Computational Science: Bioinformatics

Department of Informatics, University of Oslo

Planned start: Fall 2018

Why Bioinformatics

If you have a Bachelor's degree with a lot of computer science and mathematics and are interested in applying your knowledge on applications within the biological and medical sciences, a Master's degree in Computational Science: Bioinformatics is for you. Here you will learn to develop and apply analysis methods on data about genomes, genes and proteins that are essential for research in molecular biology and molecular medicine.

Technological advances have enabled us to detect and measure massive amounts of biological or medical data at unprecedented speed and affordability. The sequencing of the human genome and thousands of other organisms have paved the way for the identification of genetic variation at the level of individuals or even of single cells. Sequencing the DNA or RNA of tumour cells from a cancer patient may help diagnose the disease and find the best individual therapy. This is now starting to come into clinical practice, but requires efficient and robust methodology.

Hardware developments have given us smaller, faster, cheaper and more power-efficient computers. The ever increasing number of processing cores available enables massive parallelisation but may require sophisticated programming to exploit well. We use advanced algorithms and data structures, as well as mathematical models and statistical analysis to design and implement methods and tools to process, analyse and visualise molecular data. In our aim for reproducible science we prefer to work with open source software and strive for the best software development practices.

Structure and courses

The table here is an example of a suggested path for a Master of Science project, with course work the first year and thesis work the last year.

	10 ECTS	10 ECTS	10 ECTS
4th semester	Master thesis	Master Thesis	Master Thesis
3rd semester	Master thesis	Master Thesis	Master Thesis
2nd semester	Master courses	Master courses	Master courses
1st semester	Master courses	Master courses	Master courses

This program is very flexible in its structure and students may opt for starting with their thesis work from the first semester and scatter the respective course load across all four semesters. Depending on interests and specializations, there are many courses on computational science which can make up the required curriculum of course work. Furthermore, courses may be broken up in smaller modules, avoding thereby the limitation of 10 ECTS per course only. Some of these courses are listed below.

Courses. Depending on specialization, the following courses that can be included as part of the compulsory 60 ECTS needed for a Master of Science degree are described here.

Students have to choose among two of the following courses

1. CS-MATH1 (data analysis and machine learning)

2. CS-MATH2 (mathematical modelling)

3. CS-INF1 (High performance computing)

The following courses are recommended for students in this study direction:

1. INF4490 Biologically inspired computing

2. INF4350 Introductory Course in Bioinformatics

3. INF5380 High-performance computing in bioinformatics

These are also relevant, but perhaps even more relevant for the CS:Biosciences direction:

1. INF-BIO5121 High Throughput Sequencing technologies and bioinformatics analysis

2. MBV-INF4410 Bioinformatics for Molecular Biology

New courses.

1. CS-BIO1-INF: Computing for biosciences (INF4490?): INF4490 is the "Biologically inspired computing" course. It is about machine learning and artificial intelligence, classification etc. It teaches computational methods that are inspired by biological concepts like neural networks, genetic algorithms. It is highly relevant for this direction, but the new name is strongly misleading. There are also some statistics courses on prediction, classification and machine learning (e.g. STK2100, IMB9275) that should be appropriate for the CS:Bioinformatics students. Parts of those courses may be integrated. I think the course should be called just "machine learning".

2. CS-BIO2-INF: Introductory bioinformatics (INF4350+MBV3070?). INF4350 and MBV3070 are two introductory courses in bioinformatics, designed for two different audiences and with quite a different focus. Basically, INF4350 is for computer science students that will learn how bioinformatics methods work and how to make new tools, while the MBV3070 course is for biology students who would like to know which tools to use and how to apply them. The INF4350 course fits well to the CS:Bioinformatics direction, while MBV3070 fits in the CS:Bioscience direction. I think the courses should be kept mainly as they are.

3. CS-BIO3-INF: Computational Bioinformatics (INF5380?). This is a course on how to use HPC facilities and parallel programming approaches to efficiently process large datasets. It fits well into the CS:Bioinformatics direction. Perhaps it could be extended with subjects from the courses INF3380 and INF5063/9063. I think the course should keep the name "High-Performance Computing in Bioinformatics", or alternatively "Parallel computing in bioinformatics"

4. CS-BIO4-INF: Computational Physiology (INF5560?). I think this could be of interest for CS:Bioinformatics students, but fits best into the CS:Imaging and Biomedical computing or the CS:Biosciences directions.

