia.org/wiki/DNA_sequence) of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode [proteins](https://en.wikipedia.org/wiki/Protein), RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a [species](https://en.wikipedia.org/wiki/Species) or between different species can show similarities between protein functions, or relations between species (the use of [molecular systematics](https://en.wikipedia.org/wiki/Molecular_systematics) to construct [phylogenetic trees](https://en.wikipedia.org/wiki/Phylogenetic_tree)). With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, [computer programs](https://en.wikipedia.org/wiki/Computer_program) such as [BLAST](https://en.wikipedia.org/wiki/BLAST) are used daily to search sequences from more than 260 000 organisms, containing over 190 billion [nucleotides](https://en.wikipedia.org/wiki/Nucleotide).[[17]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-pmid18073190-17) These programs can compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, to identify sequences that are related, but not identical. A variant of this [sequence alignment](https://en.wikipedia.org/wiki/Sequence_alignment) is used in the sequencing process itself.

### DNA sequencing[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=7" \o "Edit section: DNA sequencing)]

*Main article:*[*DNA sequencing*](https://en.wikipedia.org/wiki/DNA_sequencing)

Before sequences can be analyzed they have to be obtained. [DNA sequencing](https://en.wikipedia.org/wiki/DNA_sequencing) is still a non-trivial problem as the raw data may be noisy or afflicted by weak signals. [Algorithms](https://en.wikipedia.org/wiki/Algorithm) have been developed for [base calling](https://en.wikipedia.org/wiki/Base_calling) for the various experimental approaches to DNA sequencing.

### Sequence assembly [[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=8)]

*Main article:*[*Sequence assembly*](https://en.wikipedia.org/wiki/Sequence_assembly)

Most DNA sequencing techniques produce short fragments of sequence that need to be assembled to obtain complete gene or genome sequences. The so-called [shotgun sequencing](https://en.wikipedia.org/wiki/Shotgun_sequencing) technique (which was used, for example, by [The Institute for Genomic Research](https://en.wikipedia.org/wiki/The_Institute_for_Genomic_Research) (TIGR) to sequence the first bacterial genome, *[Haemophilus influenzae](https://en.wikipedia.org/wiki/Haemophilus_influenzae" \o "Haemophilus influenzae)*)[[18]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-pmid7542800-18) generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the [human genome](https://en.wikipedia.org/wiki/Human_genome), it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly usually contains numerous gaps that must be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

*See also:*[*sequence analysis*](https://en.wikipedia.org/wiki/Sequence_analysis)*,*[*sequence mining*](https://en.wikipedia.org/wiki/Sequence_mining)*,*[*sequence profiling tool*](https://en.wikipedia.org/wiki/Sequence_profiling_tool)*, and*[*sequence motif*](https://en.wikipedia.org/wiki/Sequence_motif)

### Genome annotation[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=9" \o "Edit section: Genome annotation)]

*Main article:*[*Gene prediction*](https://en.wikipedia.org/wiki/Gene_prediction)

In the context of [genomics](https://en.wikipedia.org/wiki/Genomics), [annotation](https://en.wikipedia.org/wiki/Genome_project#Genome_annotation) is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of [sequencing](https://en.wikipedia.org/wiki/DNA_sequencing) has ceased to pose a bottleneck. Annotation is made possible by the fact that genes have recognisable start and stop regions, although the exact sequence found in these regions can vary between genes.

The first genome annotation software system was designed in 1995 by [Owen White](https://en.wikipedia.org/wiki/Owen_White), who was part of the team at [The Institute for Genomic Research](https://en.wikipedia.org/wiki/The_Institute_for_Genomic_Research) that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *[Haemophilus influenzae](https://en.wikipedia.org/wiki/Haemophilus_influenzae" \o "Haemophilus influenzae)*.[[18]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-pmid7542800-18) White built a software system to find the genes (fragments of genomic sequence that encode proteins), the transfer RNAs, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA, such as the [GeneMark](https://en.wikipedia.org/wiki/GeneMark" \o "GeneMark) program trained and used to find protein-coding genes in *[Haemophilus influenzae](https://en.wikipedia.org/wiki/Haemophilus_influenzae" \o "Haemophilus influenzae)*, are constantly changing and improving.

Following the goals that the Human Genome Project left to achieve after its closure in 2003, a new project developed by the National Human Genome Research Institute in the U.S appeared. The so-called [ENCODE](https://en.wikipedia.org/wiki/ENCODE) project is a collaborative data collection of the functional elements of the human genome that uses next-generation DNA-sequencing technologies and genomic tiling arrays, technologies able to generate automatically large amounts of data with lower research costs but with the same quality and viability.

### Computational evolutionary biology[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=10" \o "Edit section: Computational evolutionary biology)]

*Further information:*[*Computational phylogenetics*](https://en.wikipedia.org/wiki/Computational_phylogenetics)

[Evolutionary biology](https://en.wikipedia.org/wiki/Evolutionary_biology) is the study of the origin and descent of [species](https://en.wikipedia.org/wiki/Species), as well as their change over time. [Informatics](https://en.wikipedia.org/wiki/Informatics_(academic_field)) has assisted evolutionary biologists by enabling researchers to:

* trace the evolution of a large number of organisms by measuring changes in their [DNA](https://en.wikipedia.org/wiki/DNA), rather than through physical taxonomy or physiological observations alone,
* more recently, compare entire [genomes](https://en.wikipedia.org/wiki/Genomes), which permits the study of more complex evolutionary events, such as [gene duplication](https://en.wikipedia.org/wiki/Gene_duplication), [horizontal gene transfer](https://en.wikipedia.org/wiki/Horizontal_gene_transfer), and the prediction of factors important in bacterial [speciation](https://en.wikipedia.org/wiki/Speciation),
* build complex computational [population genetics](https://en.wikipedia.org/wiki/Population_genetics) models to predict the outcome of the system over time[[19]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-19)
* track and share information on an increasingly large number of species and organisms

Future work endeavours to reconstruct the now more complex [tree of life](https://en.wikipedia.org/wiki/Evolutionary_tree).

The area of research within [computer science](https://en.wikipedia.org/wiki/Computer_science) that uses [genetic algorithms](https://en.wikipedia.org/wiki/Genetic_algorithm) is sometimes confused with computational evolutionary biology, but the two areas are not necessarily related.

### Comparative genomics [[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=11)]

*Main article:*[*Comparative genomics*](https://en.wikipedia.org/wiki/Comparative_genomics)

The core of comparative genome analysis is the establishment of the correspondence between [genes](https://en.wikipedia.org/wiki/Genes) ([orthology](https://en.wikipedia.org/wiki/Homology_(biology)" \l "Orthology" \o "Homology (biology)) analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion.[[20]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-20) Ultimately, whole genomes are involved in processes of hybridization, polyploidization and [endosymbiosis](https://en.wikipedia.org/wiki/Endosymbiosis), often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, [heuristics](https://en.wikipedia.org/wiki/Heuristics), fixed parameter and [approximation algorithms](https://en.wikipedia.org/wiki/Approximation_algorithms) for problems based on parsimony models to [Markov Chain Monte Carlo](https://en.wikipedia.org/wiki/Markov_Chain_Monte_Carlo) algorithms for [Bayesian analysis](https://en.wikipedia.org/wiki/Bayesian_analysis) of problems based on probabilistic models.

