# BIOCHEMICAL ENGINEERING *and* BIOINFORMATICS

**PRANAV** KUMAR
FORMER FACULTY,
DEPARTMENT OF BIOTECHNOLOGY,
JAMIA MILLIA ISLAMIA (JMI),
NEW DELHI, INDIA

**Pathfinder** Publication
A unit of **Pathfinder Academy Pvt. Ltd.**, Delhi

**PRANAV** KUMAR
FORMER FACULTY,
DEPARTMENT OF BIOTECHNOLOGY,
JAMIA MILLIA ISLAMIA (JMI),
NEW DELHI, INDIA



ISBN: 978-81-95633-33-3 (paperback)

Copyright © **2026** by Pathfinder Publication, all rights reserved.

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reproduced by any mechanical, photographic, or electronic process, or in the form of a phonographic recording, nor it may be stored in a retrieval system, transmitted, or otherwise copied for public or private use, without written permission from the publisher.

**Pathfinder** Publication
A unit of **Pathfinder Academy** Pvt. Ltd., New Delhi, India.
**pathfinder**publication.in
09350208235

# Preface

The rapid advancement of biotechnology and computational biology has transformed the landscape of the life sciences, making Bioprocess engineering and Bioinformatics indispensable components of modern biological research. For aspirants preparing for the UGC-CSIR-DBT NET – Life Sciences and Biotechnology examinations, as well as GAT-B, these subjects require not only conceptual clarity but also strong analytical and problem-solving skills. This book has been carefully designed to meet the specific requirements of UGC-CSIR-DBT NET aspirants. It provides a comprehensive, exam-oriented, and student-friendly approach to mastering these subjects.

The **Bioprocess engineering** section covers core topics such as bioenergetics, biochemical engineering principles, microbial and enzymatic kinetics, bioprocess engineering, scale-up strategies, enzyme technology, transport phenomena in bioprocessing, bioprocess technology, and metabolic engineering with synthetic biology. Emphasis has been placed on developing a clear understanding of fundamental principles, mathematical derivations, and numerical problem-solving, all of which are essential for success in competitive examinations.

The **Bioinformatics** section introduces students to the fundamentals of biological data, followed by detailed discussions on sequence analysis, gene annotation, molecular phylogenetics, genomics, transcriptomics, proteomics, and computer-aided drug design. The section concludes with emerging trends in bioinformatics, highlighting recent developments such as systems biology, high-throughput data analysis, and the growing role of computational tools in modern biology. Complex algorithms and concepts are explained in a simplified, step-by-step manner to ensure accessibility for students from diverse academic backgrounds.

Throughout the book, special care has been taken to:

- Present concepts in a clear, logical, and concise manner
- Integrate theory with numerical problems and applications
- Highlight key formulas, definitions, and exam-focused points
- Support learning with diagrams, flowcharts, and summaries

This book is intended to serve as a complete learning resource as well as a reliable revision guide for UGC-CSIR-DBT NET and GAT-B aspirants. It will also be beneficial for postgraduate students and early researchers seeking a consolidated understanding of bioprocess engineering and bioinformatics. A book of this nature demands meticulous and dedicated efforts from all its contributors. I extend our heartfelt gratitude to each of them, with special thanks to Pramod Mittal and Ajay Kumar for their valuable contributions. I am particularly grateful to Pradeep Verma for their exceptional work in design, layout, and organization, which played a crucial role in bringing this book to completion.

***Pranav Kumar***

# Contents


## Chapter 1
## Biochemical engineering

1.1 Bioenergetics 1

1.2 Biochemical engineering principles 4

1.3 Bioprocess engineering 16

1.4 Scale up strategies 29

1.5 Enzyme technology 30

1.6 Transport phenomena in bioprocessing 35

1.7 Bioprocess technology 41

1.8 Metabolic engineering and synthetic biology 52

## Chapter 2
## Bioinformatics

2.1 Biological data basics 69

2.2 Sequence analysis 71

2.3 Gene annotation 74

2.4 Molecular phylogenetics 74

2.5 Genomics and Transcriptomics 78

2.6 Proteomics 79

2.7 Drug design 81

2.8 Emerging trends in bioinformatics 82

# 1



# Biochemical engineering

SUBJECT FOCUS


| | | | |
|---|---|---|---|
| 1.1 | Bioenergetics | 1.5 | Enzyme technology |
| 1.2 | Biochemical engineering principles | 1.6 | Transport phenomena in bioprocessing |
| 1.3 | Bioprocess engineering | 1.7 | Bioprocess technology |
| 1.4 | Scale up strategies | 1.8 | Metabolic engineering and synthetic biology |