5. CS-PHYS2: Computational Molecular dynamics in life science and materials science (new). I think this could be of interest for CS:Bioinformatics students, but also into several of the other directions.

6. CS-BIO5-INF Visualisation in bioinformatics (NEW) Perhaps a new course could be developed on visualisation of large biological datasets. How to present very large sets of data in a way that communicates efficiently. How to make interactive graphics, browsers etc that allow exploration of the data...

Graduate Certificates. The program plans to offer graduate certificates in

- Three of the courses with label CS-MATH gives a certificate in Computational Mathematics

- Three of the courses with label CS-PHYS gives a certificate in Computational Physics, Astrophysics, Chemistry, Materials Science and Geoscience

- Three of the courses with label CS-BIO gives a certificate in Computational life science.

- Three of the courses with label CS-INF gives a certificate in High-performance computing.

Dual Degrees. The program plans to offer dual degrees (more text to come)

Description of learning outcomes

The power of the scientific method lies in identifying a given problem as a special case of an abstract class of problems, identifying general solution methods for this class of problems, and applying a general method to the specific problem (applying means, in the case of computing, calculations by pen and paper, symbolic computing, or numerical computing by ready-made and/or self-written software). This generic view on problems and methods is particularly important for understanding how to apply available, generic software to solve a particular problem.

Computing competence represents a central element in scientific problem solving, from basic education and research to essentially almost all advanced problems in modern societies. Computing competence is simply central to further progress. It enlarges the body of tools available to students and scientists beyond classical tools and allows for a more generic handling of problems. Focusing on algorithmic aspects results in deeper insights about scientific problems.

The learning outcomes are subdivided in three general categories, knowledge, skills and general competence.

- **Knowledge**: A candidate from this program

    - has deep knowledge of the scientific method and computational science at an advanced level, meaning that the candidate
        1. has the ability to understand advanced scientific results in new fields
        2. has fundamental understanding of methods and tools
        3. can develop and apply advanced computational methods to scientific problems
        4. is capable of judging and analyzing all parts of the obtained scientific results
        5. can present results orally and in written form as scientific reports/articles
        6. can propose new hypotheses and suggest solution paths
        7. can generalize mathematical algorithms and apply them to new situations
        8. can link computational models to specific applications and/or experimental data
        9. can develop models and algorithms to describe experimental data
        10. masters methods for reproducibility and how to link this to a sound ethical scienfitic conduct
        11. has a thorough understanding of how computing is used to solve scientific problems
        12. knows fundamental algorithms in computational science
    - has a fundamental understanding and knowledge of scientific work, meaning that

1. the candidate can develop hypotheses and suggest ways to test these
2. can use relevant analytical, experimental and numerical tools and results to test the scientific hypotheses
3. can generalize from numerical and experimental data to mathematical models and underlying principles
4. can analyze the results and evaluate their relevance with respect to the actual problems and/or hypotheses
5. can present the results according to good scientific practices

- **Skills**: A candidate from this program

  - has a deep understanding of what computing means, entailing several or all of the topics listed below
    1. knows the most fundamental algorithms involved, how to optimize these and perform statistical uncertainty quantification
    2. has overview of advanced algorithms and how they can be accessed in available software and how they are used to solve scientific problems
    3. has knowledge and understands high-performance computing elements: memory usage, vectorization and parallel algorithms
    4. can use effeciently high-performance computing resources, from compilers to hardware architectures
    5. understands approximation errors and what can go wrong with algorithms
    6. has knowledge of at least one computer algebra system and how it is applied to perform classical mathematics
    7. has extensive experience with programming in a high-level language (MATLAB, Python, R)
    8. has experience with programming in a compiled language (Fortran, C, C++)
    9. has experience with implementing and applying numerical algorithms in reusable software that acknowledges the generic nature of the mathematical algorithms
    10. has experience with debugging software
    11. has experience with test frameworks and procedures
    12. has experience with different visualization techniques for different types of data
    13. can critically evaluate results and errors
    14. can develop algorithms and software for complicated scientific problems independently and in collaboration with other students
    15. masters software carpentry: can design a maintainable program in a systematic way, use version control systems, and write scripts to automate manual work