Many of these studies are based on the [homology](https://en.wikipedia.org/wiki/Homology_(biology)) detection and protein families computation.[[21]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-21)

### Pan genomics[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=12" \o "Edit section: Pan genomics)]

*Main article:*[*Pan-genome*](https://en.wikipedia.org/wiki/Pan-genome)

Pan genomics is a concept introduced in 2005 by Tettelin and Medini which eventually took root in bioinformatics. Pan genome is the complete gene repertoire of a particular taxonomic group: although initially applied to closely related strains of a species, it can be applied to a larger context like genus, phylum etc. It is divided in two parts- The Core genome: Set of genes common to all the genomes under study (These are often housekeeping genes vital for survival) and The Dispensable/Flexible Genome: Set of genes not present in all but one or some genomes under study. a bioinformatics tool BPGA can be used to characterize the Pan Genome of bacterial species.[[22]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-22)

### Genetics of disease[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=13" \o "Edit section: Genetics of disease)]

*Main article:*[*Genome-wide association studies*](https://en.wikipedia.org/wiki/Genome-wide_association_studies)

With the advent of next-generation sequencing we are obtaining enough sequence data to map the genes of complex diseases such as [diabetes](https://en.wikipedia.org/wiki/Diabetes_mellitus),[[23]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-23) [infertility](https://en.wikipedia.org/wiki/Infertility),[[24]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-Demerec1945-24) [breast cancer](https://en.wikipedia.org/wiki/Breast_cancer)[[25]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-V.C3.A9ron2013-25) or [Alzheimer's Disease](https://en.wikipedia.org/wiki/Alzheimer%27s_Disease).[[26]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-Tosto2013-26) Genome-wide association studies are a useful approach to pinpoint the mutations responsible for such complex diseases.[[27]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-Londin2013-27) Through these studies, thousands of DNA variants have been identified that are associated with similar diseases and traits.[[28]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-28)Furthermore, the possibility for genes to be used at prognosis, diagnosis or treatment is one of the most essential applications. Many studies are discussing both the promising ways to choose the genes to be used and the problems and pitfalls of using genes to predict disease presence or prognosis.[[29]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-29)

### Analysis of mutations in cancer[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=14" \o "Edit section: Analysis of mutations in cancer)]

*Main article: [Oncogenomics](https://en.wikipedia.org/wiki/Oncogenomics" \o "Oncogenomics)*

In [cancer](https://en.wikipedia.org/wiki/Cancer), the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown [point mutations](https://en.wikipedia.org/wiki/Point_mutation) in a variety of [genes](https://en.wikipedia.org/wiki/Gene) in cancer. Bioinformaticians continue to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of [human genome](https://en.wikipedia.org/wiki/Human_genome) sequences and [germline](https://en.wikipedia.org/wiki/Germline) polymorphisms. New physical detection technologies are employed, such as [oligonucleotide](https://en.wikipedia.org/wiki/Oligonucleotide) microarrays to identify chromosomal gains and losses (called [comparative genomic hybridization](https://en.wikipedia.org/wiki/Comparative_genomic_hybridization)), and [single-nucleotide polymorphism](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) arrays to detect known *point mutations*. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate [terabytes](https://en.wikipedia.org/wiki/Terabyte) of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or [noise](https://en.wikipedia.org/wiki/Noise), and thus [Hidden Markov model](https://en.wikipedia.org/wiki/Hidden_Markov_model) and change-point analysis methods are being developed to infer real [copy number](https://en.wikipedia.org/wiki/Copy_number_variation) changes.

With the breakthroughs that this next-generation sequencing technology is providing to the field of Bioinformatics, cancer genomics could drastically change. These new methods and software allow bioinformaticians to sequence many cancer genomes quickly and affordably. This could create a more flexible process for classifying types of cancer by analysis of cancer driven mutations in the genome. Furthermore, tracking of patients while the disease progresses may be possible in the future with the sequence of cancer samples.[[30]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-30)

Another type of data that requires novel informatics development is the analysis of [lesions](https://en.wikipedia.org/wiki/Lesion) found to be recurrent among many tumors.

## Gene and protein expression[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=15" \o "Edit section: Gene and protein expression)]

### Analysis of gene expression[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=16" \o "Edit section: Analysis of gene expression)]

The [expression](https://en.wikipedia.org/wiki/Gene_expression) of many genes can be determined by measuring [mRNA](https://en.wikipedia.org/wiki/Messenger_RNA) levels with multiple techniques including [microarrays](https://en.wikipedia.org/wiki/DNA_microarray), [expressed cDNA sequence tag](https://en.wikipedia.org/wiki/Expressed_sequence_tag) (EST) sequencing, [serial analysis of gene expression](https://en.wikipedia.org/wiki/Serial_analysis_of_gene_expression) (SAGE) tag sequencing, [massively parallel signature sequencing](https://en.wikipedia.org/wiki/Massively_parallel_signature_sequencing) (MPSS), [RNA-Seq](https://en.wikipedia.org/wiki/RNA-Seq), also known as "Whole Transcriptome Shotgun Sequencing" (WTSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate [signal](https://en.wikipedia.org/wiki/Signal_(information_theory)) from [noise](https://en.wikipedia.org/wiki/Noise) in high-throughput gene expression studies.[[31]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-31) Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous [epithelial](https://en.wikipedia.org/wiki/Epithelial) cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

### Analysis of protein expression[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=17" \o "Edit section: Analysis of protein expression)]

[Protein microarrays](https://en.wikipedia.org/wiki/Protein_microarray) and high throughput (HT) [mass spectrometry](https://en.wikipedia.org/wiki/Mass_spectrometry) (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

### Analysis of regulation[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=18" \o "Edit section: Analysis of regulation)]

[Regulation](https://en.wikipedia.org/wiki/Regulation_of_gene_expression) is the complex orchestration of events by which a signal, potentially an extracellular signal such as a [hormone](https://en.wikipedia.org/wiki/Hormone), eventually leads to an increase or decrease in the activity of one or more [proteins](https://en.wikipedia.org/wiki/Protein). Bioinformatics techniques have been applied to explore various steps in this process.

For example, gene expression can be regulated by nearby elements in the genome. Promoter analysis involves the identification and study of [sequence motifs](https://en.wikipedia.org/wiki/Sequence_motif) in the DNA surrounding the coding region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. [Enhancer](https://en.wikipedia.org/wiki/Enhancer_(genetics)) elements far away from the promoter can also regulate gene expression, through three-dimensional looping interactions. These interactions can be determined by bioinformatic analysis of [chromosome conformation capture](https://en.wikipedia.org/wiki/Chromosome_conformation_capture) experiments.

Expression data can be used to infer gene regulation: one might compare [microarray](https://en.wikipedia.org/wiki/Microarray) data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single-cell organism, one might compare stages of the [cell cycle](https://en.wikipedia.org/wiki/Cell_cycle), along with various stress conditions (heat shock, starvation, etc.). One can then apply [clustering algorithms](https://en.wikipedia.org/wiki/Cluster_analysis) to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented [regulatory elements](https://en.wikipedia.org/wiki/Regulatory_elements). Examples of clustering algorithms applied in gene clustering are [k-means clustering](https://en.wikipedia.org/wiki/K-means_clustering), [self-organizing maps](https://en.wikipedia.org/wiki/Self-organizing_map) (SOMs), [hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering), and [consensus clustering](https://en.wikipedia.org/wiki/Consensus_clustering) methods.

## Analysis of cellular organization[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=19" \o "Edit section: Analysis of cellular organization)]

Several approaches have been developed to analyze the location of organelles, genes, proteins, and other components within cells. This is relevant as the location of these components affects the events within a cell and thus helps us to predict the behavior of biological systems. A [gene ontology](https://en.wikipedia.org/wiki/Gene_ontology) category, *cellular compartment*, has been devised to capture subcellular localization in many [biological databases](https://en.wikipedia.org/wiki/Biological_database).

### Microscopy and image analysis[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=20" \o "Edit section: Microscopy and image analysis)]

Microscopic pictures allow us to locate both [organelles](https://en.wikipedia.org/wiki/Organelle) as well as molecules. It may also help us to distinguish between normal and abnormal cells, e.g. in [cancer](https://en.wikipedia.org/wiki/Cancer).

