Implementing a web-based introductory bioinformatics course for non-bioinformaticians that incorporates practical exercises

Antony T. Vincent[1], Yves Bourbonnais[1], Jean-Simon Brouard[1], Hélène Deveau[1], Arnaud Droit[2], Stéphane M. Gagné[1], Michel Guertin[1], Claude Lemieux[1], Louis Rathier[3], Steve J. Charette[1] and Patrick Lagüe[1*]

[1]Département de Biochimie, de Microbiologie et de Bio-informatique, Faculté des sciences et de génie, Université Laval, Québec (Québec), Canada

[2]Centre Hospitalier de l'Université Laval, Faculté de Médecine, Université Laval, Québec (Québec), Canada

[3]Équipe de soutien informatique, Faculté des sciences et de génie, Université Laval, Québec (Québec), Canada

Correspondence:

Corresponding Author

Patrick.Lague@bcm.ulaval.ca

Running title: A web-based bioinformatics course for life scientists

Keywords: Online course, Undergraduate, Life sciences, Bioinformatics.

Accepted for publication *in Biochemistry and Molecular Biology Education*

DOI: 10.1002/bmb.21086

Pubmed ID: 28902453

Abstract

A recent scientific discipline, bioinformatics, defined as using informatics for the study of biological problems, is now a requirement for the study of biological sciences. Bioinformatics has become such a powerful and popular discipline that several academic institutions have created programs in this field, allowing students to become specialized. However, biology students who are not involved in a bioinformatics program also need a solid toolbox of bioinformatics software and skills. Therefore, we have developed a completely online bioinformatics course for non-bioinformaticians, entitled "*BIF-1901 Introduction à la bio-informatique et à ses outils (Introduction to bioinformatics and bioinformatics tools*)," given by the Department of Biochemistry, Microbiology and Bioinformatics of Université Laval (Quebec City, Canada). This course requires neither a bioinformatics background nor specific skills in informatics. The underlying main goal was to produce a completely online up-to-date bioinformatics course, including practical exercises, with an intuitive pedagogical framework. The course, BIF-1901, was conceived to cover the three fundamental aspects of bioinformatics: (1) informatics, (2) biological sequence analysis and (3) structural bioinformatics. This paper discusses the content of the modules, the evaluations, the pedagogical framework, and the challenges inherent to a multidisciplinary, fully online course.

Introduction

Although the use of computer software in biology goes back to Margaret Dayhoff with the 1962 publication of the computer program "Comprotein" to aid primary protein

structure determination [1], the term bioinformatics was introduced in the scientific literature only at the beginning of the 1970s to define the study of informatics processes in biotic systems [2]. Since then, this contemporary discipline, which combines concepts of biological and computer sciences, is closely involved in several fields of modern sciences such as personal medicine [3], vaccine design [4], drug discovery [5] and the study of infectious diseases [6].

As the field of bioinformatics is still evolving [7], which skills and knowledge should be mandatory to become a specialist are still matter of debate [8,9]. For example, the recent advances in wet lab automated technologies, such as next generation sequencing (NGS) and mass spectroscopy techniques, has overwhelmingly increased the capacity to generate a huge amount of data that can only be treated using computers. As a consequence, life scientists need basic bioinformatics skills [10] to understand the underlying subtleties of using various tools and to avoid an involuntary misapplication of the methods or an erroneous interpretation of the results [11]. Further, the increasing number of applications in bioinformatics covers a wide variety of fields, from genomic data analysis to modeling of proteins and organisms, and this could help users to make the maximum possible use of the experimental data. However, bioinformatics is complex and multidisciplinary, requiring expertise from biology, physical chemistry, maths, computer science and statistics [12]. Thus, several authors now acknowledge that teaching fundamental concepts and developing basic bioinformatics skills is essential curriculum for life science students [8,13–16].

