Title: Machine learning in bioinformatics
URL: https://en.wikipedia.org/wiki/Machine_learning_in_bioinformatics
PageID: 53970843
Categories: Category:Bioinformatics, Category:Machine learning

---
Artificial general intelligence
Intelligent agent
Recursive self-improvement
Planning
Computer vision
General game playing
Knowledge representation
Natural language processing
Robotics
AI safety
Machine learning
Symbolic
Deep learning
Bayesian networks
Evolutionary algorithms
Hybrid intelligent systems
Systems integration
Open-source
Bioinformatics
Deepfake
Earth sciences
Finance
Generative AI Art Audio Music
Art
Audio
Music
Government
Healthcare Mental health
Mental health
Industry
Software development

Translation

Military

Physics

Projects

AI alignment

Artificial consciousness

The bitter lesson

Chinese room

Friendly AI

Ethics

Existential risk

Turing test

Uncanny valley

Timeline

Progress

AI winter

AI boom

AI bubble

Glossary

v

t

e

Machine learning in bioinformatics is the application of machine learning algorithms to bioinformatics , [1] including genomics , proteomics , microarrays , systems biology , evolution , and text mining . [2] [3]

Prior to the emergence of machine learning, bioinformatics algorithms had to be programmed by hand; for problems such as protein structure prediction , this proved difficult. [4] Machine learning techniques such as deep learning can learn features of data sets rather than requiring the programmer to define them individually. The algorithm can further learn how to combine low-level features into more abstract features, and so on. This multi-layered approach allows such systems to make sophisticated predictions when appropriately trained. These methods contrast with other computational biology approaches which, while exploiting existing datasets, do not allow the data to be interpreted and analyzed in unanticipated ways.

Tasks

Machine learning algorithms in bioinformatics can be used for prediction, classification, and feature selection. Methods to achieve this task are varied and span many disciplines; most well known among them are machine learning and statistics. Classification and prediction tasks aim at building models that describe and distinguish classes or concepts for future prediction. The differences between them are the following:

Classification/recognition outputs a categorical class, while prediction outputs a numerical valued feature.

The type of algorithm, or process used to build the predictive models from data using analogies, rules, neural networks, probabilities, and/or statistics.

Due to the exponential growth of information technologies and applicable models, including artificial intelligence and data mining, in addition to the access ever-more comprehensive data sets, new and better information analysis techniques have been created, based on their ability to learn. Such models allow reach beyond description and provide insights in the form of testable models.

Approaches

Artificial neural networks

Artificial neural networks in bioinformatics have been used for: [5]

Comparing and aligning RNA, protein, and DNA sequences.

Identification of promoters and finding genes from sequences related to DNA.

Interpreting the expression-gene and micro-array data.

Identifying the network (regulatory) of genes.

Learning evolutionary relationships by constructing phylogenetic trees .

Classifying and predicting protein structure .

Molecular design and docking

Feature engineering

The way that features, often vectors in a many-dimensional space, are extracted from the domain data is an important component of learning systems. [6] In genomics, a typical representation of a sequence is a vector of k-mers frequencies, which is a vector of dimension 4 k {\displaystyle 4^{k}} whose entries count the appearance of each subsequence of length k {\displaystyle k} in a given sequence. Since for a value as small as k = 12 {\displaystyle k=12} the dimensionality of these vectors is huge (e.g. in this case the dimension is 4 12 ≈ 16 × 10 6 {\displaystyle 4^{12}\approx 16\times 10^{6}}), techniques such as principal component analysis are used to project the data to a lower dimensional space, thus selecting a smaller set of features from the sequences. [6] [7]

Classification

In this type of machine learning task, the output is a discrete variable. One example of this type of task in bioinformatics is labeling new genomic data (such as genomes of unculturable bacteria) based on a model of already labeled data. [6]

