جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

قسم الانظمة الطبية الذكية

المرحلة الثانية

Fifth lecture
Advanced Data Retrieval Techniques

Class: Second

Lecturers

M.S.c. Safanah Albayati

Dr. Maytham N. Meqdad,

Lecture Five:Advanced Data Retrieval Techniques

Table of Contents

1. Advanced Data Retrieval Techniques
2. Integration of Retrieved Data

Section 3: Advanced Data Retrieval Techniques

As biological datasets continue to grow exponentially, researchers require more efficient and automated ways to retrieve and analyze data. Advanced data retrieval techniques include **APIs (Application Programming Interfaces)** and **Automated Data Mining**, which allow users to extract large-scale biological information quickly and programmatically.

1. APIs and Web Services in Bioinformatics

A. What Are APIs and How Are They Used?

APIs (Application Programming Interfaces) provide programmatic access to bioinformatics databases, allowing researchers to retrieve, filter, and analyze biological data in an automated manner. APIs replace the need for manual searches by enabling:

- **Automated Queries**: Retrieve data without manually interacting with web interfaces.
- **Bulk Data Extraction**: Retrieve large datasets efficiently.
- **Data Integration**: Combine information from multiple sources in real time.
- **Reproducibility**: Ensure consistent and reproducible data retrieval for research.

APIs are typically categorized as:

- **RESTful APIs (Representational State Transfer APIs)**: These APIs use HTTP requests (GET, POST, PUT, DELETE) to communicate with databases.
- **SOAP APIs (Simple Object Access Protocol APIs)**: Used in older systems but are being phased out in favor of RESTful APIs due to their complexity.

B. Case Studies: RESTful APIs in Bioinformatics

Several bioinformatics databases provide RESTful APIs for efficient data retrieval. Below are case studies demonstrating their usage.

Case Study 1: Retrieving Protein Data Using UniProt API

- **Database**: UniProt (Universal Protein Resource) provides protein sequence and functional information.
- **API Usage**:
  - A researcher studying **BRCA1** protein wants to retrieve its sequence and functional annotations.
  - Instead of manually searching UniProt, they use the API:

Example API Call:
https://rest.uniprot.org/uniprotkb/search?query=BRCA1&format=json

  - The API returns a **JSON-formatted** response containing BRCA1's sequence, functions, and structural details.

Benefits:

  - Enables large-scale retrieval of protein sequences.
  - Integrates protein information into computational workflows.

C. Other Bioinformatics APIs

API	Database	Use Case
NCBI E-utilities	**GenBank, PubMed**	**Retrieve nucleotide sequences, articles, and metadata**
UniProt REST API	UniProt	Fetch protein sequences and annotations
Ensembl REST API	**Ensembl Genome Browser**	**Retrieve genome annotations and variation data**
KEGG API	**Kyoto Encyclopedia of Genes and Genomes**	**Access metabolic pathway and drug interaction data**

APIs significantly enhance research workflows, allowing seamless integration of bioinformatics data into computational analyses.

2. Automated Data Mining in Bioinformatics

A. What is Automated Data Mining?

Automated data mining involves using **scripts and bioinformatics tools** to extract, filter, and analyze biological data. This technique is essential for handling **large-scale datasets**, identifying patterns, and reducing the time required for manual data retrieval.

B. Tools and Techniques for Automated Data Mining

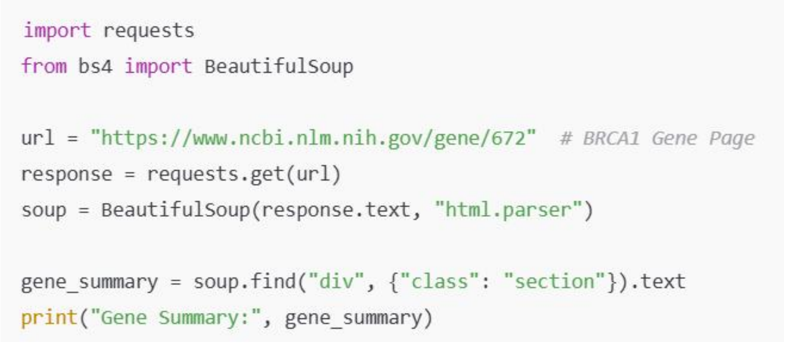
Automated data mining is performed using specialized tools and programming languages, with Python being the most widely used in bioinformatics.