Therefore, we have developed and here present an online course in bioinformatics for non-bioinformaticians, entitled "BIF-1901 *Introduction à la bio-informatique et à ses outils (Introduction to bioinformatics and bioinformatics tools)*", given by the Department of Biochemistry, Microbiology and Bioinformatics of Université Laval (Quebec City, Canada). This online course favors the development of self-directed learning competencies, such as autonomy, self-discipline, organisation and effective communication, in a supported, supervised pedagogical framework. The learning environment, the content and the course assessment are presented in the following sections.

Learning environment

The course is scheduled in the last year of undergraduate studies and is mandatory for students enrolled in biochemistry and microbiology programs and optional for those in biology program. The course aims to teach fundamental bioinformatics concepts and skills, as well as to introduce students to the different fields that use bioinformatics.

The course emphasis is on practical training for the main bioinformatics tools and resources related to the databanks of protein sequences and structures, genome sequencing and sequence assembly techniques, sequence alignment and database search, phylogenetic analysis and phylogenetic tree building, homology modeling from protein sequences, and molecular modeling of protein-ligand interactions from molecular docking. The book "*Introduction to Bioinformatics*" from J. Xiong [7] was chosen as the course book because: 1) it is easily accessible for newcomers to the discipline, 2) most

(but not all) of the modules are covered by this book, 3) the tone is uniform and consistent throughout, and 4) it is possible to subscribe to an online version of the book, and use only specific chapters for the course. The university library can then provide the subscription to students for free.

The course is divided into three parts that cover different topics of bioinformatics (Fig. 1: informatics, biological sequences and structural bioinformatics) which are further subdivided into 8 modules, each taught by different professor specialized in the subject. For each module, the theoretical concepts are presented in narrated PowerPoint/Articulate (https://articulate.com) lessons. Readings from Xiong's book are assigned and practical aspects of using the tools and resources are presented in video clips.

To allow viewing on computers, tablets and cell phones with modern operating systems, Flash and HTML5 formats are used for the multimedia files. Videos are limited to a maximum of 15–20 minutes to lighten the download of each file, optimize the focus of students on the new learning, and to facilitate the tracking of the material introduced in each capsule and/or clip, so students can easily keep track of concepts that they have studied.

The online pedagogical content is hosted on the university's integrated numerical teaching platform, called *monPortail* (monportail.ulaval.ca; for a description of the numerical environment, see www.ene.ulaval.ca/monportail-decouvrir-monportail). The platform includes a moderated forum open to teachers and students to ask questions and

post answers, and to report problems and solutions related to the course content. Students are encouraged to use the forum instead of contacting the professors directly, allowing the latter to disseminate the answers to all students enrolled in the course. Students are invited to answer their colleagues as appropriate, obviously with instructor supervision. To develop a better human connection, students are asked to upload a real photo of themselves to their account, and this picture is displayed along with their postings on the forum.

Forum users are to follow these rules: (1) read the course FAQs (a document listing all the questions/answers from the previous semesters) before asking a question, (2) include only one topic for each forum thread, (3) always be polite, (4) be as concise and clear as possible (for example by providing a transcription or screenshot of error messages for software problems) and (5) repost on the forum if a student found an answer by him- or herself.

The practical exercises that require a significant amount of computer resources or the use of specialized software are executed on a GNU/Linux server accessible to students from the internet via the user-friendly NoMachine client (https://www.nomachine.com/), which is freely available on all major operating systems (Windows, Mac OS X and GNU/Linux). This server allows students to access to an informatics environment identical to the platforms used by professional bioinformaticians without the problems related to the installation of a GNU/Linux distribution on student computers (such an installation would not be appropriate for an introductory course in bioinformatics for

students at this level). The GNU/Linux server is a virtual machine hosted on the faculty's server, and is composed of 8 CPU, 32 GB of RAM and 1 TB of disk space, with Xubuntu (https://xubuntu.org) as the operating system. These resources are sufficient for an online class of about 50 students that do not connect to the server at the same time and, as a virtual machine, the resources can be adjusted "on the fly" without rebooting. Only free professional software was chosen for the course in order to avoid any license fees and to increase the chances that students continue to use and practice their knowledge and skills after the course [17] (a software list is provided in Table S1 (supplementary material)). Students then have an opportunity to do a for-credit research project in the semester following the course, or to engage in graduate studies.