Hidden Markov models

Hidden Markov models (HMMs) are a class of statistical models for sequential data (often related to systems evolving over time). An HMM is composed of two mathematical objects: an observed state■dependent process X 1 , X 2 , … , X M {\displaystyle X_{1},X_{2},\ldots ,X_{M}} , and an unobserved (hidden) state process S 1 , S 2 , … , S T {\displaystyle S_{1},S_{2},\ldots ,S_{T}} . In an HMM, the state process is not directly observed – it is a 'hidden' (or 'latent') variable – but observations are made of a state■dependent process (or observation process) that is driven by the underlying state process (and which can thus be regarded as a noisy measurement of the system states of interest). [8] HMMs can be formulated in continuous time. [9] [10]

HMMs can be used to profile and convert a multiple sequence alignment into a position-specific scoring system suitable for searching databases for homologous sequences remotely. [11] Additionally, ecological phenomena can be described by HMMs. [12]

Convolutional neural networks

Convolutional neural networks (CNN) are a class of deep neural network whose architecture is based on shared weights of convolution kernels or filters that slide along input features, providing translation-equivariant responses known as feature maps. [13] [14] CNNs take advantage of the hierarchical pattern in data and assemble patterns of increasing complexity using smaller and simpler patterns discovered via their filters. [15]

Convolutional networks were inspired by biological processes [16] [17] [18] [19] in that the connectivity pattern between neurons resembles the organization of the animal visual cortex .

Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field . The receptive fields of different neurons partially overlap such that they cover the entire visual field.

CNN uses relatively little pre-processing compared to other image classification algorithms . This means that the network learns to optimize the filters (or kernels) through automated learning, whereas in traditional algorithms these filters are hand-engineered . This reduced reliance on prior knowledge of the analyst and on human intervention in manual feature extraction makes CNNs a desirable model. [15]

A phylogenetic convolutional neural network (Ph-CNN) is a convolutional neural network architecture proposed by Fioranti et al. in 2018 to classify metagenomics data. [20] In this approach, phylogenetic data is endowed with patristic distance (the sum of the lengths of all branches connecting two operational taxonomic units [OTU]) to select k-neighborhoods for each OTU, and each OTU and its neighbors are processed with convolutional filters.

Self-supervised learning

Unlike supervised methods, self-supervised learning methods learn representations without relying on annotated data. That is well-suited for genomics, where high throughput sequencing techniques can create potentially large amounts of unlabeled data. Some examples of self-supervised learning methods applied on genomics include DNABERT and Self-GenomeNet. [21] [22]

Random forest

Random forests (RF) classify by constructing an ensemble of decision trees , and outputting the average prediction of the individual trees. [23] This is a modification of bootstrap aggregating (which aggregates a large collection of decision trees) and can be used for classification or regression . [24] [25]

As random forests give an internal estimate of generalization error, cross-validation is unnecessary. In addition, they produce proximities, which can be used to impute missing values, and which enable novel data visualizations. [26]

Computationally, random forests are appealing because they naturally handle both regression and (multiclass) classification, are relatively fast to train and to predict, depend only on one or two tuning parameters, have a built-in estimate of the generalization error, can be used directly for high-dimensional problems, and can easily be implemented in parallel. Statistically, random forests are appealing for additional features, such as measures of variable importance, differential class weighting, missing value imputation, visualization, outlier detection, and unsupervised learning. [26]

Clustering

Clustering - the partitioning of a data set into disjoint subsets, so that the data in each subset are as close as possible to each other and as distant as possible from data in any other subset, according to some defined distance or similarity function - is a common technique for statistical data analysis.

Clustering is central to much data-driven bioinformatics research and serves as a powerful computational method whereby means of hierarchical, centroid-based, distribution-based, density-based, and self-organizing maps classification, has long been studied and used in classical machine learning settings. Particularly, clustering helps to analyze unstructured and high-dimensional data in the form of sequences, expressions, texts, images, and so on. Clustering is also used to gain insights into biological processes at the genomic level, e.g. gene functions, cellular processes, subtypes of cells, gene regulation , and metabolic processes. [27]

Clustering algorithms used in bioinformatics

Data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down).

Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it

into successively smaller clusters. Hierarchical clustering is calculated using metrics on Euclidean spaces , the most commonly used is the Euclidean distance computed by finding the square of the difference between each variable, adding all the squares, and finding the square root of the said sum. An example of a hierarchical clustering algorithm is BIRCH , which is particularly good on bioinformatics for its nearly linear time complexity given generally large datasets. [28] Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once. Most applications adopt one of two popular heuristic methods: k-means algorithm or k-medoids . Other algorithms do not require an initial number of groups, such as affinity propagation . In a genomic setting this algorithm has been used both to cluster biosynthetic gene clusters in gene cluster families(GCF) and to cluster said GCFs. [29]

Workflow

Typically, a workflow for applying machine learning to biological data goes through four steps: [2]

Recording, including capture and storage. In this step, different information sources may be merged into a single set.

Preprocessing, including cleaning and restructuring into a ready-to-analyze form. In this step, uncorrected data are eliminated or corrected, while missing data maybe imputed and relevant variables chosen.

Analysis, evaluating data using either supervised or unsupervised algorithms. The algorithm is typically trained on a subset of data, optimizing parameters, and evaluated on a separate test subset.

Visualization and interpretation, where knowledge is represented effectively using different methods to assess the significance and importance of the findings.

Data errors

Duplicate data is a significant issue in bioinformatics. Publicly available data may be of uncertain quality. [30]

Errors during experimentation. [30]

Erroneous interpretation. [30]

Typing mistakes. [30]

Non-standardized methods (3D structure in PDB from multiple sources, X-ray diffraction, theoretical modeling, nuclear magnetic resonance, etc.) are used in experiments. [30]

Applications

In general, a machine learning system can usually be trained to recognize elements of a certain class given sufficient samples. [31] For example, machine learning methods can be trained to identify specific visual features such as splice sites. [32]

Support vector machines have been extensively used in cancer genomic studies. [33] In addition, deep learning has been incorporated into bioinformatic algorithms. Deep learning applications have been used for regulatory genomics and cellular imaging. [34] Other applications include medical image classification, genomic sequence analysis, as well as protein structure classification and prediction. [35] Deep learning has been applied to regulatory genomics, variant calling and pathogenicity scores. [36] Natural language processing and text mining have helped to understand phenomena including protein-protein interaction, gene-disease relation as well as predicting biomolecule structures and functions. [37]

Precision/personalized medicine

Natural language processing algorithms personalized medicine for patients who suffer genetic diseases, by combining the extraction of clinical information and genomic data available from the patients. Institutes such as Health-funded Pharmacogenomics Research Network focus on finding breast cancer treatments. [38]

Precision medicine considers individual genomic variability, enabled by large-scale biological databases. Machine learning can be applied to perform the matching function between (groups of patients) and specific treatment modalities. [39]

Computational techniques are used to solve other problems, such as efficient primer design for PCR , biological-image analysis and back translation of proteins (which is, given the degeneration of the genetic code, a complex combinatorial problem). [2]

Genomics

While genomic sequence data has historically been sparse due to the technical difficulty of sequencing a piece of DNA, the number of available sequences is growing. On average, the number of bases available in the GenBank public repository has doubled every 18 months since 1982. [40] However, while raw data was becoming increasingly available and accessible, As of 2002 [update] , biological interpretation of this data was occurring at a much slower pace. [41] This made for an increasing need for developing computational genomics tools, including machine learning systems, that can automatically determine the location of protein-encoding genes within a given DNA sequence (i.e. gene prediction). [41]

