# PA Bioinformatique X21

**Sebastian.Will@polytechnique.edu**
*Sarah.Berkemer@polytechnique.edu*

AMIBio

April 2023

# What is (not) a PA?

PA = Programme d'Approfondissement

- not a simple continuation of 2A
- rather: first year of a Master
  - specialization
  - specific scientific field
  - building your profile
  - concrete skills

Start thinking about

- 4A, completion of the Master program
- thesis
- begin of your career

Find essential info in the livret PA and at
`https://www.enseignement.polytechnique.fr/bioinformatique`

# What is Bioinformatics?

- increasingly larger part of modern biology not possible without Computer Science:
  **Mass of Data**
  - *high-throughput* sequencing analysis: genomics, transcriptomics, ...
  - *comparative* genomics, phylogenomics
  - classification and *annotation* of proteins and RNAs
  - *structure modelling and design* of new bio-molecules; "drug design"
  - ...
- new perspectives in **Biology**, **Pharmacology** and **Medicine**
- new challenges in **Algorithmics**, **Combinatorics**, **Machine Learning**

# What is Bioinformatics?

- increasingly larger part of modern biology not possible without Computer Science:
  **Mass of Data**
  - *high-throughput* sequencing analysis: genomics, transcriptomics, ...
  - *comparative* genomics, phylogenomics
  - classification and *annotation* of proteins and RNAs
  - *structure modelling and design* of new bio-molecules; "drug design"
  - ...
- new perspectives in **Biology**, **Pharmacology** and **Medicine**
- new challenges in **Algorithmics**, **Combinatorics**, **Machine Learning**

**Special Challenge:** multi-disciplinary; know about CS and biology

**Objective: dual competence** - interact with biologists, medics, computer sciencists, ...
in academia and industry

# What is Bioinformatics?

**= analysing - modeling - structuring - explaining - predicting - treating relevant biological information**

= (Medicine +) Biology + Computer Science ( + Math, Statistics)

= knowledge of biology and sophisticated techniques, e.g.

- data search, index structure
- combinatoric optimization
- classification, machine learning
- visualization, image analysis

# Questions/challenges for Bioinformatics

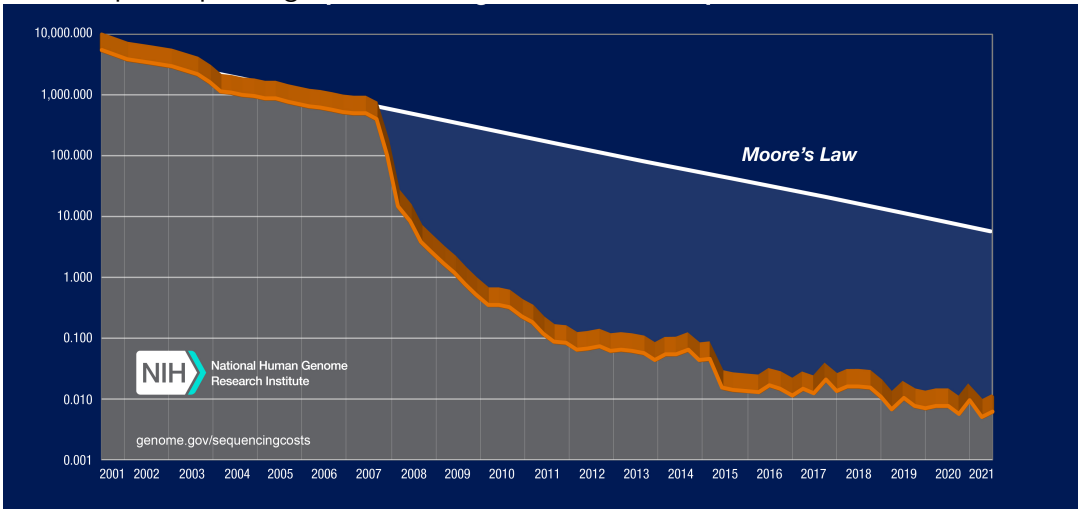
- sequences, genomics, gene identification, evolution,
- pangenomics, ..., personalized (genomic) medicine,
- structures (proteins, RNAs, complexes, ...),
- design/engineering problems: design of structures, functions, drugs,
- cell functions (metabolism, mitosis, ...),
- bio-molecule interactions (regulation of genes, regulatory networks...),
- statistics: should we be surprised to...?

