Bioinformatics BB2441 (7.5 hp) – fall 2021 *course memo v2*

Welcome to the course!

Notes regarding the COVID-19 situation

The course is given in accordance with the recommendations from the Public Health Agency of Sweden and KTH. If you have concerns related to the course and your health situation, please contact the course responsible or your study counsellor. Due to the Covid-19 situation, the setup may change during the course.

Important dates

2021-08-30	Course starts 08:15 in E1, Osquars Backe 2 (KTH main campus)
2021-TBD	Last day to sign up for Exam. You will be notified automatically.
2021-10-22	Exam 08:00-12:00

Canvas

Canvas will be used for communication during the course, including updates to this course PM. Your access to the **BB2441** event at **https://kth.instructure.com** will be activated after course registration. Log in using your kth.se-account and look for the event **BB2441 HT21-1 Bioinformatics**. Lecture notes, lab instructions, and other handouts will be provided through Canvas.

Aim

The aim of the course is that the students should acquire sufficient knowledge about fundamental bioinformatics and programming resources and methods, so as to know how to (i) explain the theory behind the methods, (ii) use the corresponding tools and resources to perform analyses of molecular biology data, and (iii) interpret the results of these analyses.

Specifically, after passing the course, the student should be able to:

1. explain the significance of bioinformatics in modern life science
2. explain the applications and limitations of bioinformatical methods
3. explain fundamental bioinformatical methods
4. use relevant bioinformatical tools to solve bioinformatical problems
5. motivate the choice of bioinformatical tools
6. interpret the results of bioinformatical analyses
7. use script programming to perform fundamental bioinformatical operations
8. use fundamental commands in Unix/Linux or equivalent

Syllabus

The course is organized in 5 modules:

1. Programming [3 lectures]
2. Alignment [5 lectures]
3. Sequence features [5 lectures]
4. Phylogeny [3 lectures]
5. Computer exercises [7 computer exercises; 3 programming, 4 bioinformatics]

There is also an introductory lecture and a Q&A session. In total 18 lectures and 7 computer exercises.

The course credits are distributed into 2 parts:

- Examination (TEN1, grade scale A-F), 5 hp
- Computer exercises (LAB1, grade scale P/F), 2.5 hp

Course contents

Bioinformatics: Pairwise sequence alignment of protein and DNA/RNA sequences, multiple sequence alignment, significance of alignment results, features of protein and DNA/RNA sequences including sequence conservation, homology, phylogeny, gene expression and differential gene expression, clustering of vectors, introduction to relevant public databases and extraction of relevant data from these. How to use command line operations for data analysis and file management (in Unix/Linux, or corresponding operating systems), as well as available web-based tools for data analysis (e.g., Galaxy).
Programming: reading from file, basic operations on data such as conditional execution, loops, and regular expression, as well as writing to file.

The course content is the union of all material presented at the lectures, at the computer exercises, and in the course literature as specified by the detailed reading assignments, as well as any additional handouts or other resources specified during the course.

Teachers

Olav Vahtras (OV), Lukas Käll (LK), Anders Andersson (AA), Olof Emanuelsson (OE)

Teaching assistants (TAs) (labb-handledare):

Patrick Truong (PT), Adelina Rabenius (AR), Ludvig Bergenstråhle (LB)

Prerequisites (recommended)

BB1150 Biochemistry 1 (or equivalent), BB1190 Gene technology (or equivalent), SF1625 Calculus in one variable (or equivalent), SF1911 Statistics for bioengineering (or equivalent), and BB1000 Programming in Python (or equivalent).

Literature

Bioinformatics book:
Exploring bioinformatics. A project-based approach. *Caroline St. Clair* and *Jonathan E. Visick* (Jones & Bartlett Learning, 2nd edition) ISBN 978-1-284-03424-0.

Recommended (not mandatory) reading for the programming part:
Online resource: Jake VanderPlas: Whirlwind Tour of Python
http://nbviewer.jupyter.org/github/jakevdp/WhirlwindTourOfPython/blob/master/Index.ipynb
Book: Al Sweigart: Automate the Boring Stuff with Python Programming.

