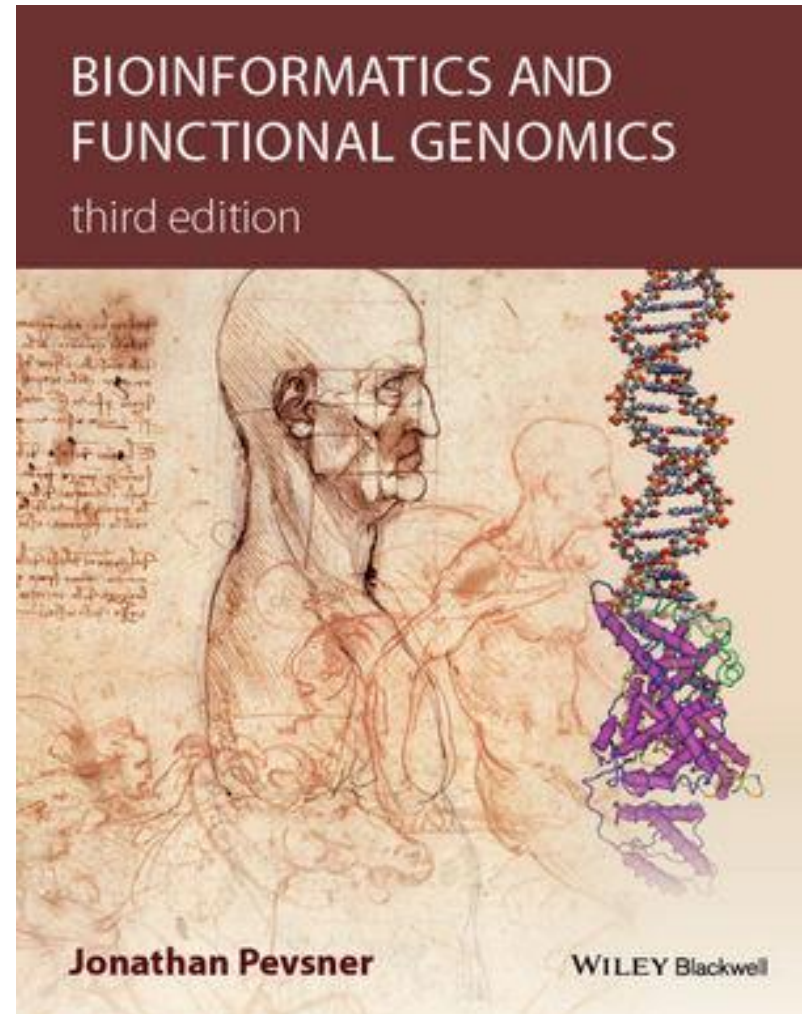


Bioinformatics

Introduction

Reference:
Bioinformatics and Functional Genomics
Jonathan Pevsner
John Wiley & Sons Inc



❑ A definition of **Bioinformatics**:

Bioinformatics is an **interdisciplinary field** that combines **biology**, **computer science**, **mathematics**, **statistics**, and **engineering** to **analyze** and **interpret biological data**.

❑ **Bioinformatics** involves:

development and application of **computational tools, algorithms, and databases**

TO UNDERSTAND

Complex biological processes:

- **genetic sequences**
- **protein structures**
- **gene expression**
- **molecular interactions**

Key components of bioinformatics include:

