# Bioinformatics Analysis: Evolutionary Conservation of Human Insulin (P01308)

**Student Name:** Muhammad Tayyab Alqan

**Student ID:** CA/SE1/6617

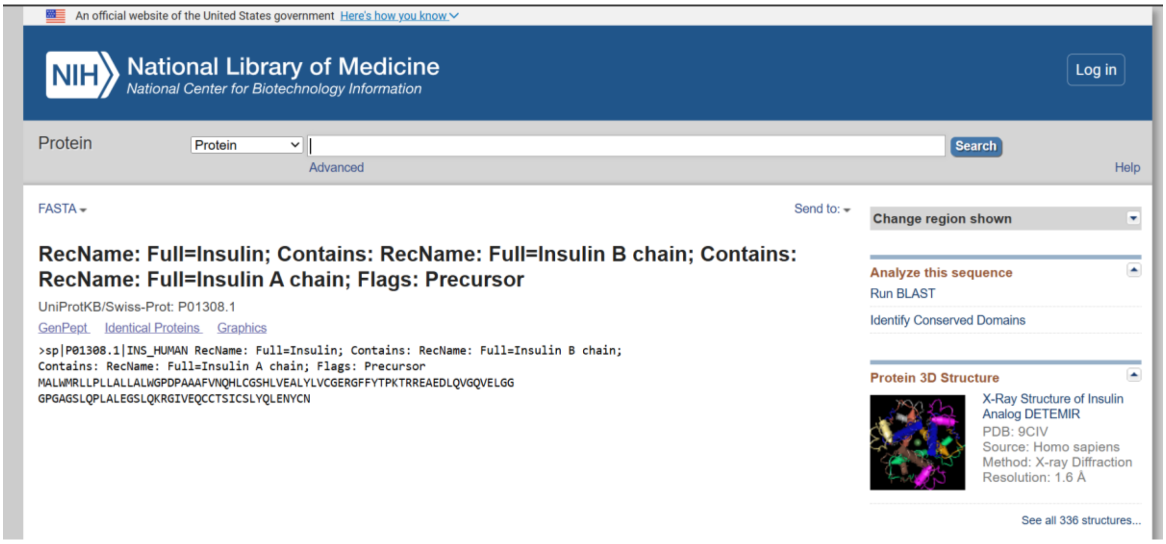
**Date:** December 19, 2025

## 1. Sequence Acquisition & Initial Characterization

The study began with the retrieval of the primary amino acid sequence for the human insulin preproprotein, identified by the **UniProt** accession **P01308.**

Sequence Summary:

- ***Length:*** **110 amino acids.**
- ***Composition:*** **The sequence represents the full precursor molecule, containing the B-chain, C-peptide, and A-chain.**
- ***Source:*** **National Center for Biotechnology Information (NCBI) / UniProtKB.**


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
Protein
Protein
Search
Advanced
Help
FASTA
Send to:
Change region shown
Analyze this sequence
Run BLAST
Identify Conserved Domains
Protein 3D Structure
See all 336 structures...

RecName: Full=Insulin; Contains: RecName: Full=Insulin B chain; Contains: RecName: Full=Insulin A chain; Flags: Precursor

UniProtKB/Swiss-Prot: P01308.1

GenPept Identical Proteins Graphics

```
>sp|P01308.1|INS_HUMAN RecName: Full=Insulin; Contains: RecName: Full=Insulin B chain;
Contains: RecName: Full=Insulin A chain; Flags: Precursor
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGG
GPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
```

X-Ray Structure of Insulin Analog DETEMIR
PDB: 9CIV
Source: Homo sapiens
Method: X-ray Diffraction
Resolution: 1.6 Å