### Protein localization[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=21" \o "Edit section: Protein localization)]

The localization of proteins helps us to evaluate the role of a protein. For instance, if a protein is found in the [nucleus](https://en.wikipedia.org/wiki/Cell_nucleus) it may be involved in [gene regulation](https://en.wikipedia.org/wiki/Regulation_of_gene_expression) or [splicing](https://en.wikipedia.org/wiki/RNA_splicing). By contrast, if a protein is found in [mitochondria](https://en.wikipedia.org/wiki/Mitochondrion), it may be involved in [respiration](https://en.wikipedia.org/wiki/Cellular_respiration) or other [metabolic processes](https://en.wikipedia.org/wiki/Metabolism). Protein localization is thus an important component of [protein function prediction](https://en.wikipedia.org/wiki/Protein_function_prediction). There are well developed [protein subcellular localization prediction](https://en.wikipedia.org/wiki/Protein_subcellular_localization_prediction) resources available, including protein subcellualr location databases and prediction tools.

### Nuclear organisation of chromatin[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=22" \o "Edit section: Nuclear organisation of chromatin)]

*Main article:*[*Nuclear organization*](https://en.wikipedia.org/wiki/Nuclear_organization)

Data from high-throughput [chromosome conformation capture](https://en.wikipedia.org/wiki/Chromosome_conformation_capture) experiments, such as [Hi-C (experiment)](https://en.wikipedia.org/wiki/Hi-C_(experiment)) and [ChIA-PET](https://en.wikipedia.org/wiki/ChIA-PET" \o "ChIA-PET), can provide information on the spatial proximity of DNA loci. Analysis of these experiments can determine the three-dimensional structure and [nuclear organization](https://en.wikipedia.org/wiki/Nuclear_organization) of chromatin. Bioinformatic challenges in this field include partitioning the genome into domains, such as [Topologically Associating Domains](https://en.wikipedia.org/wiki/Topologically_Associating_Domain) (TADs), that are organised together in three-dimensional space.[[32]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-32)

## Structural bioinformatics[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=23" \o "Edit section: Structural bioinformatics)]

*Main articles:*[*Structural bioinformatics*](https://en.wikipedia.org/wiki/Structural_bioinformatics)*and*[*Protein structure prediction*](https://en.wikipedia.org/wiki/Protein_structure_prediction)

*See also:*[*Structural motif*](https://en.wikipedia.org/wiki/Structural_motif)*and*[*Structural domain*](https://en.wikipedia.org/wiki/Structural_domain)

[](https://en.wikipedia.org/wiki/File:1kqf_opm.png)

3-dimensional protein structures such as this one are common subjects in bioinformatic analyses.

Protein structure prediction is another important application of bioinformatics. The [amino acid](https://en.wikipedia.org/wiki/Amino_acid) sequence of a protein, the so-called [primary structure](https://en.wikipedia.org/wiki/Primary_structure), can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the [bovine spongiform encephalopathy](https://en.wikipedia.org/wiki/Bovine_spongiform_encephalopathy) – a.k.a. [Mad Cow Disease](https://en.wikipedia.org/wiki/Mad_Cow_Disease) – [prion](https://en.wikipedia.org/wiki/Prion).) Knowledge of this structure is vital in understanding the function of the protein. Structural information is usually classified as one of [*secondary*](https://en.wikipedia.org/wiki/Secondary_structure), [*tertiary*](https://en.wikipedia.org/wiki/Tertiary_structure) and [*quaternary*](https://en.wikipedia.org/wiki/Quaternary_structure) structure. A viable general solution to such predictions remains an open problem. Most efforts have so far been directed towards heuristics that work most of the time.[*[citation needed](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed" \o "Wikipedia:Citation needed)*]

One of the key ideas in bioinformatics is the notion of [homology](https://en.wikipedia.org/wiki/Homology_(biology)). In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B,* whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes ([leghemoglobin](https://en.wikipedia.org/wiki/Leghemoglobin)). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.[[33]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-33)

Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

## Network and systems biology[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=24" \o "Edit section: Network and systems biology)]

*Main articles:*[*Computational systems biology*](https://en.wikipedia.org/wiki/Computational_systems_biology)*,*[*Biological network*](https://en.wikipedia.org/wiki/Biological_network)*, and [Interactome](https://en.wikipedia.org/wiki/Interactome" \o "Interactome)*

*Network analysis* seeks to understand the relationships within [biological networks](https://en.wikipedia.org/wiki/Biological_network) such as [metabolic](https://en.wikipedia.org/wiki/Metabolic_network) or [protein–protein interaction networks](https://en.wikipedia.org/wiki/Interactome). Although biological networks can be constructed from a single type of molecule or entity (such as genes), network biology often attempts to integrate many different data types, such as proteins, small molecules, gene expression data, and others, which are all connected physically, functionally, or both.

*Systems biology* involves the use of [computer simulations](https://en.wikipedia.org/wiki/Computer_simulation) of [cellular](https://en.wikipedia.org/wiki/Cell_(biology)) subsystems (such as the [networks of metabolites](https://en.wikipedia.org/wiki/Metabolic_network) and [enzymes](https://en.wikipedia.org/wiki/Enzyme) that comprise [metabolism](https://en.wikipedia.org/wiki/Metabolism), [signal transduction](https://en.wikipedia.org/wiki/Signal_transduction) pathways and [gene regulatory networks](https://en.wikipedia.org/wiki/Gene_regulatory_network)) to both analyze and visualize the complex connections of these cellular processes. [Artificial life](https://en.wikipedia.org/wiki/Artificial_life) or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

### Molecular interaction networks[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=25" \o "Edit section: Molecular interaction networks)]

[](https://en.wikipedia.org/wiki/File:The_protein_interaction_network_of_Treponema_pallidum.png)

Interactions between proteins are frequently visualized and analyzed using networks. This network is made up of protein–protein interactions from *[Treponema pallidum](https://en.wikipedia.org/wiki/Treponema_pallidum" \o "Treponema pallidum)*, the causative agent of [syphilis](https://en.wikipedia.org/wiki/Syphilis) and other diseases.

*Main articles:*[*Protein–protein interaction prediction*](https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction_prediction)*and [interactome](https://en.wikipedia.org/wiki/Interactome" \o "Interactome)*

Tens of thousands of three-dimensional protein structures have been determined by [X-ray crystallography](https://en.wikipedia.org/wiki/X-ray_crystallography) and [protein nuclear magnetic resonance spectroscopy](https://en.wikipedia.org/wiki/Protein_nuclear_magnetic_resonance_spectroscopy) (protein NMR) and a central question in structural bioinformatics is whether it is practical to predict possible protein–protein interactions only based on these 3D shapes, without performing [protein–protein interaction](https://en.wikipedia.org/wiki/Protein%E2%80%93protein_interaction) experiments. A variety of methods have been developed to tackle the [protein–protein docking](https://en.wikipedia.org/wiki/Protein%E2%80%93protein_docking) problem, though it seems that there is still much work to be done in this field.

Other interactions encountered in the field include Protein–ligand (including drug) and [protein–peptide](https://en.wikipedia.org/w/index.php?title=Protein%E2%80%93peptide&action=edit&redlink=1). Molecular dynamic simulation of movement of atoms about rotatable bonds is the fundamental principle behind computational [algorithms](https://en.wikipedia.org/wiki/Algorithm), termed docking algorithms, for studying [molecular interactions](https://en.wikipedia.org/wiki/Interactome).

