

CS 223

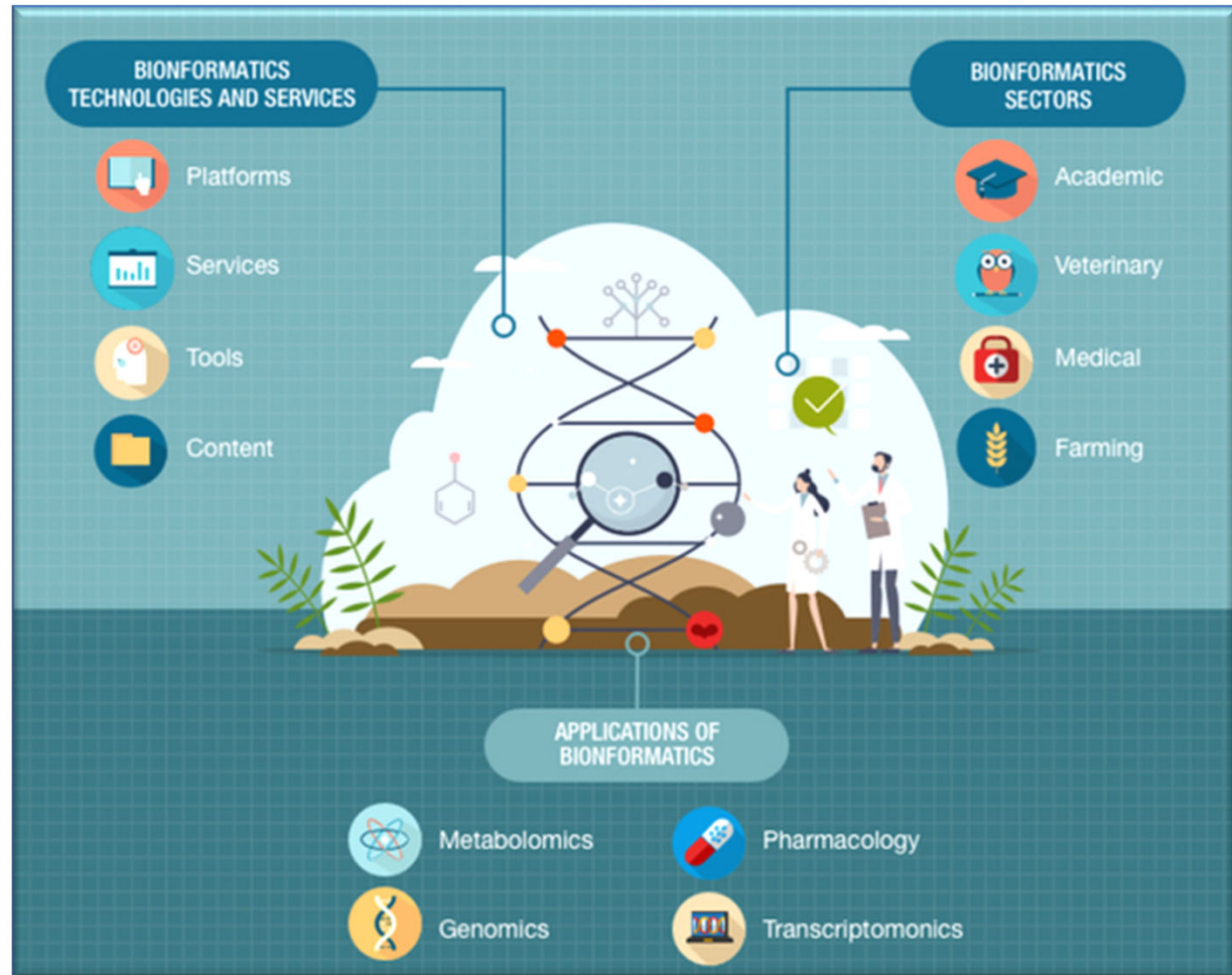
Bioinformatics

Lecture

Module 1

Week 1 p2

Leonard Wesley
Computer Science Dept
San Jose State Univ



Agenda

- Official Course Description & Canonical Syllabus
- What Industry Wants & Needs In M.S. Bioinformatics Graduates
- Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus
- Survey of Computational & Algorithmic Background Knowledge

Official Course Description & Canonical Syllabus.1

- Official Course Description:

Description

The course investigates algorithms for solving computational problems in bioinformatics, and the life processes underlying those algorithms. Metagenomics, bioinformatic machine learning, single-cell RNA sequencing, variant discovery. Possible additional topics: genome assembly, advanced phylogenetics, long-read sequencing, and structural bioinformatics.

Prerequisite(s): CS 123B / BIOL 123B / SE 123B. Allowed Declared Major: Computer Science, Bioinformatics, Data Science.