Gene prediction is commonly performed through both extrinsic searches and intrinsic searches . [41] For the extrinsic search, the input DNA sequence is run through a large database of sequences whose genes have been previously discovered and their locations annotated and identifying the target sequence's genes by determining which strings of bases within the sequence are homologous to known gene sequences. However, not all the genes in a given input sequence can be identified through homology alone, due to limits in the size of the database of known and annotated gene sequences. Therefore, an intrinsic search is needed where a gene prediction program attempts to identify the remaining genes from the DNA sequence alone. [41]

Machine learning has also been used for the problem of multiple sequence alignment which involves aligning many DNA or amino acid sequences in order to determine regions of similarity that could indicate a shared evolutionary history. [2] It can also be used to detect and visualize genome rearrangements. [42]

Proteomics

Proteins , strings of amino acids , gain much of their function from protein folding , where they conform into a three-dimensional structure, including the primary structure , the secondary structure (alpha helices and beta sheets), the tertiary structure , and the quaternary structure .

Protein secondary structure prediction is a main focus of this subfield as tertiary and quaternary structures are determined based on the secondary structure. [4] Solving the true structure of a protein is expensive and time-intensive, furthering the need for systems that can accurately predict the structure of a protein by analyzing the amino acid sequence directly. [4] [2] Prior to machine learning, researchers needed to conduct this prediction manually. This trend began in 1951 when Pauling and Corey released their work on predicting the hydrogen bond configurations of a protein from a polypeptide chain. [43] Automatic feature learning reaches an accuracy of 82-84%. [4] [44] Recent approaches have utilized deep learning techniques for state-of-the-art secondary structure predictions. For example, DeepCNF (deep convolutional neural fields) achieved an accuracy of approximately 84% when tasked to classify the amino acids of a protein sequence into one of three structural classes (helix, sheet, or coil). [44] The theoretical limit for three-state protein secondary structure is 88–90%. [4] In 2018, AlphaFold , an artificial intelligence (AI) program developed by DeepMind , placed first in the overall rankings of the 13th Critical Assessment of Structure Prediction (CASP). It was particularly successful at predicting the most accurate structures for targets rated as most difficult by the competition organizers, where no existing template structures were available from proteins with partially similar sequences. AlphaFold 2 (2020) repeated this placement in the CASP14 competition and achieved a level of accuracy much higher than any other entry. [45] [46] [47]

Machine learning has also been applied to proteomics problems such as protein side-chain prediction, protein loop modeling, and protein contact map prediction. [2]

Metagenomics

Metagenomics is the study of microbial communities from environmental DNA samples. [48] Currently, limitations and challenges predominate in the implementation of machine learning tools due to the amount of data in environmental samples. [49] Supercomputers and web servers have made access to these tools easier. [50] The high dimensionality of microbiome datasets is a major challenge in studying the microbiome; this significantly limits the power of current approaches for identifying true differences and increases the chance of false discoveries. [51] [better source needed]

Despite their importance, machine learning tools related to metagenomics have focused on the study of gut microbiota and the relationship with digestive diseases, such as inflammatory bowel disease (IBD), Clostridioides difficile infection (CDI), colorectal cancer and diabetes , seeking better diagnosis and treatments. [50] Many algorithms were developed to classify microbial communities according to the health condition of the host, regardless of the type of sequence data, e.g. 16S rRNA or whole-genome sequencing (WGS), using methods such as least absolute shrinkage and selection operator classifier, random forest , supervised classification model, and gradient boosted tree model. Neural networks , such as recurrent neural networks (RNN), convolutional neural networks (CNN), and Hopfield neural networks have been added. [50] For example, in 2018, Fioravanti et al. developed an algorithm called Ph-CNN to classify data samples from healthy patients and patients with IBD symptoms (to distinguish healthy and sick patients) by using phylogenetic trees and convolutional neural networks. [52]

In addition, random forest (RF) methods and implemented importance measures help in the identification of microbiome species that can be used to distinguish diseased and non-diseased samples. However, the performance of a decision tree and the diversity of decision trees in the ensemble significantly influence the performance of RF algorithms. The generalization error for RF measures how accurate the individual classifiers are and their interdependence. Therefore, the high dimensionality problems of microbiome datasets pose challenges. Effective approaches require many possible variable combinations, which exponentially increases the computational burden as the number of features increases. [51]

