Chapter 3: Bioinformatics

1.Setting the Stage

As biological research increasingly relies on computational innovation, bioinformatics emerges as a vital discipline, combining biology with the sophisticated tools of computer science to unlock the secrets within vast biological datasets. This chapter guides Master's students in computer science through the essential principles and practices of bioinformatics, highlighting how algorithms, data structures, and statistical approaches drive the analysis of genomic sequences, protein interactions, and high-volume biological information. Framed within English for Specific Purposes (ESP), this discussion sharpens your technical proficiency while equipping you to express intricate computational insights, engage with diverse research teams, and advance bioinformatics on a global stage through clear, specialized English.

Activity 1: Matching Key Bioinformatics Concepts

Match each term in the first column with its correct definition from the second column. Write the corresponding letter next to each number.

#	Term	Definition
1	Bioinformatics	A. A method for processing large volumes of data quickly and efficiently.
2	Algorithm	B. The physical or functional relationship between proteins in a biological system.

#	Term	Definition
3	Genomic Sequence	C. Involving collaboration across multiple fields of study.
4	Data Structure	D. A set of instructions designed to solve a problem or complete a task computationally.
5	Protein Interaction	E. A mathematical representation used to identify patterns or make predictions from data.
6	High-Throughput	F. A framework that integrates biology, computer science, and mathematics to solve biological problems.
7	Statistical Model	G. The ordered arrangement of nucleotides in DNA or RNA.
8	Interdisciplinary	H. A way of organizing data to enable efficient access and manipulation.

Activity 2:Bioinformatics Concept Matching

Below are five key terms and five relationship phrases. Your task is to connect the terms correctly by matching them with the appropriate phrase.

Key Terms:

- **Bioinformatics**
- **Computer Science**
- **Biology**
- **Data Analysis**
- **Research**

Relationship Phrases:

- **"integrates with"**
- **"relies on"**
- **"enables"**
- **"supports"**
- **"applies to"**

Activity 3: True or False Brainstorm

Read the following five statements about bioinformatics. Decide whether each is true or false based on your prior knowledge or logical reasoning as a computer science student.

1. Bioinformatics only involves analyzing DNA sequences.
2. Computer science skills like programming are essential for bioinformatics.
3. Statistical models in bioinformatics help predict biological outcomes.
4. Bioinformatics is a field limited to biologists with no need for collaboration.
5. High-throughput data refers to slow, manual data collection processes.

2.Take a Listen

Activity 4:Decoding DNA Storage

Answer the following questions based on the video. Some questions have multiple-choice options, while others require short written responses.

1. **What is 'cold data'?**
   a) Data that is frequently accessed and modified
   b) Archived data that is rarely or never accessed
   c) Data that requires constant maintenance
   d) Data stored on traditional hard drives
2. **What advantage does DNA storage have over traditional data centers?**
   a) It requires continuous electrical power to maintain data
   b) It is significantly more compact and stable over time
   c) It can process information at a much faster speed
   d) It eliminates the need for encoding
3. **Which of the following is NOT a reason why researchers are interested in DNA as a storage medium?**
   a) Its compactness allows for a high data density
   b) It has a much longer lifespan compared to traditional storage methods
   c) It enables instant access to data
   d) It is highly resistant to environmental damage
4. **What is the key difference between binary and quaternary encoding in data storage?**
   - *(Short-answer question)*

5. **What factors can degrade stored DNA molecules?**
   - *(Short-answer question – list at least three factors)*
6. **How does sequencing help in retrieving stored data from DNA?**
   - *(Short-answer question – explain in two to three sentences)*
7. **Which of the following statements about the cost of DNA storage is true?**
   a) It is currently much cheaper than traditional hard drives
   b) It costs around $1 per gigabyte in 2024
   c) It remains expensive but may become more affordable in the future
   d) It is free because DNA naturally exists in living organisms
8. **What is the primary challenge of using DNA storage for everyday computing tasks?**
   a) Limited data capacity
   b) Extremely slow reading and writing speeds
   c) High energy consumption
   d) Difficulty in encoding binary data
9. **Why do researchers need to store multiple copies of DNA-encoded data?**
   - *(Short-answer question – explain in two to three sentences)*
10. **What is the long-term goal of DNA data storage technology?**
    a) To replace all current hard drives
    b) To provide a sustainable method for storing cold data
    c) To develop biological computers
    d) To allow data storage inside living organisms

Activity 5: Understanding DNA Data Storage Basics

Choose the best option for each question.