Knowledge and skills are evaluated through mandatory online quizzes (at specific windows of time during the semester) and assignments, which are similar to the practical exercises taught in the modules (Table 1). The online quizzes, as with all pedagogical activities in this course, are completed through the online course portal hosted on *monPortail*.ulaval.ca. A large pool of questions has been produced for each quiz and they are randomly selected by the online teaching platform. Consequently, every student has a distinct quiz. The teaching platform also allows students to submit their completed assignments using electronic files (PDF files for reports, and other file formats related to the assignment topic, such as FASTA for protein sequences and PDB for protein structures). The deposited files can be tracked for delays. Teachers return the corrected PDF reports to students via *monPortail* and can also post comments on the students' section. In addition to the standard statistical measures on evaluations such as the mean

and the standard deviation, *monPortail* allows teachers to access more personalized statistics regarding the success of the course, including easily spotting which students are struggling.

Course Learning Outcomes (CLOs)

Upon completing the course, students will be able to:

1. Describe the major fields of applications and challenges in bioinformatics;
2. Effectively retrieve information from biological databases and understand their importance in bioinformatics;
3. Understand the main theoretical and practical aspects of:
   3.1 aligning and assembling DNA and protein sequences;
   3.2 similarity searches in databases;
   3.3 phylogenetics;
   3.4 predictions of the 3D structure of protein;
   3.5 molecular modeling;
4. Use various specialized resources to characterize patterns and domains of a protein, its location in the cell, and the expectable post-translational modifications;
5. Compare the experimental methods used to determine the 3D structure of proteins at the atomic level and tools for predicting the structure of proteins;
6. Understand the importance of modelling and molecular docking in the study of biological molecules.

See also Table 1 for the mapping of the specific modules to the CLOs.

.

Course Content

The following paragraphs outline the rationale and content of each module and describe the assessments. A list of freely available online bioinformatics tools used in the course is presented in Table S1 as well as a brief description of the assignments are provided as supplementary material. A detailed course syllabus is provided as supplementary material. Most of these modules are based on chapters of Xiong's textbook, and for interested readers several exercises are included in the textbook's Appendix. As the course is modular, it is possible to incorporate only the relevant modules in a particular formation, depending on the teaching expertise available. As this is an introductory course in bioinformatics, most of the modules can be taught by a single expert in the domain, if necessary. In our case, the B.Sc. in bioinformatics from our department has allowed several experts to participate in the course.

Introduction to bioinformatics, software and GNU/Linux. This module provides an introduction to the course and the pedagogical framework, followed by a definition and some applications of bioinformatics, and its importance in the context of a career in life sciences. During this module, students install and become familiar with the software they use to access the GNU/Linux server (NoMachine, https://www.nomachine.com/) and to transfer files from the server to their own computer (FileZilla, https://filezilla-project.org/). A clip is used to show the software installation steps for most current operating systems (Windows and Mac OS X). Finally, students are introduced to the Unix-like operating system environment on the server they use for the practical exercises

and homework, and the basic BASH (Bourne-Again shell, a command processor that runs in a text window) commands are presented (ls, mkdir, rmdir, cd, cp, rm and mv).

For this module, students are required to complete homework that evaluates skills in basic BASH commands. We also ensure that their NoMachine client is properly installed and functional on their computers, so they are ready for the subsequent sections. Students are given a list of command lines to execute via NoMachine on the server. To assess their work, they have to take screenshots of each step using their specific username on the GNU/Linux server. The screenshots are submitted on *monPortail* and used for correction.