**Biochemical engineering** is an interdisciplinary field that applies chemical engineering principles to biological systems and processes. The term **bioprocess engineering** is often used interchangeably with biochemical engineering, especially when the focus is on the design, development, and optimization of processes that utilize living cells or their components—such as enzymes and organelles—for industrial, medical, agricultural, or environmental applications.

Key applications of *bioprocess engineering* include:

- Designing and operating fermentation and bioreactor systems.
- Developing and improving food and pharmaceutical processing technologies.
- Applying and optimizing product separation and purification techniques.
- Designing instrumentation and control systems for monitoring biological processes.
- Scaling up laboratory processes to industrial production levels.

**Bioprocess engineers** work at the intersection of biology and engineering, transforming raw biological materials into valuable products that benefit society—such as vaccines, biofuels, enzymes, and food ingredients. They play a crucial role in advancing biotechnology and ensuring sustainable production in various sectors.

## 1.1 Bioenergetics

A living cell is a dynamic chemical system. It requires energy to perform cellular work. Cells extract, transform, and use energy. The ability to harness the energy and channel it into biological work is a fundamental property of all living organisms. It is believed that an evolved and the adapted biological system converts energy in the most efficient manner. **Bioenergetics** is the study of energy flow in living systems, grounded in the *principles of thermodynamics* and *redox chemistry*. **Thermodynamics** explains how biological processes are energetically feasible and how energy is transformed and conserved, while **redox reactions** serve as mechanisms for energy transfer within cells.

## Laws of thermodynamics and their relevance to biological processes

Living cells function as dynamic biochemical factories, where several reactions occur in a highly coordinated manner. Despite the complexity and organization of biological systems, all cellular processes strictly obey the fundamental laws of thermodynamics.

**Thermodynamics**—the science of energy transformations and their relationship to matter—is central to understanding how cells manage energy for survival, growth, and function. These principles govern every physical and chemical event within living organisms, from metabolism and molecular transport to biosynthesis and energy conversion. Although life appears highly ordered and intricate, it operates within the constraints of thermodynamic laws. Gaining a deep understanding of these laws is essential for comprehending how living systems maintain order, harness energy, and interact with their environment.

**First law of thermodynamics** (**Law of conservation of energy**): This law states that energy cannot be created or destroyed—only converted from one form to another.

*Relevance to biology*: In living organisms, energy is transformed into various forms required for life processes. For example, photosynthesis converts light energy into chemical energy, while cellular respiration converts the chemical energy stored in glucose into ATP. During these transformation, total amount of energy remains constant.

**Second law of thermodynamics** (**Law of increasing entropy**): This law states that the total entropy (disorder or randomness) of an isolated system can only increase over time or remain constant in ideal equilibrium conditions. Spontaneous processes tend to increase entropy.

*Relevance to biology*: Although living organisms are highly ordered (low entropy), they maintain this order by constantly expending energy and increasing the entropy of their surroundings. For instance, the synthesis of proteins from amino acids reduces entropy within the cell but is coupled with energy-releasing processes that increase the overall entropy of the environment. The heat released during metabolism also contributes to increased environmental entropy.

### Cell as a thermodynamic system

A **system** refers to the specific portion of the universe selected for analysis—such as a living cell or a bioreactor. Everything outside the system is termed the **surroundings**. The system boundary—which separates the system from its surroundings—can be real or imaginary. Based on the exchange of mass across this boundary, systems are classified as:

**Closed system**: A *closed system* is a system in which energy can be exchanged with the surroundings, but matter (mass) cannot cross the system boundary. Thus, the amount of matter within the system remains constant, although heat or work may be transferred into or out of the system.

**Open system**: An *open system* is a system in which both matter and energy can be exchanged with the surroundings. Most biological and ecological systems are open systems because they continuously exchange materials and energy with their environment.