Questions:

1. What is the main problem with traditional data storage mentioned in the video?
   a) It is too expensive
   b) It consumes too much energy and space
   c) It cannot store videos
   d) It is not secure
2. What percentage of stored data is considered "cold data"?
   a) 20-40%
   b) 40-60%

c) 60-80%
d) 80-100%

3. Who first proposed the idea of using DNA for data storage?
   a) Mark Antonini
   b) Richard Feynman
   c) A computer scientist in the 1990s
   d) A biologist in the 2000s
4. How is DNA structured for data storage?
   a) As a sequence of 0s and 1s
   b) As a sequence of four nucleotide bases (A, T, G, C)
   c) As a 3D hologram
   d) As magnetic particles
5. What is the first step in storing data in DNA?
   a) Synthesizing the DNA
   b) Encoding the file into a quaternary sequence
   c) Freezing the DNA
   d) Sequencing the DNA
6. Why is DNA storage considered highly compact?
   a) It uses quantum physics
   b) It requires no energy
   c) It can store data in a molecule a few micrometers wide
   d) It is biodegradable
7. What is the main challenge with DNA data storage in 2024?
   a) It is too expensive and slow
   b) It cannot store images
   c) It requires too much water
   d) It only works in living cells
8. How is DNA protected from degradation?
   a) Stored in a vacuum-sealed metal capsule
   b) Kept in a freezer
   c) Mixed with special enzymes
   d) Exposed to sunlight
9. What is the purpose of synthesizing multiple copies of DNA strands?
   a) To make the data storage faster
   b) To reduce errors when reading the data
   c) To make the DNA glow in the dark
   d) To increase storage capacity

10. Which organization is involved in developing DNA storage standards for images?
    a) NASA
    b) JPEG DNA
    c) The European Union
    d) The World Health Organization

Activity 6:DNA Data Storage Process

Read each statement and decide if it is **True (T)** or **False (F)**

Questions:

1. **T/F** DNA data storage was first proposed in the 1950s.
2. **T/F** DNA can store data more densely than a traditional hard drive.
3. **T/F** Binary code (0s and 1s) is directly written into DNA.
4. **T/F** DNA storage requires constant energy to maintain data.
5. **T/F** Errors in DNA sequencing can be reduced by storing multiple copies.
6. **T/F** DNA storage is currently faster than traditional hard drives.
7. **T/F** DNA can preserve data for thousands of years if stored correctly.
8. **T/F** The JPEG DNA group is working on image compression for DNA storage.
9. **T/F** DNA synthesis is currently cheaper than traditional data storage.
10. **T/F** DNA computers already exist and are widely used.

3. Read & Reflect

Advances in Spatial Transcriptomics and Its Applications in Cancer Research

Rao, A., et al. (2021). Nature Reviews Cancer, 21, 513–526.

Spatial transcriptomics (ST) has redefined how scientists study cancer by adding a critical dimension—location—to gene expression data. Traditional methods like bulk RNA sequencing average signals across a tissue sample, masking cellular diversity, while single-cell RNA sequencing (scRNA-seq) isolates cells but loses their spatial context. ST bridges this gap, mapping RNA transcripts to their precise positions within intact tissues. This 2021 review by Rao et al. explores how ST technologies have evolved, their methodological underpinnings, and their transformative applications in cancer research, particularly in decoding tumor heterogeneity, microenvironment dynamics, and therapeutic strategies.

The article first outlines ST's technical landscape, which splits into two camps: imaging-based and sequencing-based methods. Imaging approaches, such as single-molecule fluorescence in situ hybridization (smFISH) and MERFISH, use fluorescent probes to visualize specific RNA molecules directly in tissue sections. These offer high spatial resolution—down to single-cell or subcellular levels—but are limited to preselected gene panels, typically dozens to hundreds. Sequencing-based methods, like 10x Genomics' Visium platform or Slide-seq, rely on high-throughput sequencing (HTS). They barcode RNA molecules to spatial coordinates on arrays, capturing whole transcriptomes across thousands of spots. Early ST platforms (e.g., Ståhl et al., 2016) had coarse resolution (100 µm spots), but by 2021, innovations like Seq-Scope and Stereo-seq achieved near-single-cell precision, integrating HTS to process vast datasets efficiently. Rao et al. emphasize how these advances overcome initial hurdles—low throughput, limited gene coverage, and spatial blur—making ST a powerhouse for cancer studies.

