**BYTES AND MOLECULES - GLOSSARY**

Every living cell has its own **“hard drive”** where it stores bytes of information needed to function and grow, adapt to environmental changes and fight off dangerous infectious agents. This information is stored in molecules that are arranged in sequences of letters, referred to as the DNA code. In the 1990s, a tremendous effort by governments, researchers and companies resulted in the full sequencing of the human genome. This course is intended for all those interested in understanding how the current revolution in biomedical data impacts biomedical research and human health.

In this blog, we have listed all the terminologies and their meanings for your understanding, from the introductory lessons in the [**Bytes and Molecules**](https://learn.omicslogic.com/courses/course/course-2-bytes-and-molecules)course. Clicking on the lesson heading will redirect you back to the lesson where these terms are used.

[**01. Getting Started: Bytes and Molecules**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/01-getting-started-bytes-and-molecules)

Every living cell has its own “hard drive” where it stores bytes of information needed to function and grow, adapt to environmental changes and fight off dangerous infectious agents. In this lesson, you will learn about the role of data in biology and bioinformatics.

**Molecular data**

Molecular data (DNA or protein sequences) can be edited, manipulated, simulated and analyzed in various ways in Mesquite. Most of the features discussed elsewhere concerning editing and analysis of general categorical data also apply to molecular data; here we focus on features specifically designed for sequence data.

**Biology**

Biology is a branch of science that deals with living organisms and their vital processes. Biology encompasses diverse fields, including botany, conservation, ecology, evolution, genetics, marine biology, medicine, microbiology, molecular biology, physiology, and zoology.

**Informatics**

Informatics is the study of the structure, behaviour, and interactions of natural and engineered computational systems. ... The science of information and the engineering of information systems develop hand-in-hand. Informatics is the emerging discipline that combines the two.

**Genomics**

Genomics is the study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment.

**Transcriptomics**

Transcriptomics allows identification of genes and pathways that respond to and counteract biotic and abiotic environmental stresses. The non-targeted nature of transcriptomics allows the identification of novel transcriptional networks in complex systems.

**Proteomics**

Proteomics is the large-scale study of proteins. Proteins are vital parts of living organisms, with many functions. The proteome is the entire set of proteins produced or modified by an organism or system. Proteomics enables the identification of ever-increasing numbers of proteins.

**Metagenomics**

Metagenomics is the study of genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.

**Omics**

The word omics refers to a field of study in biological sciences that ends with -omics, such as genomics, transcriptomics, proteomics, or metabolomics. ... The metabolome represents the collection of all metabolites in a biological cell, tissue, organ, or organism, which are the end products of cellular processes.

**Bioinformatics**

Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology

**Research design**

The research design refers to the overall strategy that you choose to integrate the different components of the study in a coherent and logical way, thereby, ensuring you will effectively address the research problem; it constitutes the blueprint for the collection, measurement, and analysis of data.

**Data collection**

Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes. Data collection is a research component in all study fields, including physical and social sciences, humanities, and business.

**Analysis**

Analysis is the process of breaking a complex topic or substance into smaller parts in order to gain a better understanding of it. The technique has been applied in the study of mathematics and logic since before Aristotle, though analysis as a formal concept is a relatively recent development.

**Interpretation**

Interpretation is the act of explaining, reframing, or otherwise showing your own understanding of something. ... Interpretation requires you to first understand the piece of music, text, language, or idea, and then give your explanation of it.

[**02. Citizen Science**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/02-citizen-science)

Bioinformatics as a way to become a citizen scientist - going from a passive observer to a fascinated participant of scientific discovery! In this lesson, you will learn about genomes, citizen science activities and their role in biomedical research.

**X-ray diffraction methods**

XRD technique utilizes the X-ray scattering phenomenon to elucidate the crystal structure of crystalline/semi crystalline materials, with scattering of X-rays by periodic array of atoms giving rise to definite diffraction patterns that bestows a qualitative image of atomic arrangements within the crystal lattice.