Need for grammatical, combinatorial, probabilistic models
$\longrightarrow$ *algorithms:* efficient exact algorithms, approximations, heuristics, strategies; *machine learning*

# Challenge of big data

**data volume increases much faster than computing power**

For example, sequencing cost:



# Challenge of big data

**data volume increases much faster than computing power**

- search
- indexing
- filtering
- statistics
- extraction of textual information
- building and using ontologies
- high dimensions, large graphs, ...

and also engineering issues (hardware and software), systems architecture

# After the PA...

- M2 Bioinformatics in France, e.g.:
  - AMI2B de Paris-Saclay,
  - (in planning: M2 bioinformatics specialization @X,)
  - parcours BIM de UPMC,
  - Bordeaux, Aix-Marseille, Toulouse, ...
- abroad, MSc in 12 month (system 4+1), e.g.:
  - Cambridge: MPhil in Computational Biology,
  - Imperial College London :
  - MSc in Bioinformatics and Theoretical Systems Biology,
  - Edinburgh: MSc in Bioinformatics,
- McGill, Montreal: MSc in Computer Science / Bioinformatics,
- Europe, complete MSc: 2 years (system 3+2), e.g.:
  - ETH Zürich: Master in Computational Biology and Bioinformatics,
  - EPFL (catégorie 1), Master in Life Sciences Engineering
  - Copenhagen (U.of C. and DTU),
  - Germany, ...
- écoles, e.g. Agro, Mines, ...

# Continuation, examples

- Master of BioEngineering
  - e.g.: EPFL, Stanford, Berkeley
- Biology and Data Science / Biostatistics, e.g.:
  - Harvard: MSc of Science in Computational Biology and Quantitative Genetics ("Big Data", stats), MSc in Health Data Science, ...
  - Columbia,
  - Yale, ...
- Neurosciences, e.g.:
  - Oxford: MSc in Neuroscience,
  - EPFL: Master in Life Sciences Engineering / Neurosciences and Neuroengineering,

# Opportunities (after 4A)

PhD thesis (recommended) → appropriate modules in master (or PhD track)

- Industry :
  - pharmaceutical (custom drug),
  - agro-food (yields, climate, palatability, etc.),
  - biotechnologies (fuels, materials, etc.),
  - environment,
  - ...
  - IT (medical imaging, care systems, support, etc.) e.g. Dassault Systemes, IBM, GE, Siemens, ...
- Large institutes, e.g. : Curie, Pasteur, INRA, INSERM, ..., EMBL-EBI, SIB, NCBI, ...
- Academia

# Requirements for the PA

Prerequisites from BIO + INF:

- 1 Biology course in 2A: ..., BIO452 (recommended), ...
- 2 computer courses in 2A (excluding modal) ..., INF421, ..., INF442, ... (recommended)
- 1 CS modal or 1 project integrated into a course,
- MAP433 recommended,
- **check prerequisites per course**

# Rules for program choices

- Each period: 3 classes + 1 EA (or replace by long project)
  - total 8 including at least 3 in biology and 3 in computer science ( = several "projects" or article studies )
  - each period: possibility of a course outside the program
    ok if course BIO or INF
    others: must be motivated
  - *subject to schedule compatibility*
- Long project (BIO511, BIO512, BIO572, INF511)
  **recommended** "transversal" use of skills
  replaces 2 EAs
- 3A internship in Biology (BIO591) or Computer Science (INF591) ditto
  "transversal", final choice in the fall

More extreme choice → adaptation from the PA for Bio or the PA for Info, PA SDE, etc.