For microbiome analysis in 2020 Dang & Kishino [51] developed a novel analysis pipeline. The core of the pipeline is an RF classifier coupled with forwarding variable selection (RF-FVS), which selects a minimum-size core set of microbial species or functional signatures that maximize the predictive classifier performance. The framework combines:

identifying a few significant features by a massively parallel forward variable selection procedure

mapping the selected species on a phylogenetic tree , and

predicting functional profiles by functional gene enrichment analysis from metagenomic 16S rRNA data.

They demonstrated performance by analyzing two published datasets from large-scale case-control studies:

16S rRNA gene amplicon data for C. difficile infection (CDI) and

shotgun metagenomics data for human colorectal cancer (CRC).

The proposed approach improved the accuracy from 81% to 99.01% for CDI and from 75.14% to 90.17% for CRC.

The use of machine learning in environmental samples has been less explored, maybe because of data complexity, especially from WGS. Some works show that it is possible to apply these tools in environmental samples. In 2021 Dhungel et al., [53] designed an R package called MegaR. This package allows working with 16S rRNA and whole metagenomic sequences to make taxonomic profiles and classification models by machine learning models. MegaR includes a comfortable visualization environment to improve the user experience. Machine learning in environmental metagenomics can help to answer questions related to the interactions between microbial communities and ecosystems, e.g. the work of Xun et al., in 2021 [54] where the use of different machine learning methods offered insights on the relationship among the soil, microbiome biodiversity, and ecosystem stability.

Microarrays

Microarrays , a type of lab-on-a-chip , are used for automatically collecting data about large amounts of biological material. Machine learning can aid in analysis, and has been applied to expression pattern identification, classification, and genetic network induction. [2]

This technology is especially useful for monitoring gene expression, aiding in diagnosing cancer by examining which genes are expressed. [55] One of the main tasks is identifying which genes are expressed based on the collected data. [2] In addition, due to the huge number of genes on which data is collected by the microarray, winnowing the large amount of irrelevant data to the task of expressed gene identification is challenging. Machine learning presents a potential solution as various classification methods can be used to perform this identification. The most commonly used methods are radial basis function networks , deep learning , Bayesian classification , decision trees , and random forest . [55]

Systems biology

Systems biology focuses on the study of emergent behaviors from complex interactions of simple biological components in a system. Such components can include DNA, RNA, proteins, and metabolites. [56]

Machine learning has been used to aid in modeling these interactions in domains such as genetic networks, signal transduction networks, and metabolic pathways. [2] Probabilistic graphical models , a machine learning technique for determining the relationship between different variables, are one of the most commonly used methods for modeling genetic networks. [2] In addition, machine learning has been applied to systems biology problems such as identifying transcription factor binding sites using Markov chain optimization . [2] Genetic algorithms , machine learning techniques which are based on the natural process of evolution, have been used to model genetic networks and regulatory structures. [2]

Other systems biology applications of machine learning include the task of enzyme function prediction, high throughput microarray data analysis, analysis of genome-wide association studies to better understand markers of disease, protein function prediction. [57]

Evolution

This domain, particularly phylogenetic tree reconstruction, uses the features of machine learning techniques. Phylogenetic trees are schematic representations of the evolution of organisms. Initially, they were constructed using features such as morphological and metabolic features. Later, due to the availability of genome sequences, the construction of the phylogenetic tree algorithm used the concept based on genome comparison. With the help of optimization techniques, a comparison was done by means of multiple sequence alignment. [58]

Stroke diagnosis

Machine learning methods for the analysis of neuroimaging data are used to help diagnose stroke . Historically multiple approaches to this problem involved neural networks. [59] [60]