[**03. Molecules, Data and Life**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/03-molecules-data-and-life)

The first section of Bytes and Molecules is dedicated to the basics of cellular and molecular biology and the way DNA code makes life possible. We also discuss the most familiar use-cases for DNA variation. In this lesson, you will learn about your DNA and get some hands on experience.

**Genotype**

A genotype is an individual's collection of genes. ... The genotype is expressed when the information encoded in the genes' DNA is used to make protein and RNA molecules. The expression of the genotype contributes to the individual's observable traits, called the phenotype.

**Phenotype**

A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.

[**04. Code of Life**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/04-code-of-life)

DNA has a double-stranded structure giving it a stable and therefore suitable molecular structure for information storage. The mechanism of DNA replication and the transcription of expressed genes depend on the code contained within a DNA molecule. In this lesson, you will learn about your DNA, mutations and cell differentiation.

**Replication**

DNA replication is the process by which a double-stranded DNA molecule is copied to produce two identical DNA molecules. Replication is an essential process because, whenever a cell divides, the two new daughter cells must contain the same genetic information, or DNA, as the parent cell.

**Transcription**

Transcription is the process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). ... The newly formed mRNA copies of the gene then serve as blueprints for protein synthesis during the process of translation.

**Substitution**

Substitution is a fundamental concept in logic. A substitution is a syntactic transformation of formal expressions. To apply a substitution to an expression means to constantly replace its variable, or placeholder, symbols by other expressions.

**Insertion**

The act or process of putting something into something else : the act or process of inserting something. : something (such as a comment) that is added to a piece of writing : something that is inserted.

**Deletion**

Deletion is a type of mutation involving the loss of genetic material. It can be small, involving a single missing DNA base pair, or large, involving a piece of a chromosome.

**Nucleotides**

A nucleotide is the basic building block of nucleic acids. ... A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base. The bases used in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T).

**Base pairs**

Two nitrogen-containing bases (or nucleotides) that pair together to form the structure of DNA. The four bases in DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). These bases form specific pairs (A with T, and G with C).

**Chromatin**

Chromatin is a complex of DNA and proteins that forms chromosomes within the nucleus of eukaryotic cells. ... Under the microscope in its expanded form, chromatin looks like beads on a string. The beads are called nucleosomes. Each nucleosome is composed of DNA wrapped around eight proteins called histones.

**Transcript**

A primary transcript is the single-stranded ribonucleic acid (RNA) product synthesized by transcription of DNA, and processed to yield various mature RNA products such as mRNAs, tRNAs, and rRNAs. The primary transcripts designated to be mRNAs are modified in preparation for translation.

**Cell differentiation**

Cell differentiation is the process by which dividing cells change their functional or phenotypical type. All cells presumably derive from stem cells and obtain their functions as they mature. Cellular composition is often modeled as a hierarchical scheme with stem cells at the top of the hierarchy.

**Transcription factors**

Transcription factors are proteins involved in the process of converting, or transcribing, DNA into RNA. Transcription factors include a wide number of proteins, excluding RNA polymerase, that initiate and regulate the transcription of genes. Regulation of transcription is the most common form of gene control.

**Receptors**

In biochemistry and pharmacology, receptors are chemical structures, composed of protein, that receive and transduce signals that may be integrated into biological systems.

[**05. Protein Structure and Function**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/05-protein-structure-and-function)

Learn about the DNA, nucleotide and amino acid variation, as well as the sequence - structure - function relationship. Basic terminology of DNA variation and its impact on protein structure. In this lesson, you will learn about protein structure and their function which includes

* Structural biology,
* Introduction to Structure & Function of Protein,
* Understanding the Protein Structure,
* Understanding various functions of Protein,
* Classification of Protein Structure, and
* Examples of Structure-Function relationships

**X-Ray Crystallography**

X-ray crystallography is the experimental science determining the atomic and molecular structure of a crystal, in which the crystalline structure causes a beam of incident X-rays to diffract into many specific directions. ... The goniometer is used to position the crystal at selected orientations.