Cells, in most biological contexts, are considered *open systems* because they continuously exchange matter (nutrients, gases, waste) and energy with their environment. Understanding this is essential for analyzing cellular processes through thermodynamic perspectives.

## Free energy changes in biological reactions

**Gibbs free energy** quantifies the amount of energy available to do useful work at constant temperature and pressure. The *change in Gibbs free energy* ($\Delta G$) determines the *spontaneity* of a reaction:

- $\Delta G < 0$ (Negative): The reaction is *spontaneous* (**exergonic**) and releases free energy.
- $\Delta G > 0$ (Positive): The reaction is *non-spontaneous* (**endergonic**) and requires an input of energy.
- $\Delta G = 0$: The reaction is at *equilibrium*, with no net change in the concentrations of reactants and products.

substrates, then $\mu = \mu_m$, and the culture is in the *log phase* at its maximal growth rate. The specific growth rate is independent of substrate concentration as long as the excess substrate is present.

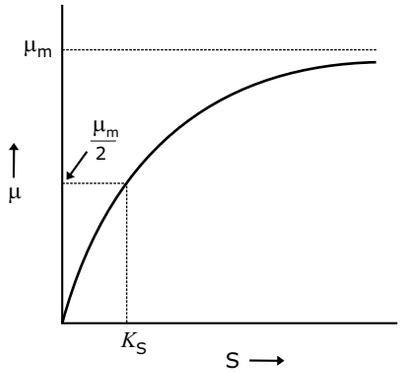



**Figure 1.2**
Dependence of the specific growth rate, on the substrate concentration. Note that μ approaches $\mu_m$ when S >> $K_S$ and becomes independent of substrate concentration. When S << $K_S$, the specific growth rate is very sensitive to the substrate concentration, exhibiting a first-order dependence.

If the concentration of one growth rate-limiting substrate is varied, while the other components are kept constant, the growth curves to substrate concentration are typically hyperbolic. Taking the reciprocal values in the Monod equation and rearranging the equation,

$$\frac{1}{\mu} = \frac{K_S + S}{\mu_m . S}$$

$$\frac{1}{\mu} = \frac{K_S}{\mu_m . S} + \frac{S}{\mu_m . S}$$

$$\frac{1}{\mu} = \frac{K_S}{\mu_m} \cdot \frac{1}{S} + \frac{1}{\mu_m}$$

The plot of $1/\mu$ against 1/[S] produces a *straight line* with intercept on the Y axis at $1/\mu_m$ and slope equal to $K_S/\mu_m$. The Monod equation is by far the most frequently-used expression relating growth rate to substrate concentration. However, it is valid only for balanced growth and should not be applied when growth conditions are changing rapidly.

*Monod equation* can be expressed in terms of concentration of cell mass (X) also.

$$\frac{dX}{dt} = \mu . X = \frac{\mu_m . S}{K_S + S} \cdot X$$

The Monod equation can also be expressed as a function of substrate utilization. We can relate changes in substrate concentration to change in 'X' through the biomass yield coefficient ($Y_{X/S}$).

$$\frac{dS}{dt} = \frac{1}{Y_{X/S}} \frac{dX}{dt} = \frac{1}{Y_{X/S}} \frac{\mu_m . S}{K_S + S} \cdot X$$

$Y_{X/S}$ is the **growth yield coefficient** (or **biomass yield coefficient**). It is defined as the unit amount of cell biomass produced per unit amount of substrate consumed. Thus, the more efficiently a substrate is degraded, the higher the value of the cell biomass yield coefficient.

## 1.3 Bioprocess engineering

**Bioprocess engineering** is the field that applies engineering principles to design, optimize, and scale up biological processes—like fermentation or cell culture—to produce useful biological products at large scale. One of the most important components of bioprocess engineering is the design of *bioreactors*, which provide controlled conditions for efficient biological production.

A **bioreactor**—also referred to as a **biochemical reactor**, **biological reactor**, **fermenter**, or **microbial reactor**—is a carefully engineered vessel designed to cultivate biological agents such as microorganisms (e.g., bacteria and fungi), animal cells, or plant cells under controlled conditions. The primary objectives are either to increase cell biomass or to produce specific target products such as enzymes, metabolites, antibodies, biofuels, or pharmaceuticals.