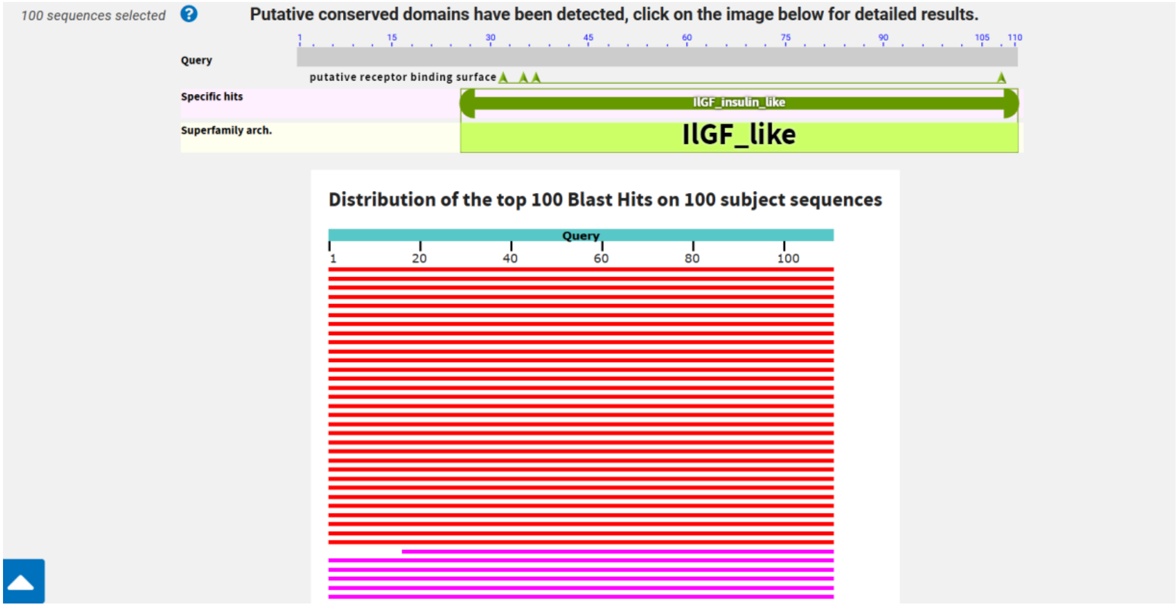
## 2. BLASTp Analysis & Homology Search

A Basic Local Alignment Search Tool (BLASTp) was utilized to identify homologous sequences across the "nr" (non-redundant) protein database. The resulting data provides a clear snapshot of how insulin has been preserved by evolutionary selection.

## A. Global Distribution of Hits

The **Graphic Summary** shows the top 100 alignments.

- **Query Coverage:** Every single red bar covers 100% of the query length, indicating that the entire insulin molecule is preserved across species rather than just small fragments.
- **Conserved Domains:** The analysis detected the **ILGF_insulin_like** superfamily domain, which is vital for metabolic regulation.


100 sequences selected

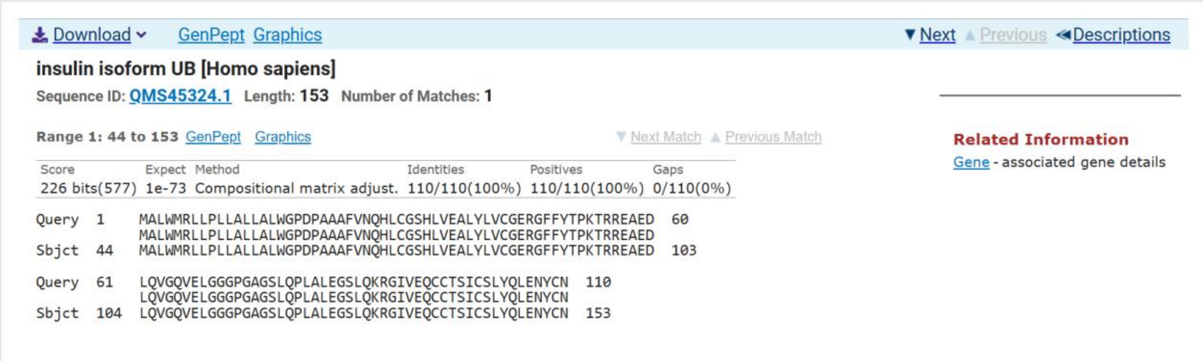
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 100 Blast Hits on 100 subject sequences

## 3. Statistical Analysis & Evolutionary Divergence

The statistical strength of these alignments demonstrates that these matches are biologically significant and not the result of random chance.