**Nuclear Magnetic Resonance**

Nuclear magnetic resonance (NMR) is a physical phenomenon in which nuclei in a strong constant magnetic field are perturbed by a weak oscillating magnetic field (in the near field) and respond by producing an electromagnetic signal with a frequency characteristic of the magnetic field at the nucleus.

**Electron Microscopy**

Electron microscopy (EM) is a technique for obtaining high resolution images of biological and non-biological specimens. ... The transmission electron microscope is used to view thin specimens (tissue sections, molecules, etc) through which electrons can pass, generating a projection image.

**Regulatory proteins**

Any protein that influences the regions of a DNA molecule that are transcribed by RNA polymerase during the process of transcription. These proteins, which include transcription factors, therefore help control the synthesis of proteins in cells.

**Receptor proteins**

Receptors are a special class of proteins that function by binding a specific ligand molecule. When a ligand binds to its receptor, the receptor can change conformation, transmitting a signal into the cell. In some cases the receptors will remain on the surface of the cell and the ligand will eventually diffuse away.

**Structural proteins**

Protein structure is the three-dimensional arrangement of atoms in an amino acid-chain molecule. Proteins are polymers – specifically polypeptides – formed from sequences of amino acids, the monomers of the polymer. A single amino acid monomer may also be called a residue indicating a repeating unit of a polymer.

**Transfer proteins**

The non-specific lipid transfer proteins (LTPs) constitute a large protein family found in all land plants. They are small proteins characterized by a tunnel-like hydrophobic cavity, which makes them suitable for binding and transporting various lipids. LTPs are abundantly expressed in most tissues.

**Sickle-cell anemia**

Sickle cell anemia is one of a group of disorders known as sickle cell disease. Sickle cell anemia is an inherited red blood cell disorder in which there aren't enough healthy red blood cells to carry oxygen throughout your body. Normally, the flexible, round red blood cells move easily through blood vessels.

[**06. Introduction to Genomics**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/06-introduction-to-genomics)

In this lesson, you will learn about fundamentals of genomics: code and analysis of DNA code, sequence, structure and function , NGS and analysis of high throughput genomic data.

**Histones**

A type of protein found in chromosomes. Histones bind to DNA, help give chromosomes their shape, and help control the activity of genes. ... Most DNA is found inside the nucleus of a cell, where it forms the chromosomes. Chromosomes have proteins called histones that bind to DNA.

**Genotype**

A genotype is an individual's collection of genes. ... The genotype is expressed when the information encoded in the genes' DNA is used to make protein and RNA molecules. The expression of the genotype contributes to the individual's observable traits, called the phenotype.

**Phenotype**

A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.

**Human Genome Project**

The Human Genome Project was an international scientific research project with the goal of determining the base pairs that make up human DNA, and of identifying, mapping and sequencing all of the genes of the human genome from both a physical and a functional standpoint.

**Central dogma of molecular biology**

The central dogma of molecular biology explains the flow of genetic information, from DNA to RNA, to make a functional product, a protein. The central dogma suggests that DNA contains the information needed to make all of our proteins, and that RNA is a messenger that carries this information to the ribosomes

**Transcription**

Transcription is the process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). ... The newly formed mRNA copies of the gene then serve as blueprints for protein synthesis during the process of translation.

**Translation**

In molecular biology and genetics, translation is the process in which ribosomes in the cytoplasm or endoplasmic reticulum synthesize proteins after the process of transcription of DNA to RNA in the cell's nucleus. The entire process is called gene expression.

**DNA polymerase**

A DNA polymerase is a member of a family of enzymes that catalyze the synthesis of DNA molecules from nucleoside triphosphates, the molecular precursors of DNA. These enzymes are essential for DNA replication and usually work in groups to create two identical DNA duplexes from a single original DNA duplex.

**Genes**

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA. Some genes act as instructions to make molecules called proteins. However, many genes do not code for proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases.

**Genomics**

Genomics is the study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment.