# 2



# Bioinformatics

SUBJECT FOCUS


2.1 Biological data basics
2.2 Sequence analysis
2.3 Gene annotation
2.4 Molecular phylogenetics
2.5 Genomics and Transcriptomics
2.6 Proteomics
2.7 Drug design
2.8 Emerging trends in bioinformatics


Bioinformatics is an interdisciplinary field that merges biology, computer science, information technology, and mathematics. It is defined as the collection, analysis, and dissemination of genetic and biological data to support the global research community. The primary goal of bioinformatics is to integrate and analyze vast amounts of biological data to assign functions to biological entities. This discipline finds applications in various areas, including:

**Genome mapping**: Creating genetic and physical maps of genomes.

**Gene discovery**: Identifying genes and determining their roles.

**Protein function** and **structure**: Inferring molecular functions and predicting the 3D structures of proteins.

**Genotype-phenotype correlation**: Understanding how gene variations influence phenotypic traits.

**Pathway reconstruction**: Reconstructing molecular interaction networks and signal transduction pathways.

**Biological system simulation**: Simulating complex biological processes to study their behavior under different conditions.

## 2.1 Biological data basics

Bioinformatics is about exploring biological information. This information is kept safely in databases. A **database** consists of an organized collection of data that provides a standardized way of locating, adding, and changing data.

The development of biological databases began with the availability of the **insulin protein** sequence in 1956, marking insulin—comprising 51 amino acid residues—as the first protein to be sequenced. This milestone laid the foundation for the systematic collection of biomolecular data. As techniques for determining protein three-dimensional structures advanced, dedicated resources for structural information were established, most notably the **PDB**, which became the first widely recognized repository for protein structural data.

### Database classification

Biological databases can be broadly classified based on the nature of the data they contain or the way in which the data is generated. Two common classification schemes are:

**By data type**: Databases can be categorized into two types: *sequence databases* and *structure databases*. **Sequence databases** store nucleic acid or protein sequences, while **structure databases** contain three-dimensional biomolecular structures. Nucleic acids are represented as linear sequences of nucleotides, whereas proteins may be represented as amino acid sequences or as three-dimensional structural coordinates. *Sequence data* provide one-dimensional information useful for tasks such as gene identification, evolutionary analysis, and functional annotation, while *structural data* capture the spatial arrangement of atoms, offering insight into a biomolecule's properties, stability, and functional mechanisms.

**By data origin**: Based on their origin, data are categorized into *primary* and *secondary databases*. **Primary databases** contain experimentally derived, raw biological data. This includes information obtained directly from laboratory experiments, such as nucleotide sequences, protein sequences, and three-dimensional macromolecular structures. These databases serve as archival repositories and generally do not alter the submitted data except for formatting and basic validation. **Secondary databases** contain derived or curated information, generated by analyzing, interpreting, or annotating data from primary databases. They often integrate computational analyses, such as multiple sequence alignments, conserved motif identification, or domain predictions, to provide functional and structural insights. For example, a secondary protein sequence database may store signature motifs, active-site residues, or conserved domains characteristic of a particular protein family—information obtained by comparing a set of related sequences.

## Types of biological data

Biological data are broadly classified into genomic, transcriptomic, proteomic, metabolomic, and structural categories.

**Genomic data** comprise nucleotide sequences derived from the genomes of organisms, including DNA and cDNA sequences. These may be complete genomes, partial sequences, or transcriptomic datasets such as RNA-Seq reads. Genomic data enable the study of gene structure, function, regulation, and evolution.

**Proteomic data** represent the protein complement of a cell, tissue, or organism. This includes the amino acid sequences of proteins, their abundance levels under specific conditions, post-translational modifications, and interaction partners. Proteomic data are typically derived from mass spectrometry, two-dimensional gel electrophoresis, or computational predictions.

**Metabolomic data** describe the complete set of metabolites (small molecules) within a biological sample. These data provide insight into metabolic pathways and physiological states, often obtained using techniques such as nuclear magnetic resonance spectroscopy or chromatography-coupled mass spectrometry.