In cancer research, ST's ability to reveal spatial patterns is a game-changer. Tumors are not uniform; their cells vary by location, influencing growth, invasion, and treatment response. The review highlights ST's role in profiling

tumor heterogeneity—for example, in colorectal cancer, ST has mapped distinct gene expression between a tumor's core and its invasive front, revealing how cancer-associated fibroblasts (CAFs) interact with malignant cells to drive progression (Peng et al., 2022, cited in related works). Similarly, in breast cancer, ST has identified hypoxic zones linked to aggressive subtypes, offering spatial clues to metastasis (Sun et al., 2021). Beyond tumor cells, ST dissects the tumor microenvironment (TME), a complex ecosystem of immune cells, stromal cells, and vasculature. In liver cancer, it's pinpointed immunosuppressive niches where macrophages dampen T-cell activity (Wu et al., 2023), while in colorectal cancer, it's uncovered immune hubs like tertiary lymphoid structures (TLSs) as potential prognostic markers (Pelka et al., 2021). These findings underscore how spatial organization shapes cancer behavior.

Therapeutically, ST is poised to transform oncology. By exposing resistance mechanisms tied to spatial niches, it guides drug development—for instance, identifying why certain tumor regions resist chemotherapy due to low oxygen or altered metabolism. In pancreatic ductal adenocarcinoma, ST paired with scRNA-seq has mapped molecular profiles to tissue architecture, suggesting targets for precision therapies (Moncada et al., 2020). The review also envisions ST enhancing diagnostics: integrating gene expression with histopathology could improve cancer staging or predict immunotherapy response. For example, TLS detection via ST might flag patients likely to benefit from checkpoint inhibitors.

Despite its promise, ST faces challenges. The technology is costly—Visium runs can exceed $10,000 per sample—and data analysis is daunting. A single experiment generates gigabytes of spatial and transcriptomic data, requiring advanced bioinformatics tools (e.g., Seurat, Scanpy) to align, cluster, and interpret. Resolution remains a trade-off: imaging methods excel at fine detail but lack scale, while sequencing methods cover more genes but blur cellular boundaries. Integration with multi-omics (e.g., proteomics, epigenomics) is another frontier, demanding computational leaps to fuse datasets seamlessly. Rao et al. note that artificial intelligence (AI) is increasingly vital, with machine learning models enhancing pattern detection and noise reduction in ST data.

Looking ahead, the authors argue ST's trajectory is steep. By 2025, falling sequencing costs, refined platforms (e.g., NanoString's CosMx), and AI-driven analysis could make ST routine in cancer labs and clinics. Its ability to link molecular insights to spatial context offers a richer understanding of cancer

biology—why tumors grow, resist, or spread—and paves the way for personalized treatments. For bioinformatics, ST is a proving ground, merging HTS, imaging, and computation into a tool that's rewriting cancer's spatial story.

Activity 7:Decoding Technologies

Below is a list of spatial transcriptomics (ST) methods/platforms and a list of features (A–H). For each method, **write the correct letters** that match its characteristics. Multiple features may apply.

Features:

A. High spatial resolution (single-cell or subcellular level)
B. Uses barcoded arrays
C. Limited to selected gene panels
D. Captures whole transcriptomes
E. Coarse resolution (~100 µm)
F. Near-single-cell precision
G. Uses fluorescent probes for visualization
H. Integrates high-throughput sequencing (HTS)

1. **smFISH:** _____
2. **MERFISH:** _____
3. **10x Genomics Visium:** _____
4. **Slide-seq:** _____
5. **Seq-Scope:** _____
6. **Stereo-seq:** _____
7. **Imaging-based ST methods:** _____
8. **Sequencing-based ST methods:** _____
9. **Early ST platforms (e.g., Ståhl et al., 2016):** _____
10. **Recent ST platforms (e.g., Stereo-seq, Seq-Scope):** _____ A, B, D, F, H

Activity 8:Understanding the Basics

Read each statement carefully and write **T** for *True* or **F** for *False* next to it.

#	Statement
1	Spatial transcriptomics (ST) provides information about where gene expression occurs in tissue.
2	Single-cell RNA sequencing preserves the spatial context of gene expression.
3	Imaging-based ST methods typically cover thousands of genes.
4	Sequencing-based methods like Visium use barcoded arrays to capture RNA transcripts.
5	Tumors are generally uniform in structure and gene expression.
6	ST can help identify which regions of a tumor resist chemotherapy.
7	The tumor microenvironment includes immune and stromal cells.
8	All ST methods provide the same level of spatial resolution.
9	Artificial intelligence is not yet used in analyzing ST data.
10	ST has no clinical diagnostic potential.