**Substitution**

Substitution is a fundamental concept in logic. A substitution is a syntactic transformation of formal expressions. To apply a substitution to an expression means to constantly replace its variable, or placeholder, symbols by other expressions.

**Insertion**

The act or process of putting something into something else : the act or process of inserting something. : something (such as a comment) that is added to a piece of writing : something that is inserted.

**Deletion**

Deletion is a type of mutation involving the loss of genetic material. It can be small, involving a single missing DNA base pair, or large, involving a piece of a chromosome.

**Duplication**

Gene duplication is a major mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.

**Mutation**

A mutation is a change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses.

**Polymorphism**

A discontinuous genetic variation resulting in the occurrence of several different forms or types of individuals among the members of a single species. A discontinuous genetic variation divides the individuals of a population into two or more sharply distinct forms.

**Missense mutation**

A missense mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. It is a type of nonsynonymous substitution.

**Nonsense mutation**

A nonsense mutation, or its synonym, a stop mutation, is a change in DNA that causes a protein to terminate or end its translation earlier than expected. This is a common form of mutation in humans and in other animals that causes a shortened or nonfunctional protein to be expressed.

**Frameshift mutation**

A frameshift mutation is a genetic mutation caused by a deletion or insertion in a DNA sequence that shifts the way the sequence is read. A DNA sequence is a chain of many smaller molecules called nucleotides.

**SNV**

Single nucleotide variants (SNVs) occur when a single nucleotide (e.g., A, T, C, or G) is altered in the DNA sequence. SNVs are by far the most common type of sequence change, and there are a number of endogenous and exogenous sources of damage that lead to the single base pair substitution mutations that create SNVs.

**SNP**

Single nucleotide polymorphisms, frequently called SNPs (pronounced “snips”), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. ... Most commonly, these variations are found in the DNA between genes.

**CNV**

Copy number variation is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals. Copy number variation is a type of structural variation: specifically, it is a type of duplication or deletion event that affects a considerable number of base pairs.

**Next Generation Sequencing**

Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.

**Short reads**

To sequence a large genome like human DNA using NGS, the DNA has to be fragmented and amplified in clones of between 75 base pairs and 400 base pairs, hence the term 'short-read sequencing' (SRS). Computer programs are then used to assemble the random clones into a contiguous sequence.

[**NGS technology**](https://www.youtube.com/watch?v=jFCD8Q6qSTM)

Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.

[**FASTA**](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp)

In bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The format originates from the FASTA software package, but has now become a near universal standard in the field of bioinformatics.

[**FAST**](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp)**Q**

FASTQ format is a text-based format for storing both a biological sequence and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

**Sequence**

A sequence is an enumerated collection of objects in which repetitions are allowed and order matters. Like a set, it contains members (also called *elements*, or *terms*). The number of elements (possibly infinite) is called the *length* of the sequence. Unlike a set, the same elements can appear multiple times at different positions in a sequence, and unlike a set, the order does matter. Formally, a sequence can be defined as a function from natural numbers (the positions of elements in the sequence) to the elements at each position. The notion of a sequence can be generalized to an indexed family, defined as a function from an index set that may not be numbers to another set of elements.

**Mapping**

Genome mapping is used to identify and record the location of genes and the distances between genes on a chromosome. ... Sequenced DNA fragments can be aligned to the genome map to aid with the assembly of the genome. Over time, as scientists learn more about a particular genome, its map becomes more accurate and detailed.

**Whole genome**

Whole genome sequencing, also known as full genome sequencing, complete genome sequencing, or entire genome sequencing, is the process of determining the entirety, or nearly the entirety, of the DNA sequence of an organism's genome at a single time.

**Genome-wide association study**

A genome-wide association study (GWAS) is an approach used in genetics research to associate specific genetic variations with particular diseases. The method involves scanning the genomes from many different people and looking for genetic markers that can be used to predict the presence of a disease.

**Phylogenetic Analysis**

Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism.