1. Web Scraping for Data Extraction

Web scraping allows researchers to extract structured biological data from web pages when APIs are not available.

- **Example: Scraping Gene Information from NCBI**
  - Python's **BeautifulSoup** and **requests** libraries can be used to scrape gene descriptions, variants, and functional annotations.

Example Python Script for Web Scraping:



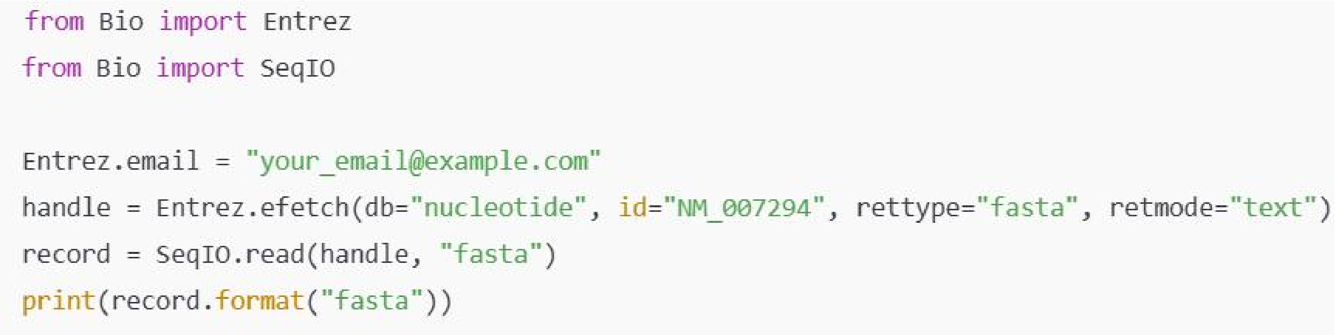
```
import requests
from bs4 import BeautifulSoup

url = "https://www.ncbi.nlm.nih.gov/gene/672"  # BRCA1 Gene Page
response = requests.get(url)
soup = BeautifulSoup(response.text, "html.parser")

gene_summary = soup.find("div", {"class": "section"}).text
print("Gene Summary:", gene_summary)
```

- **Applications**:
  - Extracting metadata from gene and protein databases.
  - Gathering large-scale research article summaries.

2. Bioinformatics Data Retrieval Using Python (Biopython)

Python-based data retrieval is preferred for automation in bioinformatics. **Biopython**, a widely used Python library, simplifies sequence retrieval, analysis, and visualization.

- **Retrieving a DNA Sequence Using Biopython**



```python
from Bio import Entrez
from Bio import SeqIO

Entrez.email = "your_email@example.com"
handle = Entrez.efetch(db="nucleotide", id="NM_007294", rettype="fasta", retmode="text")
record = SeqIO.read(handle, "fasta")
print(record.format("fasta"))
```

- **Applications**:
  - Automates retrieval of nucleotide and protein sequences.
  - Enables bulk data extraction from **NCBI, Ensembl, and UniProt**.

3. Machine Learning in Automated Data Mining

- **Clustering Algorithms**: Used for grouping genes with similar expression patterns.
- **Natural Language Processing (NLP)**: Extracts biological insights from large-scale biomedical literature.
- **Graph-Based Mining**: Identifies protein-protein interactions.