Official Course Description & Canonical Syllabus.2

- Canonical Syllabus

Learning Outcomes

Upon successful completion of this course, students will:

1. SLO-1 DEVELOPING BIOINFORMATICS COMPUTATIONAL SKILLS: Know how to analyze and predict the performance of bioinformatic algorithms, e.g., analyze recursive and iterative implementations of various algorithms used for correlation analysis, PCA, pattern analysis, ..., and how they can be used to help answer bioinformatics questions of interest.
2. SLO-2 EVIDENTIAL REASONING AND BIOINFORMATICS: Introduction to the evidential reasoning (ER) calculus that is a generalization of traditional probabilistic and statistical inference. Examples of how ER can be used to help answer bioinformatics related questions with greater fidelity than traditional methods will be presented.
3. SLO-3 ADVANCED HMMs: Discuss some of the limitations of traditional HMMs. Introduce advanced HMMs such as profile HMMs, jumping HMMs, pair-HMMs, sub-HMMs, and phylo-HMMs.
4. SLO-4 EPIGENETICS: Have a sufficient introductory understanding of epigenetics in order to equip students with the knowledge needed to investigate diseases that are caused by epigenetic mechanisms and being able to development a simple analysis pipeline that will be used throughout the remaining semester and integrated with the remaining course topics.

What Industry Wants & Needs In M.S. Bioinformatics Graduates.1

- **Employers in biotechnology, pharmaceuticals, healthcare, agriculture, and public research now rely on professionals who can turn complex biological data into usable evidence.**
 - Employment for bioinformatics degree graduates is increasing due to growing integration of data analytics in healthcare and pharmaceuticals, with a 15% growth rate projected through 2030.
 - Projected job growth is driven by advances in genomics and personalized medicine.
 - Specializing in areas like machine learning or clinical informatics significantly enhances long-term career prospects amid evolving industry demands and technological innovation.
- **Which Industries Hire the Most Bioinformatics Degree Graduates?**
 - **Bioinformatics graduates:** are hired wherever biological data must be analyzed at scale. The largest opportunities tend to cluster in industries with strong research budgets, large datasets, regulatory requirements, or direct applications in healthcare and product development.
 - **Pharmaceutical companies:** These employers use bioinformatics in drug target discovery, preclinical research, biomarker analysis, clinical trial design, and safety analysis.
 - **Biotechnology firms:** Biotech companies rely on bioinformatics for genetic testing, therapeutic development, biologics, diagnostics, and bioengineered products.
 - **Healthcare providers:** Hospitals, academic medical centers, and specialty clinics may use bioinformatics in precision medicine, oncology genomics, pathology, and population health research.
 - **Academic and government research:** Universities, public research institutes, and government-funded laboratories hire graduates to support disease research, evolutionary biology, epidemiology, genomics, and other scientific programs.
 - **Agriculture:** Agricultural organizations apply bioinformatics to crop improvement, plant genetics, disease resistance, animal genomics, and sustainable food production.

What Industry Wants & Needs In M.S. Bioinformatics Graduates.2

- **Programming expertise:** Python, R, and Java are commonly useful for scripting, statistical analysis, pipeline development, visualization, and tool building. Employers value candidates who can show working code, not just list languages on a resume.
- **Statistical competence:** Bioinformatics work often involves uncertainty, noise, sample bias, and high-dimensional data. Graduates need enough statistical knowledge to choose appropriate methods and avoid overinterpreting results.
- **Biological insight:** Molecular biology, genetics, genomics, and cell biology help professionals understand what the data represents. Without biological context, technical analysis can lead to misleading conclusions.
- **Data handling skills:** Employers need graduates who can clean, structure, store, document, and retrieve large datasets. Database familiarity, workflow reproducibility, and data integrity practices are important in both research and regulated environments.
- **Analytical problem-solving:** Bioinformatics projects rarely follow a perfect script. Professionals must troubleshoot failed pipelines, inconsistent inputs, incomplete metadata, and unexpected findings.
- **Communication abilities:** The ability to explain methods, limitations, and results in plain language is a practical career advantage. Bioinformatics professionals often serve as translators between computational teams and biological or clinical teams.

What Industry Wants & Needs In M.S. Bioinformatics Graduates.3

- **How AI Is Changing Demand for Bioinformatics Professionals:**
 - **Automation of routine tasks:** AI tools can assist with data cleaning, pattern detection, preliminary analysis, and workflow acceleration. This may reduce demand for purely repetitive tasks, but it increases the value of professionals who can validate outputs and interpret biological meaning.
 - **New specialist roles:** Employers need people who can develop, test, maintain, and explain machine learning models for biological datasets. These roles sit at the intersection of biology, computer science, statistics, and research ethics.
 - **Higher skill expectations:** Candidates are increasingly expected to understand Python, data science methods, model evaluation, and machine learning frameworks in addition to core bioinformatics concepts.
 - **Changing team structures:** Biotechnology and healthcare organizations are building teams that combine wet-lab scientists, clinicians, data engineers, AI specialists, statisticians, and bioinformaticians. Graduates who can collaborate across these groups are more competitive.
 - **Greater need for judgment:** AI outputs can be wrong, biased, poorly documented, or biologically implausible. Employers still need professionals who can question assumptions, check reproducibility, and understand the limits of automated tools.