## Others[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=26" \o "Edit section: Others)]

### Literature analysis[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=27" \o "Edit section: Literature analysis)]

*Main articles:*[*Text mining*](https://en.wikipedia.org/wiki/Text_mining)*and*[*Biomedical text mining*](https://en.wikipedia.org/wiki/Biomedical_text_mining)

The growth in the number of published literature makes it virtually impossible to read every paper, resulting in disjointed sub-fields of research. Literature analysis aims to employ computational and statistical linguistics to mine this growing library of text resources. For example:

* Abbreviation recognition – identify the long-form and abbreviation of biological terms
* Named entity recognition – recognizing biological terms such as gene names
* Protein–protein interaction – identify which [proteins](https://en.wikipedia.org/wiki/Protein) interact with which proteins from text

The area of research draws from [statistics](https://en.wikipedia.org/wiki/Statistics) and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics).

### High-throughput image analysis[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=28" \o "Edit section: High-throughput image analysis)]

Computational technologies are used to accelerate or fully automate the processing, quantification and analysis of large amounts of high-information-content [biomedical imagery](https://en.wikipedia.org/wiki/Medical_imaging). Modern image analysis systems augment an observer's ability to make measurements from a large or complex set of images, by improving [accuracy](https://en.wikipedia.org/wiki/Accuracy), [objectivity](https://en.wikipedia.org/wiki/Objectivity_(science)), or speed. A fully developed analysis system may completely replace the observer. Although these systems are not unique to biomedical imagery, biomedical imaging is becoming more important for both [diagnostics](https://en.wikipedia.org/wiki/Diagnostics) and research. Some examples are:

* high-throughput and high-fidelity quantification and sub-cellular localization ([high-content screening](https://en.wikipedia.org/wiki/High-content_screening), cytohistopathology, [Bioimage informatics](https://en.wikipedia.org/wiki/Bioimage_informatics" \o "Bioimage informatics))
* [morphometrics](https://en.wikipedia.org/wiki/Morphometrics)
* clinical image analysis and visualization
* determining the real-time air-flow patterns in breathing lungs of living animals
* quantifying occlusion size in real-time imagery from the development of and recovery during arterial injury
* making behavioral observations from extended video recordings of laboratory animals
* infrared measurements for metabolic activity determination
* inferring clone overlaps in [DNA mapping](https://en.wikipedia.org/wiki/Gene_mapping), e.g. the [Sulston score](https://en.wikipedia.org/wiki/Sulston_score)

### High-throughput single cell data analysis[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=29" \o "Edit section: High-throughput single cell data analysis)]

*Main article:*[*Flow cytometry bioinformatics*](https://en.wikipedia.org/wiki/Flow_cytometry_bioinformatics)

Computational techniques are used to analyse high-throughput, low-measurement single cell data, such as that obtained from [flow cytometry](https://en.wikipedia.org/wiki/Flow_cytometry). These methods typically involve finding populations of cells that are relevant to a particular disease state or experimental condition.

### Biodiversity informatics[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=30" \o "Edit section: Biodiversity informatics)]

*Main article:*[*Biodiversity informatics*](https://en.wikipedia.org/wiki/Biodiversity_informatics)

Biodiversity informatics deals with the collection and analysis of [biodiversity](https://en.wikipedia.org/wiki/Biodiversity) data, such as [taxonomic databases](https://en.wikipedia.org/wiki/Taxonomic_database), or [microbiome](https://en.wikipedia.org/wiki/Microbiome) data. Examples of such analyses include [phylogenetics](https://en.wikipedia.org/wiki/Phylogenetics" \o "Phylogenetics), [niche modelling](https://en.wikipedia.org/wiki/Niche_modelling), [species richness](https://en.wikipedia.org/wiki/Species_richness) mapping, [DNA barcoding](https://en.wikipedia.org/wiki/DNA_barcoding), or [species](https://en.wikipedia.org/wiki/Speciesism) identification tools.

## Databases[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=31" \o "Edit section: Databases)]

*Main articles:*[*List of biological databases*](https://en.wikipedia.org/wiki/List_of_biological_databases)*and*[*Biological database*](https://en.wikipedia.org/wiki/Biological_database)

Databases are essential for bioinformatics research and applications. Many databases exist, covering various information types: for example, DNA and protein sequences, molecular structures, phenotypes and biodiversity. Databases may contain empirical data (obtained directly from experiments), predicted data (obtained from analysis), or, most commonly, both. They may be specific to a particular organism, pathway or molecule of interest. Alternatively, they can incorporate data compiled from multiple other databases. These databases vary in their format, access mechanism, and whether they are public or not.

Some of the most commonly used databases are listed below. For a more comprehensive list, please check the link at the beginning of the subsection.

* Used in biological sequence analysis: [Genbank](https://en.wikipedia.org/wiki/Genbank" \o "Genbank), [UniProt](https://en.wikipedia.org/wiki/UniProt" \o "UniProt)
* Used in finding Protein Families and [Motif](https://en.wikipedia.org/wiki/Sequence_motif) Finding: [InterPro](https://en.wikipedia.org/wiki/InterPro" \o "InterPro), [Pfam](https://en.wikipedia.org/wiki/Pfam" \o "Pfam)
* Used for Next Generation Sequencing: [Sequence Read Archive](https://en.wikipedia.org/wiki/Sequence_Read_Archive)
* Used in Network Analysis: Metabolic Pathway Databases ([KEGG](https://en.wikipedia.org/wiki/KEGG), [BioCyc](https://en.wikipedia.org/wiki/BioCyc_database_collection" \o "BioCyc database collection)), Interaction Analysis Databases, Functional Networks
* Used in design of synthetic genetic circuits: [GenoCAD](https://en.wikipedia.org/wiki/GenoCAD" \o "GenoCAD)

## Software and tools[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=32" \o "Edit section: Software and tools)]

[Software](https://en.wikipedia.org/wiki/Software) tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services available from various [bioinformatics companies](https://en.wikipedia.org/wiki/List_of_bioinformatics_companies) or public institutions.

### Open-source bioinformatics software[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=33" \o "Edit section: Open-source bioinformatics software)]

Many [free and open-source software](https://en.wikipedia.org/wiki/Free_and_open-source_software) tools have existed and continued to grow since the 1980s.[[34]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-obf-main-34) The combination of a continued need for new [algorithms](https://en.wikipedia.org/wiki/Algorithm) for the analysis of emerging types of biological readouts, the potential for innovative [*in silico*](https://en.wikipedia.org/wiki/In_silico) experiments, and freely available [open code](https://en.wikipedia.org/wiki/Open_code) bases have helped to create opportunities for all research groups to contribute to both bioinformatics and the range of open-source software available, regardless of their funding arrangements. The open source tools often act as incubators of ideas, or community-supported [plug-ins](https://en.wikipedia.org/wiki/Plug-in_(computing)) in commercial applications. They may also provide [*de facto*](https://en.wikipedia.org/wiki/De_facto) standards and shared object models for assisting with the challenge of bioinformation integration.