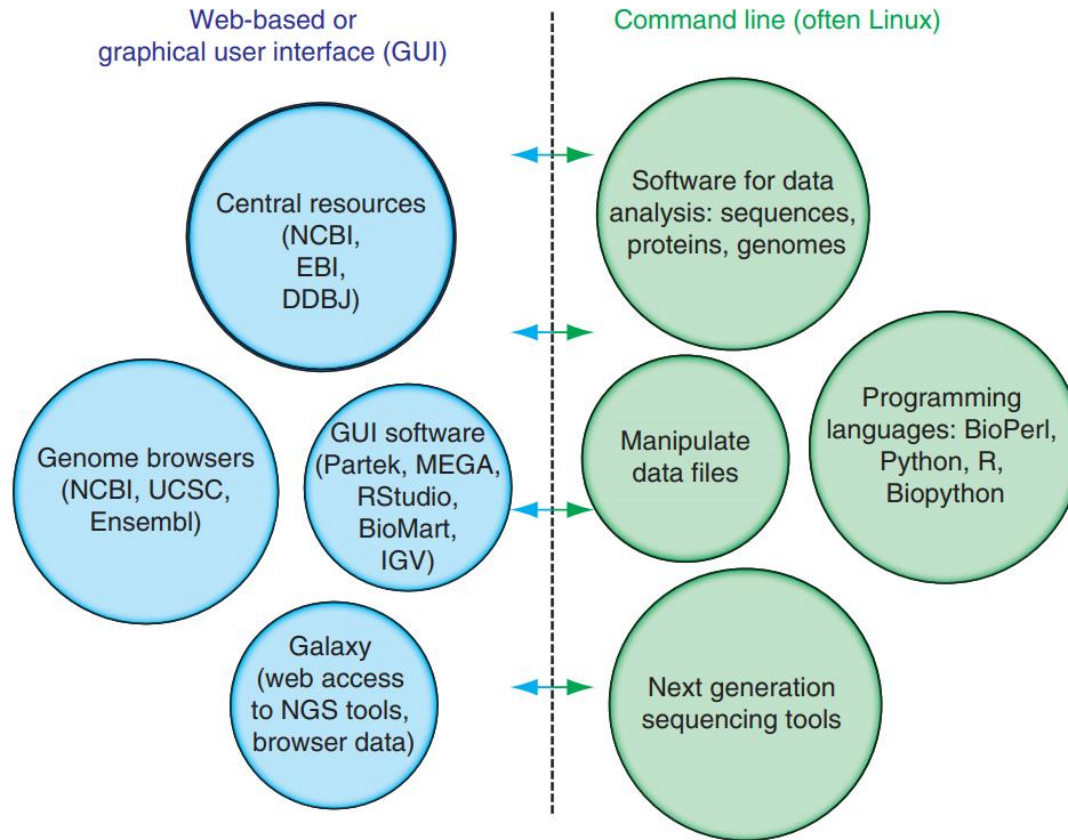
- ❑ **Data Management and Storage:** Organizing and storing **large-scale biological data**, such as **DNA sequences**, **protein structures**, and **gene expression profiles**, in databases like **GenBank**, **UniProt**, and the **Protein Data Bank (PDB)**.
- ❑ **Sequence Analysis:** Studying DNA, RNA, and protein sequences to **identify genes**, **regulatory elements**, **mutations**, and **evolutionary relationships**. Tools like **BLAST** and **ClustalW** are commonly used for **sequence alignment** and **comparison**.
- ❑ **Structural Bioinformatics:** Analyzing the **three-dimensional structures** of **proteins** and **nucleic acids** to understand their **functions** and **interactions**. Techniques include **molecular modeling**, **docking**, and **simulation**.
- ❑ **Functional Genomics:** Investigating **gene functions** and **interactions** on a **genome-wide scale** using technologies like **microarrays**, **RNA sequencing (RNA-seq)**, and **CRISPR-Cas9**.

continued...

Key components of bioinformatics include:

- ❑ **Systems Biology: Modeling and simulating biological systems** to understand **complex interactions** between **genes, proteins, and metabolites** within **cells or organisms**.
- ❑ **Pharmacogenomics and Drug Design:** Using **genomic information** to develop **personalized medicines** and **identify drug targets** through computational methods like **virtual screening** and **molecular dynamics**.
- ❑ **Metagenomics:** Analyzing **genetic material** from **environmental samples** to study **microbial communities** and their **roles in ecosystems**.
- ❑ **Evolutionary Bioinformatics:** **Reconstructing evolutionary histories** and **relationships** among **species** using **phylogenetic analysis** and **comparative genomics**.
- ❑ **Machine Learning and Artificial Intelligence:** Applying **advanced computational techniques** to **predict biological outcomes, classify data, and uncover patterns** in **large datasets**.

Bioinformatics Software: Two Cultures



Web-based or “**point-and-click**”, including:

- **major portals** (National Center for Biotechnology Information, European Bioinformatics Institute)
- **major genome browsers** (Ensembl, UCSC)
- **databases**
- **specialized websites**

Command-line resources, including:

- **programming languages** (such as **Biopython**, **BioPerl**, and the **R language**)
- command-line software (typically **accessed** using the **Linux** operating system).