**Huntington's disease**

Huntington's disease is a rare, inherited disease that causes the progressive breakdown (degeneration) of nerve cells in the brain. Huntington's disease has a broad impact on a person's functional abilities and usually results in movement, thinking (cognitive) and psychiatric disorders.

**Drug resistance**

Drug resistance is the reduction in effectiveness of a medication such as an antimicrobial or an antineoplastic in treating a disease or condition. The term is used in the context of resistance that pathogens or cancers have "acquired", that is, resistance has evolved.

**Apoptosis**

Apoptosis is the process of programmed cell death. It is used during early development to eliminate unwanted cells; for example, those between the fingers of a developing hand. In adults, apoptosis is used to rid the body of cells that have been damaged beyond repair.

**Cystic fibrosis**

Cystic fibrosis (CF) is an inherited disorder that causes severe damage to the lungs, digestive system and other organs in the body. Cystic fibrosis affects the cells that produce mucus, sweat and digestive juices. These secreted fluids are normally thin and slippery.

**Huntington's Disease**

Huntington's disease is a rare, inherited disease that causes the progressive breakdown (degeneration) of nerve cells in the brain. Huntington's disease has a broad impact on a person's functional abilities and usually results in movement, thinking (cognitive) and psychiatric disorders.

**Tuberculosis (TB)**

Tuberculosis (TB) is a potentially serious infectious disease that mainly affects the lungs. The bacteria that cause tuberculosis are spread from person to person through tiny droplets released into the air via coughs and sneezes.

**Viruses**

A virus is a small collection of genetic code, either DNA or RNA, surrounded by a protein coat. A virus cannot replicate alone. Viruses must infect cells and use components of the host cell to make copies of themselves. Often, they kill the host cell in the process, and cause damage to the host organism.

**Ebola virus**

Ebola Virus Disease (EVD) is a rare and deadly disease in people and nonhuman primates. The viruses that cause EVD are located mainly in sub-Saharan Africa. People can get EVD through direct contact with an infected animal (bat or nonhuman primate) or a sick or dead person infected with Ebola virus.

[**07. Introduction to Transcriptomics**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/07-introduction-to-transcriptomics)

Transcriptomics is the study of RNA sequences , copies of select regions of the genome transcribed in order to make proteins or carry out other functions. In this lesson, you will learn about transcriptomics, sequencing types, transcriptomic data and get introduced to machine learning for transcriptomics.

**Transcriptome**

The transcriptome is the set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells. The term can also sometimes be used to refer to all RNAs, or just mRNA, depending on the particular experiment.

**Microarray technologies**

The transcriptome is the set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells. The term can also sometimes be used to refer to all RNAs, or just mRNA, depending on the particular experiment.

**cDNA microarrays**

The cDNA microarray is the most powerful tool for studying gene expression in many different organisms. ... It is a high throughput, highly parallel RNA expression assay technique that permits quantitative analysis of RNAs transcribed from both known and unknown genes.

**Oligo DNA microarrays**

An oligo DNA microarray is a DNA microarray whose probe is the chemically synthesized oligo DNA after genes corresponding to exons in the region of 20-100 nts are selected. Since the probe is shorter than a cDNA microarray, its specificity is high and cross hybridization can be inhibited. cDNA microarrays.

**BAC microarrays**

A BAC clone chip is a DNA microarray whose probe is a template amplified by PCR. The template is a genome region incorporated into a comprehensive BAC (bacterial artificial chromosome) clone which various research institutions used in decoding the genome sequence of various organisms.

**SNP microarrays**

SNP microarray is the hybridization of fragmented single-stranded DNA to arrays containing hundreds of thousands of unique nucleotide probe sequences. ... Specialized equipment is then used to produce a measure of the signal intensity associated with each probe and its target after hybridization.

**High throughput sequencing**

High-throughput sequencing, also known as next-generation sequencing (NGS), is the comprehensive term used to describe technologies that sequence DNA and RNA in a rapid and cost-effective manner.