**Structural data** detail the three-dimensional architecture of biomolecules, primarily proteins and nucleic acids. Such data are crucial for understanding molecular function and for structure-based drug design. They are often generated by X-ray crystallography, cryo-electron microscopy, or NMR spectroscopy.

## File formats

The storage, representation, and exchange of biological data require standardized file formats to ensure compatibility and interoperability among different software tools, databases, and bioinformatics pipelines.

The **FASTA** format is one of the simplest and most widely used file formats for storing nucleotide or protein sequences. Each record begins with a header line starting with the symbol '>', followed by a sequence identifier and an optional description. The subsequent lines contain the nucleotide or amino acid sequence.

The **FASTQ** format extends the FASTA structure by including per-base quality scores, which are essential for evaluating the reliability of high-throughput sequencing reads. Each record consists of four lines: a header, the sequence, a separator line, and a string of quality score characters.

The **GFF** (General Feature Format) is used to describe genomic features such as genes, exons, coding sequences, regulatory elements, and other sequence annotations. It is a tab-delimited text format that specifies the genomic coordinates of each feature, its type, strand orientation, and additional annotation attributes.

*Other file formats*: Several specialized file formats are also widely used in bioinformatics. **PDB** files store the three-dimensional structures of proteins and other biological macromolecules. **SAM** is a text-based format used to store sequence alignments, while **BAM** is its compressed binary version. **VCF** is used to store genomic sequence variants, including single nucleotide polymorphisms, insertions, deletions, and other genetic variations.

### Data sources and Repositories

Numerous public repositories play a crucial role in curating and disseminating biological data to the global research community. These repositories not only archive raw datasets but also provide curated annotations, cross-references, and a suite of analysis tools to facilitate research.

#### Protein databases

**PDB** (Protein Data Bank): Repository of experimentally determined 3D structures of proteins, nucleic acids, and their complexes solved using X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM).

**UniProtKB** (Universal Protein Knowledgebase): Provides comprehensive protein sequences and functional annotations. It comprises Swiss-Prot (manually curated data) and TrEMBL (automatically annotated data).

**SCOP** (Structural Classification of Proteins) and CATH (Class, Architecture, Topology, Homologous superfamily): Databases that classify proteins based on structural and evolutionary relationships.

**AlphaFold protein structure database**: Houses AI-predicted 3D protein structures generated using deep learning models.

**Pfam**: A collection of protein families, domains, and conserved motifs with functional and evolutionary annotations.

#### Nucleic acid databases

**GenBank** (NCBI): One of the largest publicly available repositories of DNA and RNA sequences.

**EMBL-EBI nucleotide sequence database**: European archive of nucleotide sequences integrated with international databases.

**DDBJ** (DNA Data Bank of Japan): Collaborates with GenBank and ENA as part of the International Nucleotide Sequence Database Collaboration (INSDC).

**Rfam**: Specialized database of RNA families, encompassing noncoding RNAs and their conserved secondary structures.

**miRBase**: Focused repository for microRNA sequences and annotation data.

#### Integrated Databases (Proteins and Nucleic Acids)

**NCBI** (National Center for Biotechnology Information): Provides integrated access to GenBank, RefSeq, PubChem, BLAST, and numerous other bioinformatics resources.

**Ensembl**: A genome browser offering detailed DNA, RNA, and protein annotations across diverse species.

**KEGG** (Kyoto Encyclopedia of Genes and Genomes): Connects genes and proteins to metabolic pathways and functional systems.

## 2.2 Sequence analysis

*Sequence analysis* is a fundamental component of bioinformatics, concerned with the examination, comparison, and interpretation of biological sequences—primarily DNA, RNA, and proteins. It enables researchers to identify functional elements, infer evolutionary relationships, and predict structural and biochemical properties. Sequence analysis generally rely on sequence alignment, similarity searches, and statistical scoring systems to assess relatedness between sequences.

### DNA and Protein sequence alignment

Sequence alignment involves arranging two or more sequences to identify regions of similarity that may indicate functional, structural, or evolutionary relationships. DNA sequence alignment compares nucleotide sequences, whereas protein sequence alignment compares amino acid sequences. Protein sequence alignment offers greater evolutionary insight due to the degeneracy of the genetic code.