The [range of open-source software packages](https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software) includes titles such as [Bioconductor](https://en.wikipedia.org/wiki/Bioconductor), [BioPerl](https://en.wikipedia.org/wiki/BioPerl), [Biopython](https://en.wikipedia.org/wiki/Biopython), [BioJava](https://en.wikipedia.org/wiki/BioJava), [BioJS](https://en.wikipedia.org/wiki/BioJS), [BioRuby](https://en.wikipedia.org/wiki/BioRuby), [Bioclipse](https://en.wikipedia.org/wiki/Bioclipse), [EMBOSS](https://en.wikipedia.org/wiki/EMBOSS), [.NET Bio](https://en.wikipedia.org/wiki/.NET_Bio), [Orange](https://en.wikipedia.org/wiki/Orange_(software)) with its bioinformatics add-on, [Apache Taverna](https://en.wikipedia.org/wiki/Apache_Taverna), [UGENE](https://en.wikipedia.org/wiki/UGENE) and [GenoCAD](https://en.wikipedia.org/wiki/GenoCAD). To maintain this tradition and create further opportunities, the non-profit [Open Bioinformatics Foundation](https://en.wikipedia.org/wiki/Open_Bioinformatics_Foundation)[[34]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-obf-main-34) have supported the annual [Bioinformatics Open Source Conference](https://en.wikipedia.org/wiki/Bioinformatics_Open_Source_Conference) (BOSC) since 2000.[[35]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-obf-bosc-35)

An alternative method to build public bioinformatics databases is to use the MediaWiki engine with the *[WikiOpener](https://www.mediawiki.org/wiki/Extension:WikiOpener)* extension. This system allows the database to be accessed and updated by all experts in the field.[[36]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-36)

### Web services in bioinformatics[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=34" \o "Edit section: Web services in bioinformatics)]

[SOAP](https://en.wikipedia.org/wiki/SOAP)- and [REST](https://en.wikipedia.org/wiki/REST)-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the [EBI](https://en.wikipedia.org/wiki/European_Bioinformatics_Institute) into three categories: [SSS](https://en.wikipedia.org/wiki/Sequence_alignment_software) (Sequence Search Services), [MSA](https://en.wikipedia.org/wiki/Multiple_sequence_alignment) (Multiple Sequence Alignment), and [BSA](https://en.wikipedia.org/wiki/Bioinformatics#Sequence_analysis) (Biological Sequence Analysis).[[37]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-37) The availability of these [service-oriented](https://en.wikipedia.org/wiki/Service-orientation) bioinformatics resources demonstrate the applicability of web-based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible [bioinformatics workflow management systems](https://en.wikipedia.org/wiki/Bioinformatics_workflow_management_systems).

### Bioinformatics workflow management systems[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=35" \o "Edit section: Bioinformatics workflow management systems)]

*Main article:*[*Bioinformatics workflow management systems*](https://en.wikipedia.org/wiki/Bioinformatics_workflow_management_systems)

A [Bioinformatics workflow management system](https://en.wikipedia.org/wiki/Bioinformatics_workflow_management_systems) is a specialized form of a [workflow management system](https://en.wikipedia.org/wiki/Workflow_management_system) designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in a Bioinformatics application. Such systems are designed to

* provide an easy-to-use environment for individual application scientists themselves to create their own workflows,
* provide interactive tools for the scientists enabling them to execute their workflows and view their results in real-time,
* simplify the process of sharing and reusing workflows between the scientists, and
* enable scientists to track the [provenance](https://en.wikipedia.org/wiki/Provenance) of the workflow execution results and the workflow creation steps.

Some of the platforms giving this service: [Galaxy](https://en.wikipedia.org/wiki/Galaxy_(computational_biology)), [Kepler](https://en.wikipedia.org/wiki/Kepler_scientific_workflow_system), [Taverna](https://en.wikipedia.org/wiki/Apache_Taverna" \o "Apache Taverna), [UGENE](https://en.wikipedia.org/wiki/UGENE), [Anduril](https://en.wikipedia.org/wiki/Anduril_(workflow_engine)" \o "Anduril (workflow engine)).

## Education platforms[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=36" \o "Edit section: Education platforms)]

Software platforms designed to teach bioinformatics concepts and methods include [Rosalind](https://en.wikipedia.org/wiki/Rosalind_(education_platform)) and online courses offered through the [Swiss Institute of Bioinformatics](https://en.wikipedia.org/wiki/Swiss_Institute_of_Bioinformatics) Training Portal. The [Canadian Bioinformatics Workshops](https://en.wikipedia.org/wiki/Canadian_Bioinformatics_Workshops) provides videos and slides from training workshops on their website under a [Creative Commons](https://en.wikipedia.org/wiki/Creative_Commons) license. The 4273π project or 4273pi project[[38]](https://en.wikipedia.org/wiki/Bioinformatics" \l "cite_note-38) also offers open source educational materials for free. The course runs on low cost [Raspberry Pi](https://en.wikipedia.org/wiki/Raspberry_Pi) computers and has been used to teach adults and school pupils.[[39]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-39)[[40]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-40) 4273π is actively developed by a consortium of academics and research staff who have run research level bioinformatics using Raspberry Pi computers and the 4273π operating system.[[41]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-41)[[42]](https://en.wikipedia.org/wiki/Bioinformatics#cite_note-42)

[MOOC](https://en.wikipedia.org/wiki/Massive_open_online_course) platforms also provide online certifications in bioinformatics and related disciplines, including [Coursera](https://en.wikipedia.org/wiki/Coursera)'s Bioinformatics Specialization ([UC San Diego](https://en.wikipedia.org/wiki/University_of_California,_San_Diego)) and Genomic Data Science Specialization ([Johns Hopkins](https://en.wikipedia.org/wiki/Johns_Hopkins_University)) as well as [EdX](https://en.wikipedia.org/wiki/EdX" \o "EdX)'s Data Analysis for Life Sciences XSeries ([Harvard](https://en.wikipedia.org/wiki/Harvard_University)).

## Conferences[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=37" \o "Edit section: Conferences)]

There are several large conferences that are concerned with bioinformatics. Some of the most notable examples are [Intelligent Systems for Molecular Biology](https://en.wikipedia.org/wiki/Intelligent_Systems_for_Molecular_Biology) (ISMB), [European Conference on Computational Biology](https://en.wikipedia.org/wiki/European_Conference_on_Computational_Biology) (ECCB), and [Research in Computational Molecular Biology](https://en.wikipedia.org/wiki/Research_in_Computational_Molecular_Biology) (RECOMB).

## See also[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=38" \o "Edit section: See also)]

* [Biodiversity informatics](https://en.wikipedia.org/wiki/Biodiversity_informatics)
* [Bioinformatics companies](https://en.wikipedia.org/wiki/Bioinformatics_companies)
* [Computational biology](https://en.wikipedia.org/wiki/Computational_biology)
* [Computational biomodeling](https://en.wikipedia.org/wiki/Computational_biomodeling)
* [Computational genomics](https://en.wikipedia.org/wiki/Computational_genomics)
* [Functional genomics](https://en.wikipedia.org/wiki/Functional_genomics)
* [Health informatics](https://en.wikipedia.org/wiki/Health_informatics)
* [International Society for Computational Biology](https://en.wikipedia.org/wiki/International_Society_for_Computational_Biology)
* [Jumping library](https://en.wikipedia.org/wiki/Jumping_library)
* [List of bioinformatics institutions](https://en.wikipedia.org/wiki/List_of_bioinformatics_institutions)
* [List of open-source bioinformatics software](https://en.wikipedia.org/wiki/List_of_open-source_bioinformatics_software)
* [List of bioinformatics journals](https://en.wikipedia.org/wiki/List_of_bioinformatics_journals)
* [Margaret Oakley Dayhoff](https://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff)
* [Metabolomics](https://en.wikipedia.org/wiki/Metabolomics)
* [Nucleic acid sequence](https://en.wikipedia.org/wiki/Nucleic_acid_sequence)
* [Phylogenetics](https://en.wikipedia.org/wiki/Phylogenetics)
* [Proteomics](https://en.wikipedia.org/wiki/Proteomics)
* [Structural bioinformatics](https://en.wikipedia.org/wiki/Structural_bioinformatics)
* [Gene Disease Database](https://en.wikipedia.org/wiki/Gene_Disease_Database)