**Deep sequencing**

Deep sequencing refers to sequencing a genomic region multiple times, sometimes hundreds or even thousands of times. This next-generation sequencing (NGS) approach allows researchers to detect rare clonal types, cells, or microbes comprising as little as 1% of the original sample.

**Solexa / Illumina Sequencing**

The Solexa/Illumina sequencing method is similar to Sanger sequencing, but it uses modified dNTPs containing a terminator which blocks further polymerization- so only a single base can be added by a polymerase enzyme to each growing DNA copy strand.

**Roche 454 Sequencing**

Roche 454 sequencing can sequence much longer reads than Illumina. Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added. As in Illumina, the DNA or RNA is fragmented into shorter reads, in this case up to 1kb.

**Ion Torrent Sequencing**

Ion semiconductor sequencing is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA. This is a method of "sequencing by synthesis", during which a complementary strand is built based on the sequence of a template strand.

**ABI SOLiD Sequencing**

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a next-generation DNA sequencing technology developed by Life Technologies and has been commercially available since 2006. This next generation technology generates 108 - 109 small sequence reads at one time. It uses 2 base encoding to decode the raw data generated by the sequencing platform into sequence data.

**Omics data**

Omics technologies are defined as high-throughput biochemical assays that measure comprehensively and simultaneously molecules of the same type from a biological sample. For example, genomics profile DNA, transcriptomics measure transcripts; proteomics and metabolomics quantify proteins and metabolites, respectively.

**Next Generation Sequencing**

Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.

**mRNA expression**

When the mRNA sequence is read, each tRNA molecule delivers its amino acid to the ribosome and binds temporarily to the corresponding codon on the mRNA molecule. Once the tRNA is bound, it releases its amino acid and the adjacent amino acids all join together into a long chain called a polypeptide.

**PCA (Principal Component Analysis)**

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

**Machine learning**

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

**Hierarchical clustering**

Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

**K-means**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

**Agglomerative**

Agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

**Quasi-sample**

Under certain conditions, largely governed by the method of compiling the sampling frame or list, a systematic sample of every nth entry from a list will be equivalent for most practical purposes to a random sample. This method of sampling is sometimes referred to as quasi-random sampling.

**K-means**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

**Centroids**

the centroid or geometric center of a plane figure is the arithmetic mean position of all the points in the figure. ... If a physical object has uniform density, then its center of mass is the same as the centroid of its shape.

**Classification**

Classification means arranging or sorting objects into groups on the basis of a common property that they have. ... For example, you can classify the apples in one category, the bananas in another, and so on.

**Decision tree**

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

**Random Forest**

A random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms. This algorithm is applied in various industries such as banking and e-commerce to predict behavior and outcomes.

**LDA**

Linear discriminant analysis is primarily used here to reduce the number of features to a more manageable number before classification. Each of the new dimensions is a linear combination of pixel values, which form a template.

**SVM**

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. ... The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features.

[**08. Introduction to Metagenomics**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/08-introduction-to-metagenomics)

Explore how different types of analysis can be used to understand the microbial world and practice several bioinformatics approaches to the analysis of the microbiome. In this lesson, you will get to understand the microbiome, tools and technologies to study microorganisms and get introduced to metagenomics.

**Microbiology**

Microbiology is the study of microscopic organisms, such as bacteria, viruses, archaea, fungi and protozoa. This discipline includes fundamental research on the biochemistry, physiology, cell biology, ecology, evolution and clinical aspects of microorganisms, including the host response to these agents.

**Unicellular**

A unicellular organism, also known as a single-celled organism, is an organism that consists of a single cell, unlike a multicellular organism that consists of multiple cells. These organisms live together, and each cell must carry out all life processes to survive.

**Multicellular**

having, made up of, or involving more than one and usually many cells especially of living matter It is probable that with a few exceptions all the cells in a multicellular organism have the same genetic information encoded in the chains of nucleotide bases that make up their DNA.

**Pathogens**

A pathogen is defined as an organism causing disease to its host, with the severity of the disease symptoms referred to as virulence. Pathogens are taxonomically diverse and comprise viruses and bacteria as well as unicellular and multicellular eukaryotes.