| **Parameter** | **Value** | **Biological Interpretation** |
|---|---|---|
| **Max Score** | 226 bits | Indicates a high-quality, robust alignment. |
| **Expect Value (E)** | 1e-73 | The probability of this match occurring by chance is virtually zero. |
| **Identities** | 110/110 (100%) | Complete amino acid conservation across the entire preproprotein. |
| **Gaps** | 0% | No insertions or deletions; the peptide length is identical between species. |


Download
GenPept Graphics
▼Next ▲Previous ◀Descriptions

insulin isoform UB [Homo sapiens]

Sequence ID: QMS45324.1 Length: 153 Number of Matches: 1

Range 1: 44 to 153 GenPept Graphics

▼Next Match ▲Previous Match
Related Information
Gene - associated gene details

```
Score          Expect  Method                        Identities     Positives      Gaps
226 bits(577)  1e-73   Compositional matrix adjust.  110/110(100%)  110/110(100%)  0/110(0%)

Query  1    MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  60
            MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED
Sbjct  44   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED  103

Query  61   LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
            LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Sbjct  104  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  153
```

Detailed sequence alignment of **Human Insulin (Query)** against its closest homologs. The middle line represents the consensus, showing zero mismatches.

## 4. The "Extraordinary" Exception: Discussion

The most fascinating takeaway from this analysis is the "evolutionary stasis" of the insulin protein.

### The Human-Primates Connection:

The BLAST results confirm that human insulin and gorilla insulin (*Gorilla gorilla gorilla*) share 100.00% identity. This is an exceptional finding. Even though humans and gorillas diverged roughly 7 million years ago, the biological "machinery" for glucose regulation is so precise that evolution has not tolerated a single amino acid change in this specific protein.

### Structural Preservation:

Looking at the alignment, the "MALWMRLL..." leader sequence is identical. This sequence is responsible for the translocation of insulin into the endoplasmic reticulum. If even a single letter in this code changed, the protein would fail to be secreted, likely resulting in a non-viable organism. This extreme "Purifying Selection" is why the BLAST results look so consistently uniform across millions of years of history.

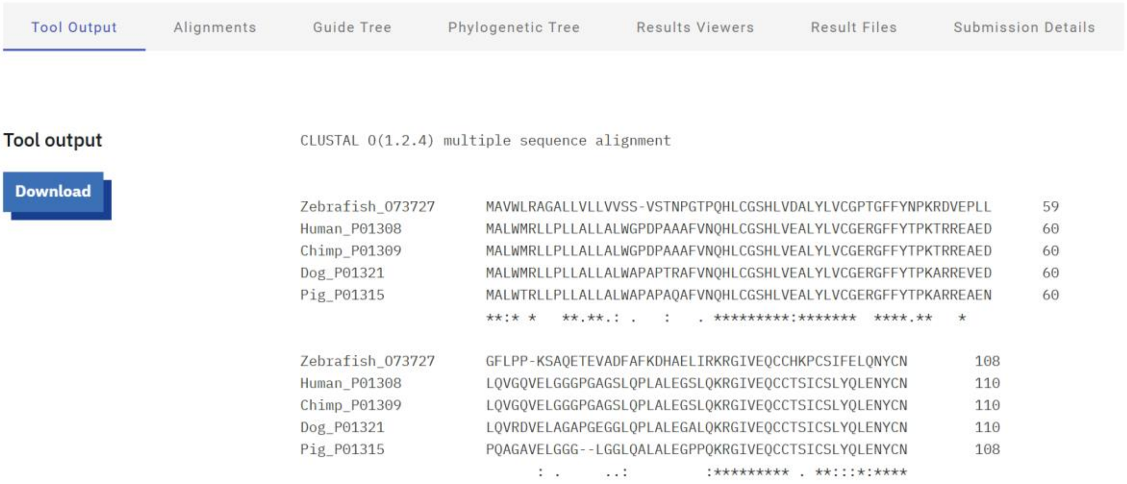
## 5. Conclusion

The analysis of P01308 demonstrates that insulin is one of the most highly conserved proteins in the mammalian lineage. The 100% identity with other primates and the near-zero E-values across the top 100 hits underscore its essential role in life. This protein is a masterpiece of biological engineering that evolution has found no reason to alter.

