

BIO-369

Randomness and information in biological data

Bitbol Anne-Florence

Cursus	Sem.	Type
Life Sciences Engineering	BA6, MA2, MA4	Opt.

Language of teaching	English
Credits	4
Session	Summer
Semester	Spring
Exam	Written
Workload	120h
Weeks	14
Hours	4 weekly
Courses	2 weekly
Exercises	2 weekly
Number of positions	

Summary

Biology is becoming more and more a data science, as illustrated by the explosion of available genome sequences. This course aims to show how we can make sense of such data and harness it in order to understand biological processes in a quantitative way.

Content

Recently, biology has become more and more a data science. For instance, progress in sequencing has caused an explosion of available genome sequences. How can we make sense of such data and harness it in order to understand biological processes in a quantitative way? In many cases, biological data can be understood as being sampled from distributions of random variables. This course will first show the importance of randomness in biology. Then it will introduce some ways of extracting information from biological data, and a simple method to approximately infer the probability distributions underlying biological data. Some notions of statistics, information theory and statistical physics will be introduced, always with concrete applications to biological data in mind. Exercises and numerical projects will allow students to apply the methods to real biological data. The course will be organized as follows:

Part I: Randomness in biological processes and biological data

1. Randomness and random variables. Luria-Delbrück experiment.
2. Importance of thermal fluctuations at the cellular scale. Biopolymers, biomembranes, chemical reactions.
3. Random walks. Population genetics. Examples of other applications (foraging, transcription factors binding to DNA).

Part II: Extracting information from biological data

1. Quantifying randomness in data: entropy. Link to thermodynamics. Entropy in neuroscience data.
2. Quantifying statistical dependence: correlation; mutual information. Applications to neuroscience and to sequencing data.
3. Inferring probability distributions from data: introduction to maximum entropy inference. Prediction of protein structure from multiple sequence alignments.
4. Finding relevant dimensions in data: dimensionality reduction methods (linear and nonlinear).

Keywords

Biological data, data science, sequencing data, neuroscience, population genetics, random variable, random walk, information theory, statistical physics, entropy, mutual information, inference, dimensionality reduction.

Learning Prerequisites**Required courses**

Analysis; probability and statistics; linear algebra; general physics; programming.

Recommended courses

Introductory machine learning.

Learning Outcomes

By the end of the course, the student must be able to:

- Manipulate notions of statistics, information theory and statistical physics.
- Apply these notions to biological data.
- Analyze biological data in a quantitative way.
- Perform data analysis in Python.

Teaching methods

Lectures, exercises, programming labs.

Assessment methods

Written final exam during the exam session, graded numerical mini-project.

Resources

Bibliography

Reference textbooks:

- P. Nelson, Physical Modeling of Living Systems, WH Freeman, 2014
- D. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2003

More advanced textbooks:

- W. Bialek, Biophysics: Searching for Principles, Princeton University Press, 2012
- T. Cover and J. Thomas, Elements of Information Theory, 2nd ed, Wiley Interscience, 2006

Ressources en bibliothèque

- [Biophysics: Searching for Principles](#)
- [Physical Modeling of Living Systems](#),
- [Information Theory Inference and Learning](#)
- [Elements of Information Theory](#),