# PA program: classes

## Term 1

*3 classes from*

- BIO551 Immunologie et agents infectieux
- BIO553 Biotechnologies pour la médecine et l'agriculture
- BIO556 Genomes, diversity, environment and human health
- BIO557 Neurosciences
- INF550 Algorithmique avancée
- INF552 Data Visualization
- INF555 Constraint-based Modeling and Algorithms for Decision Making
- INF556 Topological data analysis

## Term 2

*obligatory*

- INF589 Computational analysis of high-throughput sequencing data

*2 classes from*

- BIO562 Biologie des systèmes moléculaires
- BIO563 Epigénétique et ARN non-codants
- INF580 Large scale mathematical optimization
- INF581 (Advanced) Topics in Artificial Intelligence
- MAP566 Statistics in action
- MAP569 Machine learning II

# PA program: projects and internship

## Term 1

*1 EA (or long project)*

| | |
|---|---|
| BIO571A | Travaux expérimentaux de génie génétique |
| BIO571B | Travaux expérimentaux en imagerie quantitative |
| INF554 | [EA] Machine learning I |
| INF573 | [EA] Image Analysis and Computer Vision |

## Terms 1&2: long projects

| | |
|---|---|
| BIO511 | Projet de Biologie |
| BIO572 | Reconstitution Personnalisée du Processus Tumoral |
| INF511 | Projet de Bioinformatique |

## Term 2

*1 EA (or long project)*

| | |
|---|---|
| BIO583 | Sciences des données en imagerie biologique |
| BIO/INF588 | Projet en bioinformatique |

## Term 3

*Stage / internship*

| | |
|---|---|
| BIO591 | Biologie et Écologie |
| INF591 | Informatique |
| INF592 | Data Science |

# INF511 Projects

Examples from past years:

- A Divide and Conquer Approach for RNA Design
- Prediction of nucleotides at protein/RNA interfaces
- Classification of protein structures
- Implementation of a stochastic simulation algorithm for metabolic networks
- Classification of electroencephalograms and detection of epileptic seizures
- Executable mathematical model of a single red blood cell
- Non-invasive prenatal diagnosis of monogenic diseases
- Benchmarking of single-cell RNA-seq
- Alignment of brain imaging data

start thinking about it before summer $\rightarrow$ we need to meet to talk about it
see also BIO511, BIO512 or BIO572

# The (not so) far future: 3A internship and 4A/M2

Start thinking about

- your internship:

  BIO591 Biology and Ecology internship (Y. Mechulam)
  `yves.mechulam@polytechnique.edu`

  INF591 Computer Science internship (O. Bournez)
  `bournez@lix.polytechnique.fr`

  INF592 Data Science internship (I. Manolescu, S. Oudot)
  `ioana.manolescu@inria.fr,steve.oudot@inria.fr`

- your ideas for 4A/M2

Internship presentations in September

Start to discuss with researchers, teachers, other students...

# Stage de recherche - examples

X14
- SFU, Vancouver : Computational tools for human pathogens outbreak monitoring using whole-genome sequencing publication
- Paris : New technology for neuropathies and data mining : Application to Cystic Fibrosis and Prediabetes screening (prix de stage)

X15
- Mc Gill : Adressage des protéines
- Mc Gill : Alignment of short query sequences against large probabilistic genomes
- MIT : Machine Learning applied to the Amyotrophic Lateral Sclerosis
- IMAG : Evaluation of distance between Thresholded Boolean Automaton Networks (prix de stage), publications

X16
- U. Sherbrooke (distance, phylogénie, cancers), U.Wien (algo ARN), Mc Gill (ML/TAD), MIT (ML/ALS),
- EBI/EMBL : Analysis of Single-Cell RNA Sequencing data from human pancreas
- Mc Gill : Leveraging Affinity Information to improve Molecular Generative Models (prix de stage)

# Timeline

- 13:15 PA presentation: Computer Science
- 14:45 PA presentation: Biology
- start of inscription (Apr 19)
  - PA choice, motivation letter, course choice
  - individual discussions
- May 03, 17h - end of inscription and validation (if OK)

- Back-to-school meeting (September)
- Individual discussion (change of courses, project, internship, 4A)
- Presentations of the 3A internships (September)
- Deadlines
  - November: limit category 1
  - December: category 2 and 3 limit, internship limit
  - January: last applications abroad

# À bientôt ...

You are welcome to discuss details / get more info ...

Sebastian Will
*sebastian.will@polytechnique.edu*

Sarah Berkemer
*sarah.berkemer@polytechnique.edu*


Bionformatics team
at DIX/LIX

`https://www.enseignement.polytechnique.fr/bioinformatique`
`https://synapses.polytechnique.fr`