16. understands how to increase the efficiency of numerical algorithms and pertinent software
17. has knowledge of stringent requirements to efficiency and precision of software
18. understands tools to make science reproducible and has a sound ethical approach to scientific problems

- **General competence**: A candidate from this program

  - is able to develop professional competence through the thesis work, entailing:
    1. mature professionally and be able to work independently
    2. can communicate in a professional way scientific results, orally and in written form
    3. can plan and complete a research project
    4. can develop a scientific intuition and understanding that makes it possible to present and discuss scientific problems, results and uncertainties

  - is able to develop virtues, values and attitudes that lead to a better understanding of ethical aspects of the scientific method, as well as promoting central aspects of the scientific method to society. This means for example that the candidate
    1. can reflect on and develop strategies for making science reproducible and to promote the need for a proper ethical conduct
    2. has a deep understanding of the role basic and applied research and computing play for progress in society
    3. is able to promote, use and develop version control tools in order to make science reproducible
    4. is able to critically evaluate the consequences of own research and how this impacts society
    5. matures an understanding of the links between basic and applied research and how these shape, in a fundamental way, progress in science and technology
    6. can develop an understading of the role research and science can play together with industry and society in general
    7. can reflect over and develop learning strategies for life-long learning.

By completing a Master of Science thesis, the candidate will have developed a critical understanding of the scientific methods which have been studied, has a better understanding of the scientific process per se as well as having developed perspectives for future work and how to verify and validate scientific results.

In addition to the above more general learning outcomes, this study direction has some more specific learning outcomes. These are:

1. understand fundamental concepts in bioinformatics

2. understand basic concepts in molecular biology

3. have an overview of the most important methods and tools that are used in bioinformatics

4. understand how some of the basic methods for biological sequence analysis work

5. be capable of performing sequence analyses using existing tools

6. can argue for why methods need to be accurate and efficient

7. understand the principles of some algorithm classes like dynamic programming

8. understand machine learning concepts and methods for prediction and classification

9. be able to implement bioinformatics algorithms

10. understand how methods scale in terms of cpu and memory usage, and how this can be analysed

11. able to design tools that works efficiently on parallel computing hardware

12. able to process large datasets efficiently using parallellisation

Admission criteria: Bioinformatics

The program has a minimum course requirement of 120 ECTS (European Credit Transfer System) at the undergraduate level (bachelor degree or equivalent) in Astrophysics, bioscience, chemistry, computer science and informatics, geoscience, mathematics, materials science, mechanics and physics.

1. Of these 120 ECTS, 80 ECTS have to be within Informatics/Mathematics/Statistics (courses labeled as INF/IN, INF-MAT, MAT-INF, MAT and STK) where of 50 ECTS have to include basic mathematics and programming courses, equivalent to the University of Oslo mathematics courses MAT1100, MAT1110, MAT1120 and the corresponding computing and programming courses INF1000/INF1110 and INF1010/IN2900.

2. A total of at least 40 ECTC out of the 120 ECTC have to be advanced undergraduate courses at the 2000 and 3000 level.

3. An average mark C (European grading scale) is required for the above-specified 80 ECTS in Informatics/Mathematics/Statistics.

Study abroad and international collaborators

Students at the University of Oslo may choose to take parts of their degrees at a university abroad. The University of Oslo has exchange agreements with other universities in different parts of the world. The MN Faculty has additional exchange agreements with many universities abroad.

Career prospects

Candidates who are capable of modeling and understanding complicated systems in natural science, are in short supply in society. The computational methods and approaches to scientific problems students learn when working on their thesis projects are very similar to the methods they will use in later stages of their careers. To handle large numerical projects demands structured thinking and good analytical skills and a thorough understanding of the problems to be solved. This knowledge makes the students unique on the labor market.

Career opportunities are many, from research institutes, universities and university colleges and a multitude of companies. Examples include IBM, Hydro, Statoil, and Telenor, universities, hospitals, research institutes, biotech & pharmaceutical industry. The program gives an excellent background for further studies, with a PhD as one possible goal.

The program has also a strong international element which allows students to gain important experience from international collaborations in science, with the opportunity to spend parts of the time spent on thesis work at research institutions abroad.