Multiple approaches to detect strokes used machine learning. As proposed by Mirtskhulava, [61] feed-forward networks were tested to detect strokes using neural imaging. As proposed by Titano [62] 3D-CNN techniques were tested in supervised classification to screen head CT images for acute neurologic events. Three-dimensional CNN and SVM methods are often used. [60]

Text mining

The increase in biological publications increased the difficulty in searching and compiling relevant available information on a given topic. This task is known as knowledge extraction . It is necessary for biological data collection which can then in turn be fed into machine learning algorithms to generate new biological knowledge. [2] [63] Machine learning can be used for this knowledge extraction task using techniques such as natural language processing to extract the useful information from human-generated reports in a database. Text Nailing , an alternative approach to machine learning, capable of extracting features from clinical narrative notes was introduced in 2017.

This technique has been applied to the search for novel drug targets, as this task requires the examination of information stored in biological databases and journals. [63] Annotations of proteins in protein databases often do not reflect the complete known set of knowledge of each protein, so additional information must be extracted from biomedical literature. Machine learning has been applied to the automatic annotation of gene and protein function, determination of the protein subcellular localization , DNA-expression array analysis, large-scale protein interaction analysis, and molecule interaction analysis. [63]

Another application of text mining is the detection and visualization of distinct DNA regions given sufficient reference data. [64]

Clustering and abundance profiling of biosynthetic gene clusters

Microbial communities are complex assembles of diverse microorganisms, [65] where symbiont partners constantly produce diverse metabolites derived from the primary and secondary (specialized) metabolism, from which metabolism plays an important role in microbial interaction. [66] Metagenomic and metatranscriptomic data are an important source for deciphering communications signals.

Molecular mechanisms produce specialized metabolites in various ways. Biosynthetic Gene Clusters (BGCs) attract attention, since several metabolites are clinically valuable, anti-microbial, anti-fungal, anti-parasitic, anti-tumor and immunosuppressive agents produced by the modular action of multi-enzymatic, multi-domains gene clusters, such as Nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs). [67] Diverse studies [68] [69] [70] [71] [72] [73] [74] [75] show that grouping BGCs that share homologous core genes into gene cluster families (GCFs) can yield useful insights into the chemical diversity of the analyzed strains, and can support linking BGCs to their secondary metabolites. [69] [71] GCFs have been used as functional markers in human health studies [76] [77] and to study the ability of soil to suppress fungal pathogens. [78] Given their direct relationship to catalytic enzymes, and compounds produced from their encoded pathways, BGCs/GCFs can serve as a proxy to explore the chemical space of microbial secondary metabolism. Cataloging GCFs in sequenced microbial genomes yields an overview of the existing chemical diversity and offers insights into future priorities. [68] [70] Tools such as BiG-SLiCE and BIG-MAP [79] have emerged with the sole purpose of unveiling the importance of BGCs in natural environments.

Decodification of RiPPs chemical structures

The increase of experimentally characterized ribosomally synthesized and post-translationally modified peptides (RiPPs), together with the availability of information on their sequence and chemical structure, selected from databases such as BAGEL, BACTIBASE, MIBIG, and THIOBASE, provide the opportunity to develop machine learning tools to decode the chemical structure and classify them.

In 2017, researchers at the National Institute of Immunology of New Delhi, India, developed RiPPMiner [80] software, a bioinformatics resource for decoding RiPP chemical structures by genome mining. The RiPPMiner web server consists of a query interface and the RiPPDB database. RiPPMiner defines 12 subclasses of RiPPs, predicting the cleavage site of the leader peptide and the final cross-link of the RiPP chemical structure.

Mass spectral similarity scoring

Many tandem mass spectrometry (MS/MS) based metabolomics studies, such as library matching and molecular networking, use spectral similarity as a proxy for structural similarity. Spec2vec [81] algorithm provides a new way of spectral similarity score, based on Word2Vec . Spec2Vec learns fragmental relationships within a large set of spectral data, in order to assess spectral similarities between molecules and to classify unknown molecules through these comparisons.