## References[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=39" \o "Edit section: References)]

* 1. [*"Bioinformatics"*](https://www.britannica.com/science/bioinformatics).Encyclopædia Britannica.com*. Retrieved 17 April 2017*.
  2. ^  [to:***a***](https://en.wikipedia.org/wiki/Bioinformatics#cite_ref-:0_2-0) [***b***](https://en.wikipedia.org/wiki/Bioinformatics#cite_ref-:0_2-1) Sim, Adelene YL; Minary, Peter; Levitt, Michael (2012-06-01).[*"Modeling nucleic acids"*](http://www.sciencedirect.com/science/article/pii/S0959440X12000632).Current Opinion in Structural Biology. Nucleic acids/Sequences and topology.**22**(3): 273–278.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*4028509*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4028509)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*22538125*](https://www.ncbi.nlm.nih.gov/pubmed/22538125).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1016/j.sbi.2012.03.012*](https://doi.org/10.1016%2Fj.sbi.2012.03.012).
  3. Dawson, Wayne K.; Maciejczyk, Maciej; Jankowska, Elzbieta J.; Bujnicki, Janusz M. (2016-07-01).[*"Coarse-grained modeling of RNA 3D structure"*](http://www.sciencedirect.com/science/article/pii/S1046202316301050).Methods. Advances in RNA Structure Determination.**103**: 138–156.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1016/j.ymeth.2016.04.026*](https://doi.org/10.1016%2Fj.ymeth.2016.04.026).
  4. Kmiecik, Sebastian; Gront, Dominik; Kolinski, Michal; Wieteska, Lukasz; Dawid, Aleksandra Elzbieta; Kolinski, Andrzej (2016-06-22).[*"Coarse-Grained Protein Models and Their Applications"*](https://dx.doi.org/10.1021/acs.chemrev.6b00163).Chemical Reviews.**116**: 7898–936.[*ISSN*](https://en.wikipedia.org/wiki/International_Standard_Serial_Number) [*0009-2665*](https://www.worldcat.org/issn/0009-2665).[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*27333362*](https://www.ncbi.nlm.nih.gov/pubmed/27333362).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1021/acs.chemrev.6b00163*](https://doi.org/10.1021%2Facs.chemrev.6b00163).
  5. Wong, KC (2016).[*Computational Biology and Bioinformatics: Gene Regulation*](https://www.crcpress.com/Computational-Biology-and-Bioinformatics-Gene-Regulation/Wong/p/book/9781498724975). CRC Press (Taylor & Francis Group).[*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*9781498724975*](https://en.wikipedia.org/wiki/Special:BookSources/9781498724975).
  6. Joyce, Adam P.; Zhang, Chi; Bradley, Philip; Havranek, James J. (2015-01-01).[*"Structure-based modeling of protein: DNA specificity"*](http://bfg.oxfordjournals.org/content/14/1/39).Briefings in Functional Genomics.**14**(1): 39–49.[*ISSN*](https://en.wikipedia.org/wiki/International_Standard_Serial_Number) [*2041-2649*](https://www.worldcat.org/issn/2041-2649).[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*4366589*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4366589)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*25414269*](https://www.ncbi.nlm.nih.gov/pubmed/25414269).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1093/bfgp/elu044*](https://doi.org/10.1093%2Fbfgp%2Felu044).
  7. Spiga, Enrico; Degiacomi, Matteo Thomas; Dal Peraro, Matteo (2014-01-01). Karabencheva-Christova, Tatyana, ed.[*Chapter Three - New Strategies for Integrative Dynamic Modeling of Macromolecular Assembly*](http://www.sciencedirect.com/science/article/pii/S1876162314000091). Biomolecular Modelling and Simulations.**96**. Academic Press. pp. 77–111.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1016/bs.apcsb.2014.06.008*](https://doi.org/10.1016%2Fbs.apcsb.2014.06.008).
  8. Hogeweg P (2011). Searls, David B., ed.[*"The Roots of Bioinformatics in Theoretical Biology"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068925).PLoS Computational Biology.**7**(3): e1002021.*[Bibcode](https://en.wikipedia.org/wiki/Bibcode" \o "Bibcode)*:*[2011PLSCB...7E0020H](http://adsabs.harvard.edu/abs/2011PLSCB...7E0020H)*.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*3068925*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068925)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*21483479*](https://www.ncbi.nlm.nih.gov/pubmed/21483479).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1371/journal.pcbi.1002021*](https://doi.org/10.1371%2Fjournal.pcbi.1002021).
  9. Hesper B, Hogeweg P (1970). "Bioinformatica: een werkconcept".**1**(6). Kameleon: 28–29.
  10. Hogeweg P (1978). "Simulating the growth of cellular forms".Simulation.**31**(3): 90–96.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1177/003754977803100305*](https://doi.org/10.1177%2F003754977803100305).
  11. Moody, Glyn (2004).Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business.[*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-471-32788-2*](https://en.wikipedia.org/wiki/Special:BookSources/978-0-471-32788-2).
  12. Dayhoff, M.O. (1966) Atlas of protein sequence and structure. National Biomedical Research Foundation, 215 pp.
  13. Eck RV, Dayhoff MO (1966). "Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences".Science.**152**(3720): 363–6.[*Bibcode*](https://en.wikipedia.org/wiki/Bibcode):*[1966Sci...152..363E](http://adsabs.harvard.edu/abs/1966Sci...152..363E)*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*17775169*](https://www.ncbi.nlm.nih.gov/pubmed/17775169).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1126/science.152.3720.363*](https://doi.org/10.1126%2Fscience.152.3720.363).
  14. Johnson G, Wu TT (January 2000).[*"Kabat Database and its applications: 30 years after the first variability plot"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102431).Nucleic Acids Res.**28**(1): 214–218.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*102431*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102431)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*10592229*](https://www.ncbi.nlm.nih.gov/pubmed/10592229).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1093/nar/28.1.214*](https://doi.org/10.1093%2Fnar%2F28.1.214).
  15. Attwood TK, Gisel A, Eriksson NE, Bongcam-Rudloff E (2011).[*"Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective"*](http://www.intechopen.com/articles/show/title/concepts-historical-milestones-and-the-central-place-of-bioinformatics-in-modern-biology-a-european-).Bioinformatics – Trends and Methodologies. InTech*. Retrieved 8 Jan 2012*.
  16. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (February 1977). "Nucleotide sequence of bacteriophage phi X174 DNA".Nature.**265**(5596): 687–95.[*Bibcode*](https://en.wikipedia.org/wiki/Bibcode):*[1977Natur.265..687S](http://adsabs.harvard.edu/abs/1977Natur.265..687S)*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*870828*](https://www.ncbi.nlm.nih.gov/pubmed/870828).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1038/265687a0*](https://doi.org/10.1038%2F265687a0).
  17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (January 2008).[*"GenBank"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238942).Nucleic Acids Res.**36**(Database issue): D25–30.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*2238942*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238942)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*18073190*](https://www.ncbi.nlm.nih.gov/pubmed/18073190).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1093/nar/gkm929*](https://doi.org/10.1093%2Fnar%2Fgkm929).
  18. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM (July 1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd".Science.**269**(5223): 496–512.[*Bibcode*](https://en.wikipedia.org/wiki/Bibcode):*[1995Sci...269..496F](http://adsabs.harvard.edu/abs/1995Sci...269..496F)*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*7542800*](https://www.ncbi.nlm.nih.gov/pubmed/7542800).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1126/science.7542800*](https://doi.org/10.1126%2Fscience.7542800).
  19. Carvajal-Rodríguez A (2012).[*"Simulation of Genes and Genomes Forward in Time"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2851118).Current Genomics. Bentham Science Publishers Ltd.**11**(1): 58–61.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*2851118*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2851118)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*20808525*](https://www.ncbi.nlm.nih.gov/pubmed/20808525).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.2174/138920210790218007*](https://doi.org/10.2174%2F138920210790218007).
  20. Brown, TA (2002). "Mutation, Repair and Recombination".Genomes(2nd ed.). Manchester (UK): Oxford.
  21. Carter, N. P.; Fiegler, H.; Piper, J. (2002). "Comparative analysis of comparative genomic hybridization microarray technologies: Report of a workshop sponsored by the Wellcome trust".Cytometry, Part A.**49**(2): 43–8.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1002/cyto.10153*](https://doi.org/10.1002%2Fcyto.10153).