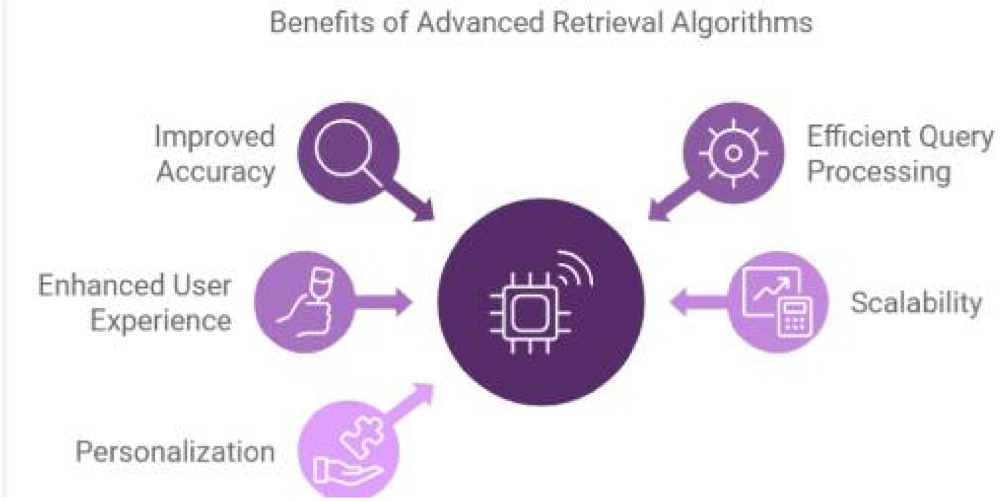
C. Comparison of APIs and Automated Data Mining Techniques

Technique	Advantages	Disadvantages
APIs	Structured data retrieval, fast and reliable, well-documented	Limited to predefined query formats
Web Scraping	Extracts data when APIs are unavailable	Can break if website structure changes
Biopython	Automates large-scale sequence retrieval and analysis	Requires Python programming knowledge

3. Practical Applications of Advanced Data Retrieval Techniques

- **Large-Scale Genome Annotation**: Fetch gene annotations from Ensembl via REST API.
- **Drug Discovery**: Mine KEGG database for drug-target interactions.
- **Disease Gene Identification**: Extract mutations from ClinVar using NCBI E-utilities.

Lecture Five:Advanced Data Retrieval Techniques


Benefits of Advanced Retrieval Algorithms
Improved Accuracy
Efficient Query Processing
Enhanced User Experience
Scalability
Personalization

Section 4: Integration of Retrieved Data

Once biological data has been retrieved from various sources using APIs, web scraping, or automated mining techniques, the next crucial step is **data integration**. This process involves consolidating heterogeneous data from multiple sources, ensuring consistency, and making it usable for further bioinformatics analysis. Data integration is essential for large-scale biological studies, such as multi-omics research, comparative genomics, and clinical applications.

1. Data Integration Challenges

Despite advancements in data retrieval, integrating large datasets poses several challenges.

A. Common Challenges in Integrating Data from Multiple Sources

1. Data Format Inconsistency

1. Different databases store data in various formats (FASTA, CSV, JSON, XML, etc.), requiring standardization before analysis.
2. **Example**: UniProt provides protein sequence data in FASTA format, whereas Ensembl provides genomic annotations in GTF/GFF formats.

2.Data Redundancy and Duplication

1. The same datasets may be stored across multiple databases, leading to redundancy.
2. **Example**: A gene's sequence may be available in both GenBank and Ensembl, but each might have different annotations.

3. Conflicting Annotations and Nomenclature Issues

1. Different research groups may annotate the same gene differently, leading to inconsistencies.
2. **Example**: The TP53 gene may have different annotations in RefSeq (NCBI) vs. Ensembl.

4. Scalability and Storage Issues

1. With the growth of high-throughput sequencing, integrating and processing large-scale datasets requires scalable computing resources.
2. **Solution**: Cloud-based platforms like **Google Cloud Genomics** and **AWS Bioinformatics** are increasingly used for large-scale data integration.