What Industry Wants & Needs In M.S. Bioinformatics Graduates.4

• REFERENCES:

- Bioinformatics - Career overview <https://www.opasis.com/public/school/blogdetails/Bioinformatics+-+Career+overview>
- Bioinformatics 2025 Outlook: Thoughts from Bioinformaticians <https://www.fiosgenomics.com/bioinformatics-2025-outlook-thoughts-from-bioinformaticians/>
- National Employment Trends: 19-1029.01 - Bioinformatics Scientists <https://www.onetonline.org/link/localtrends/19-1029.01>
- AI's Impact on Jobs in Computational Biology & Drug Discovery <https://www.symbioti.ca/how-ai-is-creating-and-replacing-jobs-in-computational-biology-and-drug-discovery>
- AI Compute Demand in Biotech: 2025 Report & Statistics | IntuitionLabs <https://intuitionlabs.ai/articles/ai-compute-demand-biotech>
- Dr. Omics Education; <https://dromicsedu.com/blog-details/bioinformatics-salary-report-2024-highest-paying-jobs--skills>
- Bioinformatics Salary Guide 2025: What to Expect? <https://www.biotechnika.org/2025/03/bioinformatics-salary-guide-2025-what-to-expect/>
- Studying Bioinformatics: Is it Worth it? <https://www.datascienceblog.net/post/commentary/studying-bioinformatics-is-it-worth-it/>
- Top 3 Skills in Bioinformatics to Secure a Job in 2026 - Bversity <https://bversity.io/blog/skills-in-bioinformatics-to-secure-a-job-in-2026>
- Occupation Profile for Bioinformatics Scientists | CareerOneStop <https://www.careeronestop.org/Toolkit/Careers/Occupations/occupation-profile.aspx>

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus

- Which topics underlie all course alternatives & tweaks:
 - Explore how to model complex data, analyze algorithm performance, and implement custom structures from scratch.
 - Python/R programming, Big-O analysis, recursion, linear and hierarchical data structures.
 - Implementations of advanced trees, priority queues, and graphs using Pythonic paradigms and
 - Emphasis will be placed on algorithmic analysis through divide-and-conquer, greedy, and dynamic programming strategies.
 - Evaluate algorithm complexity using Big-O notation, solve recurrence relations, and explore NP-completeness to identify the computational limits of modern problem-solving.
- That Is: All alternatives will begin with “SLO-1 DEVELOPING BIOINFORMATICS COMPUTATIONAL SKILLS:”
 - Know how to analyze and predict the performance of bioinformatic algorithms, e.g., analyze recursive and iterative implementations of various algorithms used for correlation analysis, PCA, pattern analysis, ..., and how they can be used to help answer bioinformatics questions of interest.
 - DURATION: ~2 weeks.

ALTERNATIVE 1

Stay With Current Course Description & Canonical Syllabus

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.1

- **Alternative 1: Current Course Description & Canonical Syllabus**
 - **SLO 2: EVIDENTIAL REASONING AND BIOINFORMATICS:** Introduction to the evidential reasoning (ER) calculus that is a generalization of traditional probabilistic and statistical inference. Examples of how ER can be used to help answer bioinformatics related questions with greater fidelity than traditional methods will be presented.
 - **Computational Aspects:** Iterative vs Recursive time-complexity analysis, from equations to algorithm design & development, Python module & pipeline development.
 - **Duration:** ~2 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.2

- **Alternative 1: Current Course Description & Canonical Syllabus**

- **SLO 3: ADVANCED HMMs:** Discuss some of the limitations of traditional HMMs. Introduce advanced HMMs such as profile HMMs, jumping HMMs, pair-HMMs, sub-HMMs, and phylo-HMMs.
- **Computational Aspects:** Trees, tree search & retrieval, cluster building and analysis, Python module & pipeline development.
- **Duration:** ~4 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.3

- **Alternative 1: Current Course Description & Canonical Syllabus**

- **SLO 4: EPIGENETICS:** Have a sufficient introductory understanding of epigenetics and single-cell RNA-seq and variant analysis in order to equip students with the knowledge needed to investigate diseases that are caused by epigenetic mechanisms and being able to development a simple analysis pipeline that will be used throughout the remaining semester and integrated with the remaining course topics.
- **Computational Aspects:** API development, advanced tree/hash search & retrieval, divide and conquer, time and complexity analysis, Python module & pipeline development.
- **Duration:** ~4 weeks.