# Task 2: Multiple Sequence Alignment (MSA)

## 2.1 Methodology

Using Clustal Omega, a Multiple Sequence Alignment was performed on five diverse species: *Homo sapiens*, *Pan troglodytes*, *Canis lupus familiaris*, *Sus scrofa*, and *Danio rerio*.

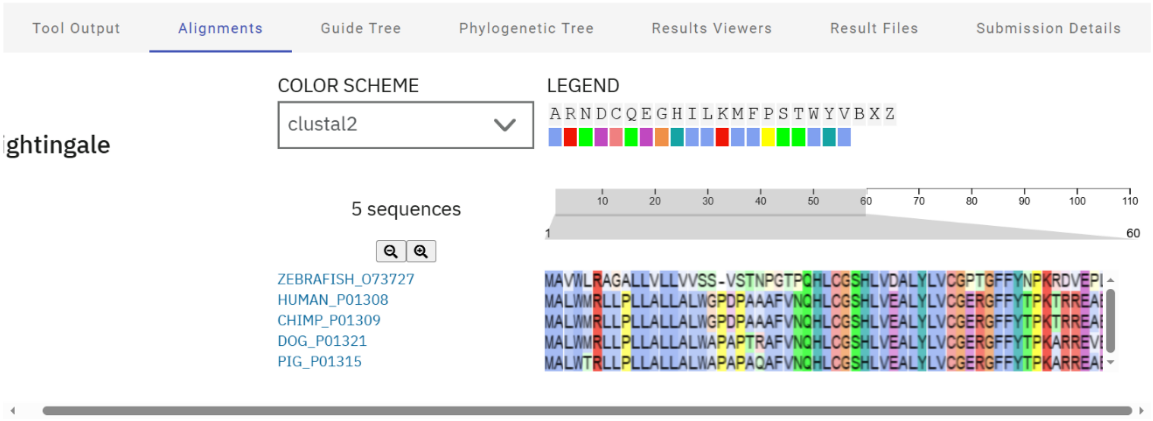


```
CLUSTAL O(1.2.4) multiple sequence alignment


Zebrafish_073727      MAVWLRAGALLVLLVVSS-VSTNPGTPQHLCGSHLVDALYLVCGPTGFFYNPKRDVEPLL      59
Human_P01308          MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED      60
Chimp_P01309          MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED      60
Dog_P01321            MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED      60
Pig_P01315            MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAEN      60
                      **:* *   **.**.: .   :   . *********:*******  ****.**   *

Zebrafish_073727      GFLPP-KSAQETEVADFAFKDHAELIRKRGIVEQCCHKPCSIFELQNYCN      108
Human_P01308          LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN      110
Chimp_P01309          LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN      110
Dog_P01321            LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN      110
Pig_P01315            PQAGAVELGGG--LGGLQALALEGPPQKRGIVEQCCTSICSLYQLENYCN      108
                            : .     ..:         :********* . **:::*:****
```

### 2.2 Interpretation of Conserved Regions & Motifs

The MSA reveals critical biological insights:

1. **Strict Conservation:** The A and B chains (represented by the clusters of asterisks *) are highly conserved across all mammals and even the zebrafish.
2. **Cysteine Motifs:** The six Cysteine residues required for disulfide bridge formation are 100% invariant. These are the "structural anchors" of the hormone.
3. **Divergence in Zebrafish:** *Danio rerio* shows significant gaps and substitutions in the C-peptide region (the middle section), as seen in the color-coded "Nightingale" viewer. This confirms that while functional chains are rigid, the connecting peptides are more tolerant of mutations.



COLOR SCHEME: clustal2

5 sequences

```
ZEBRAFISH_073727   MAVWLRAGALLVLLVVSS-VSTNPGTPQHLCGSHLVDALYLVCGPTGFFYNPKRDVEPL
HUMAN_P01308       MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAE
CHIMP_P01309       MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAE
DOG_P01321         MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVE
PIG_P01315         MALWTRLLPLLALLALWAPAPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAE
```

