Course Overview - Modeling Biological Data

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

February 2025

EPFL - BMI - UPLAMANNO

Table of contents

Why BIOENG-210?

Why BIOENG-210?

- **Modern Biology Challenge:**
  - Increasing reliance on large datasets
  - Complex data processing needs
  - Multi-dimensional analysis
- **Course Focus:**
  - Theoretical foundations
  - Analytical techniques
  - Software tools
  - Data-driven reasoning

Learning Outcomes

- **Technical Skills:**
  - Analyze multidimensional biological data
  - Apply statistical models
  - Design analysis pipelines
- **Critical Thinking:**
  - Select appropriate methods
  - Interpret results meaningfully
  - Connect theory to practice
- **End Goal:**
  - Independent data analysis capability
  - Sound statistical reasoning
  - Biological insight extraction

Course Structure

Weekly Format:

- 90 min Lecture
- 100 min Exercise session

Learning Flow:

- Theory in lectures
- Practice in exercises
- Continuous feedback cycle

Lecture Format

- **Before Lecture:**
  - Slides available one week ahead
  - Preview material
  - Optional reading suggestions
- **During Lecture:**
  - Theory and concepts
  - Interactive discussions
  - Mock exam questions
- **After Lecture:**
  - Detailed lecture notes
  - Exercise material release
  - Supplementary readings

Exercise Sessions

- **Format:**
  - Jupyter notebooks in Python
  - Real-world biological problems
  - Two main problems per session
- **Workflow:**
  - Material released on lecture day
  - 50% completion during class
  - Two-week submission window
- **Submission:**
  - PDF with requested plots
  - Feedback within one week
  - No late submissions accepted

Assessment Structure

- **Exercise Submissions (35%):**
  - Focus on data visualization
  - Grading scale: 3-6 per plot
  - Regular feedback cycles
- **Final Exam (65%):**
  - Multiple Choice Questions
  - Closed-book format
  - Mock questions in lectures
  - Practice exam session

Using AI Tools

- **Acceptable Use:**
  - Code completion
  - API documentation help
  - Bug fixing assistance
  - Syntax suggestions
- **Not Allowed:**
  - Complete exercise solutions
  - Analysis interpretations
  - Direct problem solving
- **Recommended Setup:**
  - VSCode + Github Copilot
  - Basic AI models (not Chat)

Data in Biological Sciences

What is Biological Data Science?

"From measurements to insights"

- **Key Components:**
  - Data collection & experimental design
  - Statistical modeling & inference
  - Machine learning & prediction
  - Biological interpretation

Fundamental Data Types of Biological Data

"Different data types require different approaches"

Data Type	Example	Format	Storage	Key Properties
Strings	DNA sequence	ATCG alphabet	FASTA	Discrete, ordered
Binary	Mutations	0/1 states	VCF	Sparse, categorical
Scalars	qPCR readout	Real numbers	Tables	Continuous, noise
Vectors	Expression	Feature arrays	CSV	High-dimensional
Tables	scRNA-seq	Cells $\times$ genes	MTX/h5	Sparse, structured
Images	Microscopy	X$\times$Y pixels	TIFF	Resolution, contrast
Movies	Live imaging	X$\times$Y$\times$T frames	TIFF stacks	Temporal dynamics

Common Biological Experiments and Their Data

Experiment	Primary Data	Dimensions	Covariates
RNA sequencing	Expression table	20K genes $\times$ samples	Conditions, time
Flow cytometry	Marker intensities	15 proteins $\times$ cells	Populations, treatment
Calcium imaging	Fluorescence traces	103 neurons $\times$ 104 frames	Stimuli, behavior
Cell tracking	Position matrices	102 cells $\times$ time $\times$ XY	Division events, speed
Histology	Tissue sections	103$\times$103 pixels $\times$ markers	Region, pathology

What is Data Modeling?

The Art of Data Modeling

"Different paths from data to understanding"

"Data modeling in biological data science is the process of creating mathematical abstractions that capture relevant patterns in experimental measurements, while accounting for biological variability and technical noise. It bridges the gap between raw observations and scientific understanding."

Two Complementary Approaches:

- **Hypothesis-Driven (Top-down):**
  - Prior knowledge → Mathematical model
  - Parameters with biological meaning
  - Test specific mechanisms
  - **Example:** Neuron firing models
- **Data-Driven (Bottom-up):**
  - Pattern discovery in data
  - Flexible statistical models
  - Learn structure automatically
  - **Example:** Cell type identification

Both approaches aim to:

- Separate signal from noise
- Find meaningful structure
- Make predictions
- Guide new experiments

Multiple Perspectives in Data Analysis

- Data analysis methods have evolved independently in different fields - statistics, machine learning, signal processing, and applied mathematics.
- The same concepts and procedures often appear with different names and interpretations.Understanding these multiple perspectives and correspondences helps us:
  - Choose the right tools for each problem
  - Recognize connections between methods
  - Translate between different fields
  - Build deeper intuition

Multiple Perspectives in Data Analysis

- Data analysis methods have evolved independently in different fields - statistics, machine learning, signal processing, and applied mathematics.
- The same concepts and procedures often appear with different names and interpretations.Understanding these multiple perspectives and correspondences helps us:
  - Choose the right tools for each problem
  - Recognize connections between methods
  - Translate between different fields
  - Build deeper intuition

Perspectives to Data Analysis

Multiple Perspectives in Data Analysis

Geometric View

- Data points as coordinates in space
- Similar samples = close points
- Interpret procedures geometrically:
  - Dimension reduction = projection
  - Data transformation = rotation
  - Normalization = rescaling

Probabilistic View

- Data as random samples
- Uncertainty in measurements
- Interpret results probabilistically:
  - Predictions → probability distributions
  - Differences → statistical significance
  - Clusters → membership probability

Multiple Perspectives in Data Analysis

Geometric View

- Data points as coordinates in space
- Similar samples = close points
- Interpret procedures geometrically:
  - Dimension reduction = projection
  - Data transformation = rotation
  - Normalization = rescaling

Probabilistic View

- Data as random samples
- Uncertainty in measurements
- Interpret results probabilistically:
  - Predictions → probability distributions
  - Differences → statistical significance
  - Clusters → membership probability

Multiple Perspectives in Data Analysis

Signal Processing View

- Biological signal + technical noise
- Multiple sources of variation
- Focus on extraction/denoising:
  - Separate components
  - Filter noise
  - Enhance features

Optimization View

- Let algorithms find optimal solution
- Balance multiple objectives
- Express goals mathematically:
  - Classification → minimize errors
  - Reconstruction → maximize fidelity
  - Clustering → maximize separation

Multiple Perspectives in Data Analysis

Signal Processing View

- Biological signal + technical noise
- Multiple sources of variation
- Focus on extraction/denoising:
  - Separate components
  - Filter noise
  - Enhance features

Optimization View

- Let algorithms find optimal solution
- Balance multiple objectives
- Express goals mathematically:
  - Classification $\rightarrow$ minimize errors
  - Reconstruction $\rightarrow$ maximize fidelity
  - Clustering $\rightarrow$ maximize separation