Activity 9:Key Terms in Context

Fill in each blank using the appropriate word from the word bank. Each word is used **only once**.

1. Spatial transcriptomics adds a _____ dimension to gene expression data.
2. Tumor _____ refers to the variation of cells and genes in different areas of a tumor.
3. The _____ platform is a sequencing-based method developed by 10x Genomics.
4. Imaging-based methods use _____ probes to visualize RNA molecules.
5. ST technologies vary in their spatial _____, from coarse to near-single-cell.
6. Some tumor regions form a resistant _____ where treatment is less effective.
7. In liver cancer, _____ suppress T-cell activity in immunosuppressive zones.
8. The tumor _____ is a mix of cancer cells, immune cells, and other components.
9. _____ is a sequencing-based spatial transcriptomics technology known for high-throughput mapping.

10. ST bridges the gap between imaging and high-throughput _____ techniques.

4. Grammar in Action: Passive Voice in Scientific Writing

✓ Form:

Passive = be (in the tense of the verb in the active form) + past participle

✓ Why Use It:

- Focuses on **action or result**, not the doer
- Sounds **objective, formal**, and **scientific**
- Useful when the **actor is unknown, unimportant, or obvious**

✓ Tips:

- Always match the verb **"be"** to the tense of the active sentence.
- Only **transitive verbs** (those that take an object) can be turned into passive voice.
- Often, the **agent ("by...")** can be omitted if it's not essential.

Examples

Researchers sequence RNA. → **RNA is sequenced.**

- *The team developed ST platforms.* → **ST platforms were developed.**
- *Scientists will integrate multi-omics data.* → **Multi-omics data will be integrated.**

Changing Tense from Active to Passive Voice

Tense	Active Voice	Passive Voice
Present Simple	Scientists **analyze** the data.	The data **are analyzed** (by scientists).

Tense	Active Voice	Passive Voice
Present Continuous	Scientists **are analyzing** the data.	The data **are being analyzed**.
Past Simple	The team **mapped** the tissue.	The tissue **was mapped**.
Past Continuous	The lab **was processing** the samples.	The samples **were being processed**.
Present Perfect	Experts **have developed** new tools.	New tools **have been developed**.
Past Perfect	Researchers **had used** the software.	The software **had been used**.
Future Simple	AI **will process** the results.	The results **will be processed**.
Future Perfect	They **will have sequenced** the genes.	The genes **will have been sequenced**.
Modal + Verb	They **can detect** mutations.	Mutations **can be detected**.

Activity 10:the Passive Voice in Scientific Discourse

Transform each **active voice sentence** into a **passive voice sentence**. Omit the agent ("by…") **only when it is not necessary** or if it sounds redundant or irrelevant. Use correct tense and scientific style.

1. Researchers are revolutionizing cancer diagnostics with spatial transcriptomics.
2. Scientists have developed high-resolution sequencing platforms like Slide-seq and Stereo-seq.
3. The 10x Genomics team introduced the Visium platform in earlier studies.
4. Tumor-associated macrophages suppress T-cell activity in immunosuppressive niches.
5. The review by Rao et al. outlines the technological evolution of ST methods.

6. Laboratories will integrate ST data with histopathology to refine cancer staging.
7. AI algorithms are enhancing pattern recognition in spatial transcriptomic datasets.
8. Bioinformaticians had overlooked the spatial dimension in early transcriptomic analyses.
9. Multi-omics researchers will have identified key protein-RNA interactions by 2026.
10. The article discusses how ST has revealed metabolic heterogeneity across tumor zones.

5. Speak & Debate

Activity 11:What is Spatial Transcriptomics?

- **Scenario**:
  Imagine you're explaining to a friend who knows little about science what **spatial transcriptomics** is and how it helps in cancer research.

1. **Task**:
   In **1–2 minutes**, explain:
   - What spatial transcriptomics means.
   - How it helps scientists understand cancer better.
   -
2. **Follow-up Question**:
   After your explanation, your partner will ask one simple question:
   - "Why is spatial transcriptomics important in cancer research?"

6. Writing for Impact

Write a paragraph to explain how computers and software tools help scientists analyze spatial transcriptomics (ST) data in cancer research. In your response, explain the role of technology in processing large and complex data, and discuss how this helps scientists improve their understanding of cancer and its treatments.