For systemic annotation, some metabolomics studies rely on fitting measured fragmentation mass spectra to library spectra or contrasting spectra via network analysis. Scoring functions are used to determine the similarity between pairs of fragment spectra as part of these processes. So far, no research has suggested scores that are significantly different from the commonly utilized

cosine-based similarity . [82]

Databases

An important part of bioinformatics is the management of big datasets, known as databases of reference. Databases exist for each type of biological data, for example for biosynthetic gene clusters and metagenomes.

General databases by bioinformatics

National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI) [83] provides a large suite of online resources for biological information and data, including the GenBank nucleic acid sequence database and the PubMed database of citations and abstracts for published life science journals. Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized data sets. Resources include PubMed Data Management, RefSeq Functional Elements, genome data download, variation services API, Magic-BLAST, QuickBLASTp, and Identical Protein Groups. All of these resources can be accessed through NCBI. [84]

Bioinformatics analysis for biosynthetic gene clusters

antiSMASH

antiSMASH allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. It integrates and cross-links with a large number of in silico secondary metabolite analysis tools. [85]

gutSMASH

gutSMASH is a tool that systematically evaluates bacterial metabolic potential by predicting both known and novel anaerobic metabolic gene clusters (MGCs) from the gut microbiome .

MIBiG

MIBiG, [86] the minimum information about a biosynthetic gene cluster specification, provides a standard for annotations and metadata on biosynthetic gene clusters and their molecular products. MIBiG is a Genomic Standards Consortium project that builds on the minimum information about any sequence (MIxS) framework. [87]

MIBiG facilitates the standardized deposition and retrieval of biosynthetic gene cluster data as well as the development of comprehensive comparative analysis tools. It empowers next-generation research on the biosynthesis, chemistry and ecology of broad classes of societally relevant bioactive secondary metabolites , guided by robust experimental evidence and rich metadata components. [88]

SILVA

SILVA [89] is an interdisciplinary project among biologists and computers scientists assembling a complete database of RNA ribosomal (rRNA) sequences of genes, both small (16S , 18S , SSU) and large (23S , 28S , LSU) subunits, which belong to the bacteria, archaea and eukarya domains. These data are freely available for academic and commercial use. [90]

Greengenes

Greengenes [91] is a full-length 16S rRNA gene database that provides chimera screening, standard alignment and a curated taxonomy based on de novo tree inference. [92] [93] Overview:

1,012,863 RNA sequences from 92,684 organisms contributed to RNAcentral.

The shortest sequence has 1,253 nucleotides, the longest 2,368.

The average length is 1,402 nucleotides.

Database version: 13.5.

Open Tree of Life Taxonomy

Open Tree of Life Taxonomy (OTT) [94] aims to build a complete, dynamic, and digitally available Tree of Life by synthesizing published phylogenetic trees along with taxonomic data. Phylogenetic trees have been classified, aligned, and merged. Taxonomies have been used to fill in sparse regions and gaps left by phylogenies. OTT is a base that has been little used for sequencing analyzes of the 16S region, however, it has a greater number of sequences classified taxonomically down to the genus level compared to SILVA and Greengenes. However, in terms of classification at the edge level, it contains a lesser amount of information [95]

Ribosomal Database Project

Ribosomal Database Project (RDP) [96] is a database that provides RNA ribosomal (rRNA) sequences of small subunits of domain bacterial and archaeal (16S); and fungal rRNA sequences of large subunits (28S). [97]

References

v

t

e

Differentiable programming

Information geometry

Statistical manifold

Automatic differentiation

Neuromorphic computing

Pattern recognition

Ricci calculus

Computational learning theory

Inductive bias

IPU

TPU

VPU

Memristor

SpiNNaker

TensorFlow

PyTorch

Keras

scikit-learn

Theano

JAX

Flux.jl

MindSpore

Portals Computer programming Technology

Computer programming

Technology