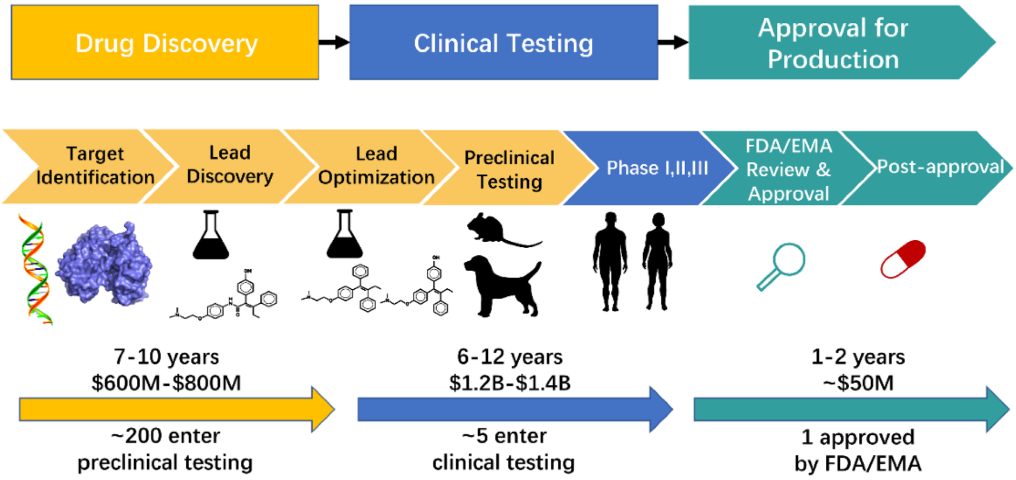
# Topic Choice: Role of Bioinformatics in Drug Discovery

## Introduction & Overview

- **Introduction:** Define bioinformatics as the intersection of biology, statistics, and computer science. Contrast traditional drug discovery (which takes 10–15 years and billions of dollars) with Bioinformatics-aided Drug Discovery (BADD).

- **The Drug Discovery Pipeline:** Briefly outline the stages:

Target Identification → Target Validation → Lead Discovery → Lead Optimization →Pre-clinical/Clinical trials.



## Target Identification and Structural Bioinformatics

- **Target Identification:** Discuss how genomic and proteomic databases (NCBI, UniProt) are mined to find genes/proteins related to diseases. Mention tools like **BLAST** for sequence similarity.

- **Structural Bioinformatics:** Explain the importance of 3D structures. Discuss the **Protein Data Bank (PDB)** and how tools like **AlphaFold** have revolutionized our ability to predict protein structures for targets where experimental data is missing.

## Virtual Screening and Molecular Docking

- **Computer-Aided Drug Design (CADD):** Describe the transition to "in silico" methods.

- **Virtual Screening:** How computers screen "libraries" of millions of chemical compounds against a target.

- **Molecular Docking:** Detail how software (e.g., AutoDock, GOLD) predicts the orientation and binding affinity of a drug candidate to its target protein.
- **QSAR Models:** Explain Quantitative Structure-Activity Relationship (QSAR) models used to predict the biological activity of new molecules based on their chemical structure.

## ADMET Prediction and Case Studies

- **ADMET Prediction:** Bioinformatics tools predict **A**bsorption, **D**istribution, **M**etabolism, **E**xcretion, and **T**oxicity. This prevents "attrition" (failure) in late-stage clinical trials by identifying toxic compounds early.
- **Case Study:** Use a real-world example.
  - *Example:* **Imatinib (Gleevec)** for Leukemia was one of the first successes of rational structure-based drug design.
  - *Example:* **COVID-19 Vaccines**: Mention how bioinformatics allowed for rapid viral sequencing and epitope mapping.

### Challenges, Conclusion, and References

- **Challenges:** Discuss data noise, the complexity of biological networks (systems biology), and the need for massive computing power.
- **Conclusion:** Summarize that bioinformatics is no longer optional but a core pillar of the pharmaceutical industry.

*Suggested sources:* Nature Reviews Drug Discovery, Journal of Chemical Information and Modeling, NCBI Bookshelf.

