# Bioinformatics

## Guide to AI and Data Courses · Part VI: Systems and Applications

Biological data | Sequence alignment | Genome assembly | Gene expression | Protein structure | Neighbouring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides: https://sercankulcu.github.io/files/ai_courses/slides/Chapter_39_Bioinformatics.html

INTRODUCTION

# What Question Does It Ask?

> **How are biological sequences and molecular data analyzed computationally?**

Bioinformatics analyses biological data such as DNA, RNA and protein sequences with algorithms and statistics. The course covers problems such as **sequence alignment**, genome assembly, gene expression analysis and protein structure prediction, using methods that cope with the size and noise of this data.

## At a glance

| | | |
|---|---|---|
| **APPROACH**<br>biology, algorithms and statistics | **INPUT**<br>DNA, protein sequences, expression data | **OUTPUT**<br>alignment, structure, function prediction |
| **SUCCESS MEASURE**<br>biological validity | **MATH LOAD**<br>medium-high | **DATA NEEDS**<br>huge sequence datasets |

# A Short History

**1953**
Watson and Crick, drawing also on Franklin's data, propose the double-helix structure of DNA.

**1970**
Needleman and Wunsch publish a global sequence alignment algorithm based on dynamic programming.

**1981**
Smith and Waterman publish the local sequence alignment algorithm.

**1990**
BLAST is published: fast, heuristic similarity search in large databases.

**2003**
The Human Genome Project is completed; sequencing costs begin to fall rapidly.

**2020**
AlphaFold 2 predicts protein structure at CASP14 with accuracy close to experimental methods.

THE COURSE

# Core Concepts

**Sequence alignment**
Global and local alignment; scoring matrices.

**Sequence databases**
GenBank, UniProt, PDB; similarity search.

**Hidden Markov model**
Gene finding and protein family profiles.

**Genome assembly**
Building a genome from short reads; de Bruijn graphs.

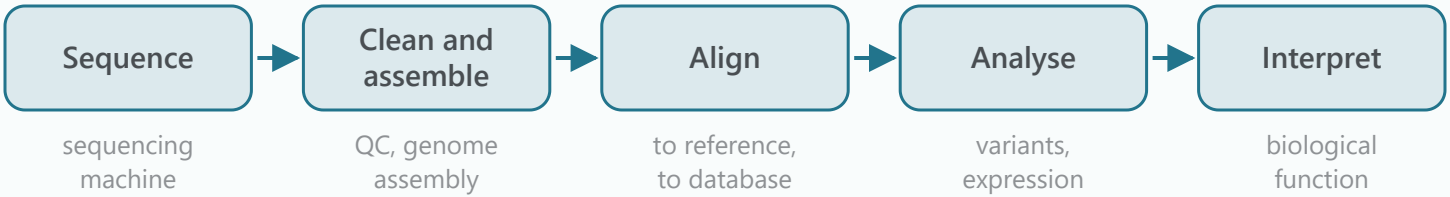
**Gene expression**
Measuring how active each gene is with RNA-seq.

**Phylogenetics**
Building the family tree of species from sequences.

**Protein structure**
Predicting 3D shape from the amino acid sequence.

**Multiple testing**
Controlling the false discovery rate over thousands of genes.

# A Typical Workflow




**Example:** A bacterium taken from a patient is sequenced, the short reads are assembled into a genome, and its genes are compared with known sequences in a database using BLAST. If antibiotic resistance genes are found, the physician is told which drugs will not work.

THE COURSE

# What Does the Course Cover?

1. A molecular biology primer and biological databases
2. Pairwise alignment: Needleman-Wunsch, Smith-Waterman
3. Scoring matrices and similarity search with BLAST
4. Multiple sequence alignment
5. Hidden Markov models and gene finding
6. Motif finding
7. Genome assembly: de Bruijn graphs
8. Read mapping and variant calling
9. Phylogenetic trees
10. Gene expression with RNA-seq; multiple testing
11. Clustering and classifying gene expression
12. Protein structure prediction and AlphaFold

## Skills you will gain

handling sequences with Biopython | similarity search with BLAST | read mapping with BWA and SAMtools | expression analysis with Bioconductor | implementing an alignment algorithm

A typical semester; the order and coverage vary between universities.

TRY IT

# Align Two Gene Fragments

How similar are DNA fragments of the same gene in two species? The Needleman-Wunsch table finds the best alignment. Change the sequences and scores, then click a cell to see how it was computed.

| | – | G | C | A | T | G | C | A |
|---|---|---|---|---|---|---|---|---|
| – | **0** | −1 | −2 | −3 | −4 | −5 | −6 | −7 |
| G | −1 | **1** | **0** | −1 | −2 | −3 | −4 | −5 |
| A | −2 | 0 | 0 | **1** | 0 | −1 | −2 | −3 |
| T | −3 | −1 | −1 | **0** | 2 | 1 | 0 | −1 |
| T | −4 | −2 | −2 | −1 | **1** | 1 | 0 | −1 |
| A | −5 | −3 | −3 | −1 | 0 | **0** | 0 | 1 |
| C | −6 | −4 | −2 | −2 | −1 | −1 | **1** | 0 |
| A | −7 | −5 | −3 | −1 | −2 | −2 | 0 | **2** |

Seq 1 GATTACA Seq 2 GCATGCA

Match **+1**

Mismatch **−1**

Gap **−1**

```
G-ATTACA
| | |·||
GCA-TGCA
```

Score **2**: 5 matches, 1 mismatches, 2 gaps (orange path).