ALTERNATIVE 2

Biopharmaceutical Small Molecule Drug Development

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.1

- **Alternative 2: Biopharmaceutical Small Molecule Drug Development**

- **SLO 2:** Chemical DBs and Properties of Small Molecule Compounds:
- **Computational Aspects:** Iterative vs Recursive time-complexity analysis, tree and table design, development, and implementation. APIs, Data Analytics (data collection, data cleaning), Python module & pipeline development.
- **Duration:** ~2 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.2

- **Alternative 2: Biopharmaceutical Small Molecule Drug Development**
 - **SLO 3:** Data Analytics:
 - **Computational Aspects:** Iterative vs Recursive time-complexity analysis, hash table design, development, and implementation. Data Analytics (data normalization, analysis, visualization, and interpretation), Python module & pipeline development.
 - **Duration:** ~3-4 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.3

- **Alternative 2: Biopharmaceutical Small Molecule Drug Development**

- **SLO 4:** Machine-Learning-Based Drug Prediction :
- **Computational Aspects:** Iterative vs Recursive time-complexity analysis, hash table design, development, and implementation. Data Analytics (data normalization, analysis, visualization, and interpretation), Python module & pipeline development.
- **Duration:** ~3-4 weeks.

ALTERNATIVE 3

AI-Assisted Bioinformatics

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.1

- **Alternative 3: AI-Assisted Bioinformatics**
 - **SLO 2: Intro To AI and AI-Assisted Bioinformatics Portals:**
 - **Computational Aspects:** AI-Basics, Search, Representation, and Reasoning algorithms in AI.
 - **Duration:** ~1 to 2 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.2

- **Alternative 3: AI-Assisted Bioinformatics**

- **SLO 3: APIs To AI-Assisted Bioinformatics Portals:**

- **Computational Aspects:** Develop Python-Based APIs to Stanford U.’s Biomni Lab, Arc Institute’s EVO/EVO2, and/or Cong-Lab’s CRISPR-GPT Portals, and pipeline development.

- **Example Industry Use:**

- Ginko Bioworks, Boston MA – Uses Biomni for complex transcriptomic and cell-painting workflows.
 - Ono Pharma, Osaka, Japan – Uses Biomni for drug discovery
 - Goodfire, San Francisco CA – Uses EVO/EVO2 to build mechanistic interpretability tools for peering into the inner feature representations of genomic AI models.
 - Stanford University, UC Berkeley, and UCSF - Use EVO/EVO2 for designing synthetic bacteriophages and assessing genetic variants.

- **Duration:** ~4 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.3

- **Alternative 3: AI-Assisted Bioinformatics**
 - **SLO 4: Using AI-Assisted Bioinformatics APIs To Investigate A Bioinformatics Problem:**
 - **Computational Aspects:** Using Python-Based APIs to Stanford U.’s Biomni Lab, Arc Institute’s EVO/EVO2, and/or Cong-Lab’s CRISPR-GPT & pipeline development. Implementations of advanced algorithmic analysis to investigate selected bioinformatics problems.
 - **Duration:** ~6 weeks.

ALTERNATIVE 4

Analysis Of Genomic Sequences

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.1

- **Alternative 4: Analysis of Genomic Sequences**
 - **SLO 2: Intro To RNA-seq, BAM Files, Gene quantification.**
 - **Computational Aspects:** Data Structures and related Python libraries to represent and manipulate genomic data/
 - **Duration:** ~1 to 2 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.2

- **Alternative 4: Analysis of Genomic Sequences**
 - **SLO 3: Differential Expression and Pipeline Integration.**
 - **Computational Aspects:** Pythonic data structures and algorithms suitable for gene quantification, cleaning, sorting, plotting to support differential expression analysis.
 - **Duration:** ~2 to 3 weeks.

Alternatives & “Tweaks” To The Official Course Description & Canonical Syllabus.3

- **Alternative 4: Analysis of Genomic Sequences**
 - **SLO 4: Pipeline Integration for functional enrichment and review.**
 - **Computational Aspects:** Continued pipeline development, data structures and algorithms to semi-automate conclusions and decisions RE datasets.
 - **Duration:** ~3 to 4 weeks.

Your Tasks

- Meet with classmates to decide which syllabus to use. If needed, some time during class on Tuesday 8/25/2026 can be used for discussion just among students, i.e., if desired, the instructor will leave to provide privacy.
- By Wednesday, 8/26/2026, complete the anonymous “Syllabus Selection” survey.
 - If one or more students do not complete this assignment or select the canonical syllabus, the canonical syllabus will be used.
 - If one or more students do not agree to follow the syllabus that the majority of students wish to use, the canonical syllabus will be used.
- Complete a hardcopy of the “Computational Background” survey today before leaving class.

Q&A
DONE