Articles: (will be provided to you; part of course requirements)
Eddy: Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology*, vol **22**, pp 1035-1036 (2004)
Wang: RNA-Seq: a revolutionary tool for transcriptomics. *Nature Review Genetics*, vol **10**, pp 57-63 (2009)
Sherlock: Analysis of large-scale gene expression data. *Current Opinion in Immunology*, vol **12**, pp 201-205 (2000)

Additional articles or other resources may be specified during the course.

Approximate reading assignments (chapters/articles) are provided in this course PM. Detailed reading assignments (actual pages/sections) will be provided during the course. Handouts, lab instructions, articles, and other resources specified during the course are also part of the required reading.

Requirements for examination

A. **Written exam (TEN1).** 5.0 hp. 4 hours written exam with theory questions and problem solving. Grades: A, B, C, D, E, Fx, F, where A-E correspond to pass, F is fail, and Fx is fail with the possibility to complement.

B. **Computer exercises (LAB1)**. 2.5 hp. Grades: Pass/Fail. Pass means that all computer exercises have been performed and that a report for each exercise has been handed in *in time* and has been approved.

To receive a final grade at the course you must pass both the written exam (with grades A-E) and receive a Pass grade at the computer exercises. Your final grade will be the same as the grade on your written exam. If you receive an Fx at the written exam you may request to be assigned a complementary assignment to complete and hand in within the stipulated time, and if deemed satisfactory your grade will be changed from Fx to E. If you receive a Fail at the computer exercises, you must consult with the teaching assistants to find out what you need to do to change that into a Pass.

Written assignments may give you bonus points on the exam.

For the bioinformatics modules (Alignment, Sequence features, Phylogeny) you are asked to write questions that you would like to know the answer to, related to the content of the module. This should be done in Canvas under Assignments.
There will in total be 9 occasions:

- Alignment module: **4 occasions**. Questions or answers should be entered **before** the 4 latter of the 5 lectures in the module (i.e., L7, L8, L9 and L10).
- Sequence features module: **3 occasions**. Questions or answers should be entered **after** 3 of the lectures (L5, L11, L13), and within the given time frame.
- Phylogeny module: **2 occasions**. Questions or answers should be entered **after** 2 of the lectures (exactly when will be communicated at the start of the module), and within the given time frame.

You may choose to write a question *or* answer a question that someone else already wrote. Either will count towards getting the bonus points.
These assignments are not mandatory. They will, however, give you bonus points on the final written exam:

- If you provide a question or comment at **8 or 9** occasions you will get **4 bonus points**
- If you provide a question or comment at **5-7** occasions you will get **2 bonus points**
- If you provide a question or comment at **4 or less** occasions you will **not get any** bonus points

This is provided that (i) the question or answer is relevant and of sufficient quality indicating you have given the concept some thought, and that (ii) the question or answer is provided within the stipulated time frame.
You are of course welcome to provide both a question and answer(s) to other questions on the same occasion.

The final written exam

The final written exam **2021-10-22** will contain theory questions and problem solving based on the contents of the bioinformatics part of the course. You are allowed to use a calculator including graph calculator where the memory is empty at the start of the exam. The exam consists of two parts, A and B. Part A consists of 12 multiple-choice questions (one alternative per question) worth 2 points per question (either 0 or 2). Part B consists of 4 essay questions worth 5 points per question. To pass the exam (grade E) you need at least 16 points on Part A (8 correctly answered questions). **If you have less than 16 points on Part A, the Part B questions will not be corrected**. The final grade (given that you pass Part A) will be based on all your points (Part A + Part B). Bonus points from the written assignments (module questions) will be counted into Part A (and into the total score if you pass Part A). Thus, if you received bonus points from the written assignments, you need at least 6 or 7 correctly answered questions on Part A. If you did not receive bonus points from the written assignments, you need at least 8 correctly answered questions on Part A.
NOTE 1: You must sign up for the final written exam in advance.
NOTE 2: Rules in order to be allowed into the room where the exam is given:

- A valid **photo identification** is required, no exceptions.
- Be present 10 minutes before actual exam starting time: Entry to the exam is organized in two intakes. Intake 1: **10 minutes *before* the start**. Intake 2: 30 minutes after start. No admission in between, or after intake 2.

- If you haven't registered for the exam you may still be allowed to take it provided: (i) there is an available seat; (ii) you present a valid course registration certificate upon admission; (iii) you adhere to all other rules including the two mentioned above.

Teaching activities

The **lectures** will consist of teacher-led lecturing of concepts as well as various student activities.