Bioinformatics turns evolution into a **scoring model**: filling the table takes m×n steps, while trying every alignment is exponential. The same idea underlies tools such as BLAST.

CONTEXT

# Prerequisites

Courses worth knowing before this one, and why they are needed.

| Course | Why it is needed |
| --- | --- |
| Machine Learning | gene expression classification, protein structure |
| Data Mining | sequence motifs, clustering |
| Algorithms | dynamic programming: the method behind sequence alignment |
| Data Structures | hash tables, trees and graphs: k-mer counting and genome assembly |
| Biology | the basics of DNA, RNA, proteins and gene expression |

CONTEXT

# Are You Ready for This Course?

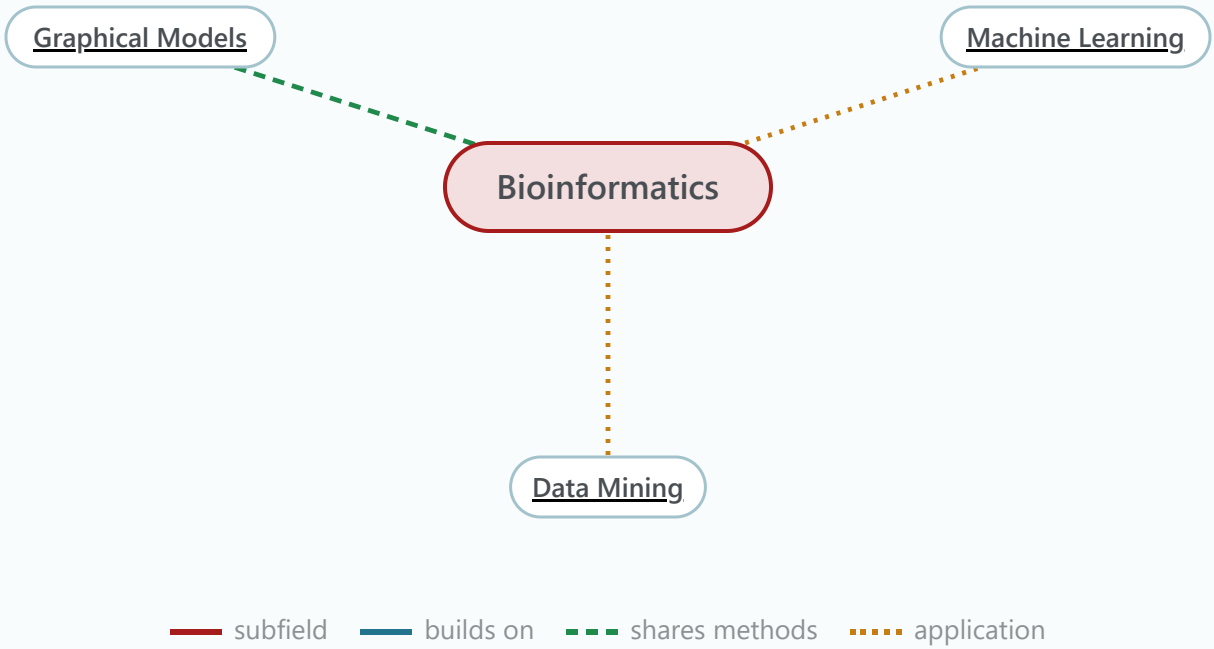
Ask yourself; for every "no", the right-hand column shows where to start.

| Ask yourself | If not, start with |
|---|---|
| Can you find the longest common subsequence of two strings by dynamic programming? | Dynamic programming from an algorithms course |
| Can you explain how information flows from DNA to RNA to protein? | High school biology or a short molecular biology primer |
| Can you count all substrings of length k in a text with a hash table? | Hash tables from a data structures course |
| Can you train a classifier and interpret the result of a clustering? | Machine Learning |

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

# Neighbouring Courses



# Overlapping Courses

| Relation | Course | Shared ground |
|---|---|---|
| application | This course applies the methods of Machine Learning | gene expression classification, protein structure |
| application | This course applies the methods of Data Mining | sequence motifs, clustering |
| shares methods | Shares methods with Probabilistic Graphical Models | hidden Markov models for sequence alignment |