**Microbiome**

The microbiome consists of microbes that are both helpful and potentially harmful. Most are symbiotic (where both the human body and microbiota benefit) and some, in smaller numbers, are pathogenic (promoting disease). In a healthy body, pathogenic and symbiotic microbiota coexist without problems.

**Bacteriology**

Bacteriology is the branch and specialty of biology that studies the morphology, ecology, genetics and biochemistry of bacteria as well as many other aspects related to them. This subdivision of microbiology involves the identification, classification, and characterization of bacterial species.

**Mycology**

Mycology is the branch of biology concerned with the study of fungi, including their genetic and biochemical properties, their taxonomy and their use to humans as a source for tinder, traditional medicine, food, and entheogens, as well as their dangers, such as toxicity or infection.

**Virology**

Virology is the scientific discipline concerned with the study of the biology of viruses and viral diseases, including the distribution, biochemistry, physiology, molecular biology, ecology, evolution and clinical aspects of viruses.

**Phycology**

Phycology is the scientific study of algae. Also known as algology, phycology is a branch of life science. Algae are important as primary producers in aquatic ecosystems. Most algae are eukaryotic, photosynthetic organisms that live in a wet environment.

**Nematology**

Nematology is the scientific discipline concerned with the study of nematodes, or roundworms. Although nematological investigation dates back to the days of Aristotle or even earlier, nematology as an independent discipline has its recognizable beginnings in the mid to late 19th century.

**Protozoology**

Protozoology is the study of protozoa, the "animal-like" (i.e., motile and heterotrophic) protists. The Protozoa are considered to be a sub-kingdom of Protista. They are free-living organisms that are found in almost every habitat. ... Example: Toxoplasmosis and giardiasis are diseases caused by protozoa.

**Parasitology**

Parasitology is the scientific discipline concerned with the study of the biology of parasites and parasitic diseases, including the distribution, biochemistry, physiology, molecular biology, ecology, evolution and clinical aspects of parasites, including the host response to these agents.

**Dysbiosis**

Dysbiosis is a condition when the gut bacteria become imbalanced. As a result, a wide range of digestive disturbance symptoms occurs, including diarrhea, cramping, constipation, bloating, and indigestion.

**Bacterial culture**

A microbiological culture, or microbial culture, is a method of multiplying microbial organisms by letting them reproduce in a predetermined culture medium under controlled laboratory conditions. ... Microbial cultures are used to determine the type of organism, its abundance in the sample being tested, or both.

**Gram staining**

A Gram stain is a test that checks for bacteria at the site of a suspected infection such as the throat, lungs, genitals, or in skin wounds. ... When the stain combines with bacteria in a sample, the bacteria will either stay purple or turn pink or red. If the bacteria stays purple, they are Gram-positive.

**Metagenomics**

Metagenomics is the study of genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.

**Genomic analysis**

Genomic analysis is the identification, measurement or comparison of genomic features such as DNA sequence, structural variation, gene expression, or regulatory and functional element annotation at a genomic scale.

**Sequence-based Metagenomics**

Sequence-based metagenomics involves sequencing and analysis of DNA from environmental samples. Sequence-based metagenomics studies can be used to assemble genomes, identify genes, find complete metabolic pathways, and compare organisms of different communities.

**Function-based Metagenomics**

Functional metagenomics involves isolating DNA from microbial communities to study the functions of encoded proteins. ... Using this function-based approach allows for discovery of novel enzymes whose functions would not be predicted based on DNA sequence alone.

**Metatranscriptomic analysis**

Metatranscriptomics analysis enables understanding of how the microbiome responds to the environment by studying the functional analysis of genes expressed by the microbiome. It can also estimate the taxonomic composition of the microbial population.