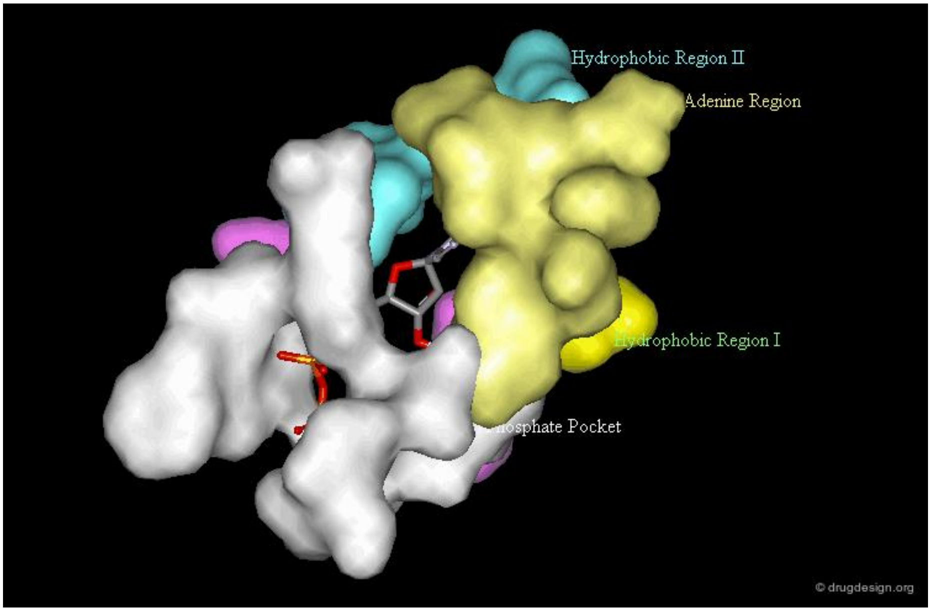
## 1. Comparative study of bioinformatics databases

- **Introduction:** Why databases are the backbone of bioinformatics.
- **NCBI (National Center for Biotechnology Information):** Discuss its role as an "umbrella" database (GenBank, PubMed, BLAST).
- **UniProt (Universal Protein Resource):** Focus on its curated, high-quality protein sequence and functional information.

- **PDB (Protein Data Bank):** Focus on 3D structural data (X-ray, NMR).
- **Comparative Table:** Create a table comparing them based on: *Primary Data Type, Update Frequency, Search Tools, and Accessibility.*
- **Conclusion:** How researchers use these databases in tandem to solve biological problems.

## 2. Applications of Machine Learning in Bioinformatics

- **Introduction:** The explosion of "Big Data" in biology.
- **Supervised vs. Unsupervised Learning:** Examples of each in biology (e.g., Clustering for gene expression vs. Classification for disease diagnosis).



- **Key Applications:**
  - **Protein Folding:** Mention Google DeepMind’s AlphaFold.
  - **Genomics:** Detecting mutations and predicting gene functions.
  - **Medical Imaging:** Using CNNs to detect tumors in scans.
- **Future Trends:** Deep learning and personalized medicine.

## General Bioinformatics & Overview

- **Attwood, T. K., & Parry-Smith, D. J. (1999).** *Introduction to Bioinformatics*. Pearson Education. (A foundational textbook for defining bioinformatics).

- **Bayat, A. (2002).** Science, medicine, and the future: Bioinformatics. *BMJ: British Medical Journal*, 324(7344), 1018. Link to Article
- **Mount, D. W. (2004).** *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.

## Role in Drug Discovery (CADD & Virtual Screening)

- **Kapetanovic, I. M. (2008).** Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2), 165-176.
- **Ou-Yang, S. S., Lu, J. Y., Kong, X. Q., Liang, Z. J., Luo, C., & Jiang, H. (2012).** Computational drug discovery. *Acta Pharmacologica Sinica*, 33(9), 1131-1140. Link to Article
- **Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014).** Computational methods in drug discovery. *Pharmacological Reviews*, 66(1), 334-395.

## Databases & Tools (NCBI, PDB, UniProt)

- **Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000).** The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235-242. Link to PDB
- **Sayers, E. W., et al. (2022).** Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 50(D1), D20-D26.
- **The UniProt Consortium. (2023).** UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523-D531.

## Modern Breakthroughs (AlphaFold & AI)

- **Jumper, J., Evans, R., Pritzel, A., et al. (2021).** Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
- **Vamathevan, J., et al. (2019).** Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463-477.