# Where It Differs (1/2)

| Criterion | Bioinformatics | Machine Learning | Data Mining |
|---|---|---|---|
| **Core question** | How are biological sequences and molecular data analyzed computationally? | How does a program generalize from examples? | How are hidden, useful patterns found in large piles of data? |
| **Approach** | biology, algorithms and statistics | statistical learning | discovery-driven analysis |
| **Input** | DNA, protein sequences, expression data | labeled or unlabeled examples | large tables or transaction data |
| **Output** | alignment, structure, function prediction | a model that makes predictions | rules, clusters, outliers |
| **Success measure** | biological validity | error on unseen data | interestingness, support, confidence |
| **Math load** | medium-high | high: linear algebra, probability, optimization | medium |
| **Data needs** | huge sequence datasets | moderate to large | large |

- **Machine Learning:** ML develops general-purpose learning methods; this course applies them to sequence and expression data together with biological knowledge.
- **Data Mining:** Data mining looks for patterns in any data; this course uses alignment and motif methods specific to DNA and protein sequences.

# Where It Differs (2/2)

| Criterion | Bioinformatics | Graphical Models | Medical Imaging |
|---|---|---|---|
| **Core question** | How are biological sequences and molecular data analyzed computationally? | How do we model and query many interdependent, uncertain variables? | How do X-ray, CT and MRI images help diagnosis and treatment? |
| **Approach** | biology, algorithms and statistics | probability and graphs combined | image processing plus clinical knowledge |
| **Input** | DNA, protein sequences, expression data | variables, dependencies, data | CT, MRI, X-ray, ultrasound |
| **Output** | alignment, structure, function prediction | conditional probability, best explanation | segmentation, measurement, diagnostic suggestion |
| **Success measure** | biological validity | likelihood, inference accuracy | sensitivity, specificity, Dice |
| **Math load** | medium-high | high: probability | medium-high |
| **Data needs** | huge sequence datasets | moderate | little; labels are expensive |

- **Graphical Models:** Probabilistic graphical models are a general framework; this course uses hidden Markov models for gene finding and protein families.
- **Medical Imaging:** Medical image processing analyses images of organs and tissue; this course analyses sequence and expression data at the molecular level.

PRACTICE

# Applications

**Personalized medicine**
Choosing treatment by genetic variants.

**Outbreak tracking**
Tracing spread through virus genomes.

**Drug discovery**
Predicting target protein and molecule interactions.

**Cancer genomics**
Finding tumour-specific mutations.

**Agriculture**
Breeding productive and resilient crop varieties.

**Evolution studies**
Building species relationships from sequences.

PRACTICE

# Risks and Ethics

**Genetic privacy**
A genome identifies a person and their relatives; identities have been recovered from genetic data thought anonymous.

**Ancestry-biased data**
Genomic databases come mostly from people of European ancestry; risk scores are less accurate for other ancestries.

**Batch effects**
When samples are processed on different days or machines, a technical difference can look like a biological finding.

**Clinical overinterpretation**
Treating a variant of uncertain significance as a cause of disease can lead a patient to needless treatment or worry.

In depth: Explainable AI and AI Ethics.

PRACTICE

# Tools and Datasets

## Tools and libraries

BLAST | Biopython | Bioconductor (R) | BWA | SAMtools | HMMER | PyMOL

## Well-known datasets

GenBank | UniProt | Protein Data Bank (PDB) | Gene Expression Omnibus (GEO)

1000 Genomes Project | The Cancer Genome Atlas (TCGA)

PRACTICE

# Common Misconceptions

| Common belief | In fact |
|---|---|
| One gene means one trait. | Most traits come from many genes and the environment together; few are set by a single gene. |
| Once a genome is sequenced, its function is known. | A sequence is only a string of letters; finding what genes do is a separate, hard analysis. |
| Bioinformatics is only for biologists. | The core of the field is algorithms, data structures and statistics; computer engineers are central. |
| When thousands of genes are tested, the significant ones are real. | Multiple testing inflates false positives; corrections such as the false discovery rate are needed. |

NEXT STEPS

# After This Course

## Courses that build on or extend this one

This course is mostly an end point: it builds on the others.

## Related careers

Bioinformatician | Computational biologist | Genomics data scientist | Biostatistician | AI researcher in drug discovery

Bioinformatics brings machine learning, data mining and probabilistic models together in one application area. After this course, computational genomics, structural biology and protein language models, or neighbouring health courses such as medical image processing, are natural next steps.

NEXT STEPS

# Further Reading


- Durbin, Eddy, Krogh and Mitchison, *Biological Sequence Analysis*, Cambridge University Press, 1998.
- Jones and Pevzner, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004.
- Compeau and Pevzner, *Bioinformatics Algorithms: An Active Learning Approach*, 3rd ed., Active Learning Publishers, 2018.
- Altschul et al., "Basic Local Alignment Search Tool", *Journal of Molecular Biology* 215(3), 1990.

# The End

## Chapter 39: Bioinformatics

← Previous chapter · Map of the field · Next chapter →