The **computer exercises** in the Programming part (LAB1-LAB3) are carried out individually. You are allowed to collaborate and get help and inspiration from others, but you must all submit your solutions individually. Each solution file must further be individually hand typed and not copy-pasted from others. The written reports should be handed in *within two weeks after the exercise* using the Canvas report submission system.

LAB1: Programming. Introduction to shell scripting in bash
LAB2: Programming. Basic Python programming
LAB3: Programming. Regular expressions

The **computer exercises** in the Bioinformatics part (LAB4-LAB7) are supposed to be carried out in teams of two students. You will be responsible for finding your own collaborator. The written reports should be handed in *within two weeks after the exercise* using the Canvas report submission system. Exercise instructions and data sets will be provided through Canvas.

The four bioinformatics exercises all revolve around a common scenario, which will be available in Canvas. You must read this scenario before the first Bioinformatics exercise.

LAB4: Bioinformatics. Gene finding, Blast and sequence alignment
LAB5: Bioinformatics. rRNA finding, taxonomic classification and multiple sequence alignment
LAB6: Bioinformatics. Protein sequence feature prediction, multiple sequence alignment and phylogenetics
LAB7: Bioinformatics. RNA sequencing, differential gene expression

Further practical information regarding the computer exercises can be found at Canvas. Specifically, you will be asked to use your own computer for the labs. Since it may take some time to setup your computer for the lab work, we require that you do this in advance of the first Bioinformatics lab. We have therefor designed a first assignment in Canvas called **“Preparing your computer environment for the first lab session”** that your group should do on your own, deadline Wed 26 Sep. If you or your lab group does not have access to a computer (Windows, Mac or Linux) to perform the computer labs, please inform the course responsible as soon as possible.

On site or online teaching?

Lectures will be on site, but in rooms large enough to enable physical distance between participants. This is provided the current Covid-19 restriction level remains in effect, but in case of a worsening pandemic situation it might be changed with short notice.
Lectures will not be recorded unless otherwise specifically indicated by the teacher.

Computer labs will be online. You will use your own computer - we will provide you with instructions for how to install the required (free) software. During the labs, you will be able to communicate with the lab TAs via Zoom. The Bioinformatics labs (LAB4-LAB7) should be carried out in groups of 2 students - you are responsible for organizing the communication (e.g., using Zoom or Skype) with your lab partner if you are working at separate places.

Contact and information

Initial information regarding the course is provided in this course PM. All subsequent communication, including updates to this course PM, will be communicated through the **BB2441 HT21-1 Bioinformatics** event in Canvas (**https://kth.instructure.com).**

Course coordinator: Olof Emanuelsson, olofem@kth.se. Please check Canvas first, maybe there is already an answer to your question there.

Lab TAs, for questions regarding the Bioinformatics labs:
Patrick Truong (patrick.truong@scilifelab.se, main responsible TA), Adelina Rabenius (adelina.rabenius@scilifelab.se), Ludvig Bergenstråhle (ludvig.bergenstrahle@scilifelab.se)

For questions regarding the Programming labs: Olav Vahtras (vahtras@kth.se)

For questions regarding registration etc., please contact Studentexpeditionen (floor 5, AlbaNova) at studentexpeditionen@biotech.kth.se. Please note that there is no drop-in, any appointments must be scheduled in advance.

Disability

If you have a disability, you may receive support from Funka: https://www.kth.se/en/student/studentliv/funktionsnedsattning. We recommend you inform the teachers regarding any need you may have. Funka does not automatically inform the teachers (not even the course coordinator).