B. Solutions for Data Inconsistency and Redundancy

1. Data Standardization Using Ontologies

1. **Gene Ontology (GO)**: Standardizes descriptions of gene functions.
2. **Medical Subject Headings (MeSH)**: Helps classify biomedical literature.

2.Cross-Referencing Data Using Identifiers

1. **Ensembl Gene ID**, **NCBI RefSeq ID**, and **UniProt Accession Numbers** help researchers unify annotations.

3.Data Cleaning and Merging Techniques

Lecture Five:Advanced Data Retrieval Techniques

1. **ETL (Extract, Transform, Load) Pipelines**: Used to clean and integrate datasets from multiple sources.
2. **Example**: **BioMart** allows integration of gene, protein, and variation data.

2. Tools and Software for Data Integration

Several bioinformatics tools have been developed to simplify data integration.

A. BioMart: A Data Integration Platform

Overview: BioMart is an open-source system that provides a unified interface to query multiple biological databases.

Applications:

- Retrieve gene annotations from Ensembl.
- Cross-reference genomic data with disease databases.
- Integrate SNP (single nucleotide polymorphism) data with gene expression profiles.

Demonstration:

- A researcher looking for human *BRCA1* variants can use BioMart to:

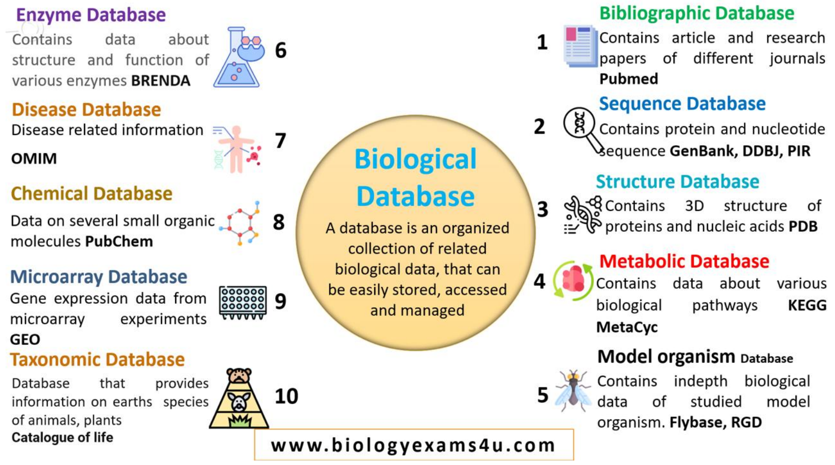
- Select Ensembl Genes as the dataset.
- Apply filters for the *BRCA1* gene in *Homo sapiens*.
- Retrieve associated SNPs, functional annotations, and pathway data.

B. Galaxy: A Workflow-Based Integration Platform

Overview: Galaxy is a web-based platform that provides an easy-to-use interface for integrating bioinformatics tools.

Applications:

- Perform whole-genome sequence analysis.
- Integrate RNA-seq data with functional annotations.
- Automate multi-step workflows for large-scale studies.


Biological Database
A database is an organized collection of related biological data, that can be easily stored, accessed and managed
1 Bibliographic Database
Contains article and research papers of different journals Pubmed
2 Sequence Database
Contains protein and nucleotide sequence GenBank, DDBJ, PIR
3 Structure Database
Contains 3D structure of proteins and nucleic acids PDB
4 Metabolic Database
Contains data about various biological pathways KEGG MetaCyc
5 Model organism Database
Contains indepth biological data of studied model organism. Flybase, RGD
6 Enzyme Database
Contains data about structure and function of various enzymes BRENDA
7 Disease Database
Disease related information OMIM
8 Chemical Database
Data on several small organic molecules PubChem
9 Microarray Database
Gene expression data from microarray experiments GEO
10 Taxonomic Database
Database that provides information on earths species of animals, plants Catalogue of life
www.biologyexams4u.com