[**09. Bioinformatics and Precision Medicine**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/09-bioinformatics-and-precision-medicine)

Precision Medicine is changing the way we understand, diagnose and treat cancer. The transformation is driven by high-throughput molecular data from patients, animal models, and large-scale cell line experiments. In this lesson, you will learn about precision medicine and the role of bioinformatics for Precision Medicine.

**Precision Medicine**

Precision medicine is a medical model that proposes the customization of healthcare, with medical decisions, treatments, practices, or products being tailored to a subgroup of patients, instead of a one‐drug‐fits‐all model.

**Phenylketonuria**

PKU is caused by a defect in the gene that helps create the enzyme needed to break down phenylalanine. Without the enzyme necessary to process phenylalanine, a dangerous buildup can develop when a person with PKU eats foods that contain protein or eats aspartame, an artificial sweetener.

**Sickle cell anemia**

Sickle cell anemia is one of a group of disorders known as sickle cell disease. Sickle cell anemia is an inherited red blood cell disorder in which there aren't enough healthy red blood cells to carry oxygen throughout your body. Normally, the flexible, round red blood cells move easily through blood vessels.

**Down syndrome**

Down syndrome is a condition in which a person has an extra chromosome. Chromosomes are small “packages” of genes in the body. They determine how a baby's body forms and functions as it grows during pregnancy and after birth. Typically, a baby is born with 46 chromosomes.

**DNA methylation**

DNA methylation is a biological process by which methyl groups are added to the DNA molecule. Methylation can change the activity of a DNA segment without changing the sequence. When located in a gene promoter, DNA methylation typically acts to repress gene transcription.

**Pharmacogenomics**

Pharmacogenomics is the study of how genes affect a person's response to drugs. This relatively new field combines pharmacology (the science of drugs) and genomics (the study of genes and their functions) to develop effective, safe medications and doses that will be tailored to a person's genetic makeup.

**Precision Oncology**

Precision oncology, defined as molecular profiling of tumors to identify targetable alterations, is rapidly developing and has entered the mainstream of clinical practice.

**Clinical matching**

Clinical trial matching services facilitate patient enrollment in clinical trials by identifying potential trials for interested patients and their proxies (e.g., caregivers and providers), and in some cases by providing other support services such as educational materials or personnel who can answer questions or assist patients. While all services involve providing a list of trials to those using them, the actual goals of the services vary significantly, from largely fostering patient interest in clinical trials by showing them the breadth of trials open for their cancer, to full matching and enrollment assistance using detailed assessment of eligibility criteria, patient clinical data, and patient preferences.

**High-throughput molecular data**

“High-throughput data”, the information generated in a massive, fast manner by 'omics' technologies - transcriptomics, metabolomics and proteomics - have opened a new era in biomedical research allowing an exponential increase of biomedical discoveries.

**Bioinformatics**

Bioinformatics is defined as the application of tools of computation and analysis to the capture and interpretation of biological data. It is an interdisciplinary field, which harnesses computer science, mathematics, physics, and biology

**Informatics**

Informatics is the study of the structure, behaviour, and interactions of natural and engineered computational systems. ... It has computational, cognitive and social aspects. The central notion is the transformation of information - whether by computation or communication, whether by organisms or artifacts.

**Data science**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

**Multiple omics**

Multiomics, multi-omics, integrative omics, "panomics" or 'pan-omics' is a biological analysis approach in which the data sets are multiple "ones", such as the genome, proteome, transcriptome, epigenome, metabolome, and microbiome

[**10. Space Omics and Astronaut Health**](https://learn.omicslogic.com/Learn/course-2-bytes-and-molecules/lesson/10-space-omics-and-astronaut-health)

Astrobiology is a science which addresses the origin of life on Earth and potential for life elsewhere in the universe. Space Omics is dedicated to omics data to study molecular mechanisms affected by space travel & critical for astronaut health. In this lesson, you will learn about the effects of spaceflight on astronaut health, get an overview of omics for space, simulated vs real spaceflight and data repositories for Space Omics

**Gene expression**

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product that enables it to produce end products, protein or non-coding RNA, and ultimately affect a phenotype, as the final effect.