Schedule

Programming Alignment Sequence features Phylogeny

Event	Date	Time	Room	Lecturer	Contents and *approximate* reading assignments
L1	Mon 30 aug	08:15 - 09:00	E1	LK/OE	Introduction
	Wed 1 Sep	TBD	TBD		(moved to 8 Sep)
L2	Thu 2 sep	08:15 - 10:00	FR4	OV	Introduction to shell scripting in bash
LAB1	Fri 3 sep	**08:00 – 12:00**	**501Spe, 502Spo**	**OV**	**(P) - Introduction to shell scripting in bash**
L3	Mon 6 sep	10:15 - 12:00	FR4	OV	Basic Python programming
LAB2	**Mon 6 sep**	**13:00 - 17:00**	**4V3Ora, 4V4Gul**	**OV**	**(P) Basic Python programming**
L5	Tue 7 sep	08:15 - 10:00	E1	OE	Genome features; Prokaryotic gene prediction [SV: 1:1-15+Biobackground, 2:figures p 27+31 +Biobackground, 5:101, 9:173-186+Biobackground]
L4	Wed 8 Sep	15:15-17:00	Digital	OV	Regular expressions
L6	Thu 9 sep	08:15 - 10:00	FR4	LK	DNA alignment (pairwise) [SV 3]
L7	Fri 10 sep	15:15 - 17:00	FR4	LK	Protein alignment incl. subst. matrices [SV 5, Eddy]
LAB3	**Mon 13 sep**	**15:00 - 19:00**	**501Spe, 502Spo**	**OV**	**(P) - Regular expressions**
L8	Tue 14 sep	08:00 - 10:00	M1	LK	Multiple sequence alignment, BLAST [SV 4]
-	Tue 14 sep	15:15 - 17:00	M2	- - -	[Backup, if used will this be announced on Canvas]
L9	Thu 16 sep	08:15 - 10:00	FR4	LK	Statistical significance of alignment matches I
L10	Mon 20 sep	13:15 - 15:00	FR4	LK	Statistical significance of alignment matches II
L11	Tue 21 sep	08:15 - 10:00	FR4	OE	Motifs, PSSMs, Sequence logos; Eukaryotic gene prediction [SV: 9:187-189, 10:193-213+Biobackground]

LAB4	**Tue 21 sep**	**13:00 - 17:00**	**5O1Spe, 5O2Spo**	**TAs**	**(B) - Gene finding, Blast and sequence alignment.**
L12	Wed 22 sep	10:15 - 12:00	E1	OE	Protein sequence features; Predictive discriminative models [SV: 5:91, 11:219-222+Biobackground]
L13	Thu 23 sep	08:15 - 10:00	FR4	OE	Gene expression; RNA-seq [SV: 8:152-156+Biobackground, Wang]
L14	Mon 27 sep	10:15 - 12:00	FR4	OE	Clustering [SV: 7:126-128, Sherlock]
LAB5	**Tue 28 sep**	**08:00 - 12:00**	**5O1Spe, 5O2Spo**	**TAs**	**(B) - rRNA finding, taxonomic classification and multiple sequence alignment**
L15	Tue 28 sep	15:15 - 17:00	M1	AA	Phylogeny - distances and trees [SV 6, 7]
L16	Thu 30 sep	08:15 - 10:00	FR4	AA	Phylogeny - distances and trees [SV 6, 7]
L17	Fri 1 oct	08:15 - 10:00	E1	AA	Phylogeny – distances and trees [SV 6, 7]
-	Mon 4 oct	10:15 - 12:00	M2	- - -	[Backup, if used will this be announced on Canvas]
LAB6	**Tue 5 oct**	**08:00 - 12:00**	**5O1Spe, 5O2Spo**	**TAs**	**(B) - Protein sequence feature prediction, multiple sequence alignment and phylogenetics**
-	Wed 6 oct	08:15 - 10:00	FR4	- - -	[Backup, if used will this be announced on Canvas]
LAB7	**Tue 12 oct**	**13:00 - 17:00**	**5O1Spe, 5O2Spo**	**TAs**	**(B) - RNA sequencing, differential gene expression**
L18	Fri 15 oct	15:15 - 17:00	FR4	LK/AA/OE	Questions and Answers
Exam	Fri 22 oct	08:00 - 12:00	FB51, FB54, FB55, FD41		Written exam

L = lecture
LAB = computer exercise
(P) - Programming computer exercise
(B) - Bioinformatics computer exercise
SV= Chapters in St. Clair & Visick
Eddy = Eddy article
Wang = Wang article
Sherlock = Sherlock article
TBD = to be decided

Reading instructions and lecture content are subject to change.
FB42, FB51, FB52, FB53, FB54, FB55, FD41, FR4 - AlbaNova UniversityCenter, Roslagstullsbacken 21
E1 - Osquars Backe 2
M1 - Brinellvägen 64
5O1Spe, 5O2Spo - Valhallavägen campus, Lindstedtsvägen 9
4V3Ora, 4V4Gul - Valhallavägen campus, Osquars backe 2