  22. Chaudhari Narendrakumar M., Kumar Gupta Vinod, Dutta Chitra (2016). "BPGA-an ultra-fast pan-genome analysis pipeline".Scientific Reports.**6**.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1038/srep24373*](https://doi.org/10.1038%2Fsrep24373).
  23. Ionescu-Tîrgovişte, Constantin; Gagniuc, Paul Aurelian; Guja, Cristian.[*"Structural Properties of Gene Promoters Highlight More than Two Phenotypes of Diabetes"*](http://journals.plos.org/plosone/article?id=10.1371%252Fjournal.pone.0137950).PLOS ONE.**10**(9): e0137950.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*4574929*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574929)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*26379145*](https://www.ncbi.nlm.nih.gov/pubmed/26379145).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1371/journal.pone.0137950*](https://doi.org/10.1371%2Fjournal.pone.0137950).
  24. Aston KI (2014). "Genetic susceptibility to male infertility: News from genome-wide association studies".Andrology.**2**(3): 315–21.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*24574159*](https://www.ncbi.nlm.nih.gov/pubmed/24574159).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1111/j.2047-2927.2014.00188.x*](https://doi.org/10.1111%2Fj.2047-2927.2014.00188.x).
  25. Véron A, Blein S, Cox DG (2014). "Genome-wide association studies and the clinic: A focus on breast cancer".Biomarkers in Medicine.**8**(2): 287–96.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*24521025*](https://www.ncbi.nlm.nih.gov/pubmed/24521025).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.2217/bmm.13.121*](https://doi.org/10.2217%2Fbmm.13.121).
  26. Tosto G, Reitz C (2013).[*"Genome-wide association studies in Alzheimer's disease: A review"*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809844).Current Neurology and Neuroscience Reports.**13**(10): 381.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*3809844*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809844)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*23954969*](https://www.ncbi.nlm.nih.gov/pubmed/23954969).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1007/s11910-013-0381-0*](https://doi.org/10.1007%2Fs11910-013-0381-0).
  27. Londin E, Yadav P, Surrey S, Kricka LJ, Fortina P (2013). "Use of Linkage Analysis, Genome-Wide Association Studies, and Next-Generation Sequencing in the Identification of Disease-Causing Mutations".Pharmacogenomics. Methods in Molecular Biology.**1015**: 127–46.[*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-1-62703-434-0*](https://en.wikipedia.org/wiki/Special:BookSources/978-1-62703-434-0).[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*23824853*](https://www.ncbi.nlm.nih.gov/pubmed/23824853).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1007/978-1-62703-435-7\_8*](https://doi.org/10.1007%2F978-1-62703-435-7_8).
  28. Hindorff, L.A.,; et al. (2009).[*"Potential etiologic and functional implications of genome-wide association loci for human diseases and traits."*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2687147).Proc. Natl. Acad. Sci. USA.**106**: 9362–9367.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*2687147*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2687147)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*19474294*](https://www.ncbi.nlm.nih.gov/pubmed/19474294).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1073/pnas.0903103106*](https://doi.org/10.1073%2Fpnas.0903103106).
  29. Hall, L.O. (2010). "Finding the right genes for disease and prognosis prediction.".System Science and Engineering (ICSSE),2010 International Conference: 1–2.*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1109/ICSSE.2010.5551766*](https://doi.org/10.1109%2FICSSE.2010.5551766).
  30. Hye-Jung, E.C.; Jaswinder, K.; Martin, K.; Samuel, A.A; Marco, A.M (2014). "“Second-Generation Sequencing for Cancer Genome Analysis”". In Dellaire, Graham; Berman, Jason N.; Arceci, Robert J.Cancer Genomics. Boston (US): Academic Press. pp. 13–30.[*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*9780123969675*](https://en.wikipedia.org/wiki/Special:BookSources/9780123969675).[*doi*](https://en.wikipedia.org/wiki/Digital_object_identifier):[*10.1016/B978-0-12-396967-5.00002-5*](https://doi.org/10.1016%2FB978-0-12-396967-5.00002-5).
  31. Grosse I.,, Grau J., Ben-Gal I., Posch S., (2006.[*"VOMBAT: Prediction of Transcription Factor Binding Sites using Variable Order Bayesian Trees,"*](http://www.eng.tau.ac.il/~bengal/VOMBAT.pdf)(PDF). Nucleic Acids Research, vol. 34, issue W529–W533, 2006.
  32. Ay, Ferhat; Noble, William S. (2 September 2015). "Analysis methods for studying the 3D architecture of the genome".Genome Biology.**16**(1).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/s13059-015-0745-7*](https://doi.org/10.1186%2Fs13059-015-0745-7).
  33. Hoy, JA; Robinson, H; Trent JT, 3rd; Kakar, S; Smagghe, BJ; Hargrove, MS (3 August 2007). "Plant hemoglobins: a molecular fossil record for the evolution of oxygen transport.".Journal of Molecular Biology.**371**(1): 168–79.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*17560601*](https://www.ncbi.nlm.nih.gov/pubmed/17560601).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1016/j.jmb.2007.05.029*](https://doi.org/10.1016%2Fj.jmb.2007.05.029).
  34. [*"Open Bioinformatics Foundation: About us"*](http://www.open-bio.org/wiki/Main_Page).Official website.[*Open Bioinformatics Foundation*](https://en.wikipedia.org/wiki/Open_Bioinformatics_Foundation)*. Retrieved 10 May 2011*.
  35. [*"Open Bioinformatics Foundation: BOSC"*](http://www.open-bio.org/wiki/BOSC).Official website.[*Open Bioinformatics Foundation*](https://en.wikipedia.org/wiki/Open_Bioinformatics_Foundation)*. Retrieved 10 May 2011*.
  36. Brohée, Sylvain; Barriot, Roland; Moreau, Yves.[*"Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases"*](http://bioinformatics.oxfordjournals.org/content/26/17/2210.full).Bioinformatics. Oxford Journals*. Retrieved 5 May 2015*.
  37. Nisbet, Robert (14 May 2009). "BIOINFORMATICS".[*Handbook of Statistical Analysis and Data Mining Applications*](https://books.google.com/books?id=U5np34a5fmQC&pg=PA328&dq=bioinformatics+service+categories+EBI&hl=en&sa=X&ei=Jk9tU9KoKu_ksAS-x4DIDA&ved=0CD4Q6AEwAA#v=onepage&q=bioinformatics%20service%20categories%20EBI&f=false). John Elder IV, Gary Miner. Academic Press. p. 328*. Retrieved 9 May 2014*.
  38. Barker, D; Ferrier, D.E.K.; Holland, P.W; Mitchell, J.B.O; Plaisier, H; Ritchie, M.G; Smart, S.D. (2013).[*"4273π : bioinformatics education on low cost ARM hardware"*](http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-243).BMC Bioinformatics.**14**: 243.[*PMC*](https://en.wikipedia.org/wiki/PubMed_Central) [*3751261*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3751261)*Freely accessible*.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*23937194*](https://www.ncbi.nlm.nih.gov/pubmed/23937194).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/1471-2105-14-243*](https://doi.org/10.1186%2F1471-2105-14-243).
  39. Barker, D; Alderson, R.G; McDonagh, J.L; Plaisier, H; Comrie, M.M; Duncan, L; Muirhead, G.T.P; Sweeny, S.D. (2015).[*"University-level practical activities in bioinformatics benefit voluntary groups of pupils in the last 2 years of school"*](http://stemeducationjournal.springeropen.com/articles/10.1186/s40594-015-0030-z).International Journal of STEM Education.**2**(17).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/s40594-015-0030-z*](https://doi.org/10.1186%2Fs40594-015-0030-z).
  40. McDonagh, J.L; Barker, D; Alderson, R.G. (2016).[*"Bringing computational science to the public"*](http://springerplus.springeropen.com/articles/10.1186/s40064-016-1856-7).SpringerPlus.**5**(259).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/s40064-016-1856-7*](https://doi.org/10.1186%2Fs40064-016-1856-7).
  41. Robson, J.F.; Barker, D (2015).[*"Comparison of the protein-coding gene content of Chlamydia trachomatis and Protochlamydia amoebophila using a Raspberry Pi computer"*](http://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-015-1476-2).BMC Research Notes.**8**(561).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/s13104-015-1476-2*](https://doi.org/10.1186%2Fs13104-015-1476-2).
  42. Wregglesworth, K.M; Barker, D (2015).*["A comparison of the protein-coding genomes of two green sulphur bacteria, Chlorobium tepidum TLS and Pelodictyon phaeoclathratiforme BU-1"](http://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-015-1535-8)*.BMC Research Notes.**8**(565).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1186/s13104-015-1535-8*](https://doi.org/10.1186%2Fs13104-015-1535-8).