- ❑ **Command-line tools** may have a **steeper learning curve**, but almost always offer **more options** for **executing programs**.
- ❑ They are **more appropriate** for **analyzing large-scale datasets** that are now routinely encountered in bioinformatics.
- ❑ Even for **smaller datasets**, command-line approaches can offer **more flexibility** and **precision** in accomplishing your tasks and **more reproducible** research since you can **document** your **analysis steps**.

Web-Based Software

- ❑ The field of **bioinformatics** relies **heavily** on the **Internet** as a **place** to **access**
 - **sequence data**
 - **software** that is useful to analyze molecular data
 - as a **place** to **integrate** different kinds of **resources** and **information** relevant to biology

- ❑ The main **publicly accessible databases** that serve as **repositories** for **DNA** and **protein** data:
 - The **National Center for Biotechnology Information (NCBI)**, which hosts **GenBank** and other resources
 - The **European Bioinformatics Institute (EBI)**
 - **Ensembl**, which includes a **genome browser** and **resources** to study **dozens** of **genomes**
 - The **University of California at Santa Cruz (UCSC)** Genome bioinformatics site, including a **web browser** and **table browser** for a variety of species

The **main advantages** offered by **websites**:
easy access, rapid updates, good visibility to the community, and **ease of use** (since in general programming skills, command line skills, and the use of Linux-type operating systems are not required).

Command-Line Software

Command-line tools offer **distinct, critical advantages**. **High-throughput approaches** to biology result in the **creation** of both **large** and **small datasets** which **require sophisticated analyses**.

❑ We can think about command-line software in several ways:

- The **operating system** is often **Linux** (a Unix-like environment).
- The **Mac O/S** is compatible with Linux as well (and is **POSIX-compliant**).
- The **Windows**-type operating systems are **popular**, they are **not appropriate** for the **majority of command- line programs**.
- **Programming languages** are commonly used in bioinformatics. Examples are **Perl** (or its relative **BioPerl**), **Python** (as well as **Biopython**), and **R** to manipulate data. (see [myCompiler](#))
- The **command line of Unix systems** offers **Bash**, a **default shell** for **Linux** and **Mac OS X** operating systems. Bash **includes** a series of **utilities** that can accomplish tasks such as **sorting** a table of data, **transposing** it, **counting** the numbers of rows and columns, **merging** data, or working with regular expressions.

Overview of some **web-based** (or **graphical user interface (GUI)**) and **command-line** software

Topic	Web-based or GUI software	Command-line software
Access to information	BioMart Genome Workbench	EDirect
Pairwise alignment	BLAST	BLAST+ Biopython needle (EMBOSS) water (EMBOSS)
BLAST	BLAST	BLAST+
Database searching	DELTA-BLAST Megablast	HMMER
Multiple alignment	Pfam, MUSCLE	MAFFT
Phylogeny	MEGA	MrBayes
Chromosomes	Galaxy	geecee (EMBOSS) isochore (EMBOSS)
Next-generation sequencing	Galaxy, SIFT, PolyPhen2	SAMTools, tabix, VCFtools
RNA	RNAfam, tRNAscan	
RNAseq	Galaxy	affy (R package), RSEM
Proteomics	ExPASy	pepstats (EMBOSS)
Protein structure	Cn3D, Pymol	psiphi (EMBOSS)
Functional genomics	FLink, Cytoscape	
Tree of life		Velvet (assembly)
Viruses		MUMmer (alignment)
Bacteria and archaea	MUMmer	GLIMMER (gene-finding)
Fungi	YGOB	Ensembl (variants)
Eukaryotic genomes		
Human genome		PLINK
Human disease	OMIM, BioMart	EDirect, MitoSeek

Ten rules for online learning:

- make a plan
- be selective
- organize your learning environment
- do the readings
- do the exercises
- do the assessments
- exploit the advantages (e.g., convenience)
- reach out to others
- document your achievements
- be realistic

Critical assessment competitions in bioinformatics

Name/Acronym	Competition
Alignathon	Compare whole-genome alignment methods
EGASP	ENCODE Genome Annotation Assessment Project
Assemblathon	Compare the performance of genome assemblers
GAGE	Genome Assembly Gold-standard Evaluations
ABRF	Association of Biomolecular Resource Facilities (ABRF) assessment of phosphorylation
CASP	Critical Assessment of Structure Prediction
CAFA	Critical Assessment of Protein Function
CAGI	Critical Assessment of Genome Interpretation