## Further reading[[edit](https://en.wikipedia.org/w/index.php?title=Bioinformatics&action=edit&section=40" \o "Edit section: Further reading)]

* Sehgal et al. : Structural, phylogenetic and docking studies of D-amino acid oxidase activator(DAOA ), a candidate schizophrenia gene. Theoretical Biology and Medical Modelling 2013 10 :3.
* Raul Isea [The Present-Day Meaning Of The Word Bioinformatics](http://gjar.org/publishpaper/vol2issue1/d75r28.pdf), Global Journal of Advanced Research, 2015
* Ilzins, O., Isea, R. and Hoebeke, J. [Can Bioinformatics Be Considered as an Experimental Biological Science](http://www.openscienceonline.com/journal/archive2?journalId=705&paperId=2496) 2015
* Achuthsankar S Nair [Computational Biology & Bioinformatics – A gentle Overview](http://print.achuth.googlepages.com/BINFTutorialV5.0CSI07.pdf), Communications of Computer Society of India, January 2007
* Aluru, Srinivas, ed. *Handbook of Computational Molecular Biology*. Chapman & Hall/Crc, 2006. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [1-58488-406-1](https://en.wikipedia.org/wiki/Special:BookSources/1-58488-406-1) (Chapman & Hall/Crc Computer and Information Science Series)
* Baldi, P and Brunak, S, *Bioinformatics: The Machine Learning Approach*, 2nd edition. MIT Press, 2001. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-262-02506-X](https://en.wikipedia.org/wiki/Special:BookSources/0-262-02506-X)
* Barnes, M.R. and Gray, I.C., eds., *Bioinformatics for Geneticists*, first edition. Wiley, 2003. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-470-84394-2](https://en.wikipedia.org/wiki/Special:BookSources/0-470-84394-2)
* Baxevanis, A.D. and Ouellette, B.F.F., eds., *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, third edition. Wiley, 2005. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-471-47878-4](https://en.wikipedia.org/wiki/Special:BookSources/0-471-47878-4)
* Baxevanis, A.D., Petsko, G.A., Stein, L.D., and Stormo, G.D., eds., [*Current Protocols*](https://en.wikipedia.org/wiki/Current_Protocols)*in Bioinformatics*. Wiley, 2007. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-471-25093-7](https://en.wikipedia.org/wiki/Special:BookSources/0-471-25093-7)
* Cristianini, N. and Hahn, M. [*Introduction to Computational Genomics*](http://www.computational-genomics.net/), Cambridge University Press, 2006. ([ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [9780521671910](https://en.wikipedia.org/wiki/Special:BookSources/9780521671910) | [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-521-67191-4](https://en.wikipedia.org/wiki/Special:BookSources/0-521-67191-4))
* Durbin, R., S. Eddy, A. Krogh and G. Mitchison, *Biological sequence analysis*. Cambridge University Press, 1998. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-521-62971-3](https://en.wikipedia.org/wiki/Special:BookSources/0-521-62971-3)
* Gilbert D (2004).[*"Bioinformatics software resources"*](http://bib.oxfordjournals.org/cgi/content/abstract/5/3/300).Briefings in Bioinformatics.**5**(3): 300–304.[*PMID*](https://en.wikipedia.org/wiki/PubMed_Identifier) [*15383216*](https://www.ncbi.nlm.nih.gov/pubmed/15383216).*[doi](https://en.wikipedia.org/wiki/Digital_object_identifier" \o "Digital object identifier)*:[*10.1093/bib/5.3.300*](https://doi.org/10.1093%2Fbib%2F5.3.300).
* Keedwell, E., *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*. Wiley, 2005. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-470-02175-6](https://en.wikipedia.org/wiki/Special:BookSources/0-470-02175-6)
* Kohane, et al. *Microarrays for an Integrative Genomics.* The MIT Press, 2002. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-262-11271-X](https://en.wikipedia.org/wiki/Special:BookSources/0-262-11271-X)
* Lund, O. et al. *Immunological Bioinformatics.* The MIT Press, 2005. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-262-12280-4](https://en.wikipedia.org/wiki/Special:BookSources/0-262-12280-4)
* [Pachter, Lior](https://en.wikipedia.org/wiki/Lior_Pachter) and [Sturmfels, Bernd](https://en.wikipedia.org/wiki/Bernd_Sturmfels). "Algebraic Statistics for Computational Biology" Cambridge University Press, 2005. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-521-85700-7](https://en.wikipedia.org/wiki/Special:BookSources/0-521-85700-7)
* Pevzner, Pavel A. *Computational Molecular Biology: An Algorithmic Approach* The MIT Press, 2000. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-262-16197-4](https://en.wikipedia.org/wiki/Special:BookSources/0-262-16197-4)
* Soinov, L. [Bioinformatics and Pattern Recognition Come Together](http://jprr.org/index.php/jprr/article/view/8/5) Journal of Pattern Recognition Research ([JPRR](http://www.jprr.org/)), Vol 1 (1) 2006 p. 37–41
* Stevens, Hallam, *Life Out of Sequence: A Data-Driven History of Bioinformatics*, Chicago: The University of Chicago Press, 2013, [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [9780226080208](https://en.wikipedia.org/wiki/Special:BookSources/9780226080208)
* Tisdall, James. "Beginning Perl for Bioinformatics" O'Reilly, 2001. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number_(identifier)) [0-596-00080-4](https://en.wikipedia.org/wiki/Special:BookSources/0-596-00080-4)
* [Dedicated issue of *Philosophical Transactions B* on Bioinformatics freely available](http://publishing.royalsociety.org/bioinformatics)[[*permanent dead link*](https://en.wikipedia.org/wiki/Wikipedia:Link_rot)]
* [Catalyzing Inquiry at the Interface of Computing and Biology (2005) CSTB report](http://www.nap.edu/catalog/11480.html)
* [Calculating the Secrets of Life: Contributions of the Mathematical Sciences and computing to Molecular Biology (1995)](http://www.nap.edu/catalog/2121.html)
* [Foundations of Computational and Systems Biology MIT Course](https://web.archive.org/web/20071222091912/http:/ocw.mit.edu/OcwWeb/Biology/7-91JSpring2004/LectureNotes/index.htm)
* [Computational Biology: Genomes, Networks, Evolution Free MIT Course](http://compbio.mit.edu/6.047/)