Biological databanks. A huge amount of data is generated through research activities, presenting challenges for modern biologists [18]. Biological databases are used to organize and share research data gathered by the scientific community. In this module, students are introduced to the main biological databases in the genomics and post-genomics era and how to use them. The goal of this module is to show how to efficiently query the main biological databases (GenBank, EMBL, RefSeq, UniProt, PDB, etc.), particularly through the use of advanced search queries that allow filtering directly on specific fields. Students also learn how to differentiate primary, secondary and specialized databases. Since data quality can vary greatly depending on the source, it is important to develop a critical sense in regard to the various available online resources. The source of the data determines whether it is primary data, or data that was processed with bioinformatics tools or validated by experts. Finally, students explore the most common file formats used to represent genomic data and metadata. An online quiz allows

students to review the course material, but since the content is used and evaluated later in the course, the quiz does not contribute to the final score.

Sequencing and sequence assembly techniques. In recent years, the development of high-throughput sequencing technologies has revolutionized the field of genomic research. Several features of these technologies pose challenges to the genome assembly algorithms. This module of the course explores some applications of the NGS techniques and the challenges they pose in terms of data storage and analysis, then presents two sequencing techniques (454 and Illumina) that have been widely used. This module also introduces the process of reads assembly, defines important terms (reads, contigs, etc.) and makes clear the distinction between *de novo* assembly and assembly with a reference genome. Finally, the module ends with an introduction to the assembly algorithms with a focus on those designed for NGS data (use of graphs). At the end of the module, students should be able to (1) compare 454's and Illumina's technologies (2) describe the main steps and challenges associated with the assembly process (3) calculate the N50 of an assembly and (4) understand how graphs are used by assembly algorithms to represent the relationship between reads. We also introduce Tablet [19], a user-friendly software for visualizing sequence alignments.

Sequence alignment and database search. When a sequence, whether in nucleotides or amino acids, is obtained, one of the first steps is to compare it with sequences already available in public databases. A database search uses pairwise alignment algorithms to find homologous sequences. It is also possible to compare different homologous

sequences by performing a multiple sequence alignment (MSA). As discussed here and applied in other sections, this type of alignment is used when predicting the structure and function of a protein and is also the basis of phylogenetic studies, where evolutionary links are inferred from sequence alignments.

In this module, we begin by presenting the alignment of two sequences with each other. We also distinguish between homology, similarity and identity, three terms widely used in these comparisons. The different methods used for pairwise alignments are then described. This module also includes an introduction to the various substitution matrices used in alignments to assign the score employed to quantify the degree of likelihood between these sequences. Finally, we see how common MSA method works, like the progressive alignment of sequences. This module concludes by presenting software commonly used when aligning and/or searching against databases such as BLAST [20], MUSCLE [21], ClustalW [22], Jalview [23].

At the end of this section, students have to complete their second mandatory course homework assignment. They should choose a curated data set of sequences from the NCBI Protein Clusters database (https://www.ncbi.nlm.nih.gov/proteinclusters). Students can choose any data set that contains between 15 and 30 sequences of lengths of 300 to 600 amino acids. However, they should choose carefully since the same sequences are also used for the homework on molecular phylogeny. The sequences must be aligned using MUSCLE and a figure of the resulting alignment produced with Jalview in an EPS format, with some graphical modifications such coloring the residues with a percentage

of identity greater than 80%. Since this section explains the cornerstone of bioinformatics, students must complete an online quiz.

Model organism databases and prediction of protein motifs and functions. For a long time, model organisms have been used to study fundamental mechanisms preserved during evolution. Using these organisms has several advantages such as low cost, rapid growth, the availability of genetic tools, and the possibility of doing high-throughput experiments. The numerous studies that have been done about these organisms, however, cause an issue of compiling the colossal amount of results available. In this section of the course, students learn about the online resources regarding model organisms. To do this, we explore two typical organism-oriented databases: the Saccharomyces Genome Database (SGD) [24] and dictyBase [25]. The latter is devoted to the amoeba *Dictyostelium discoideum*.

By the end of this section, students are able to (1) navigate within various model organism databases, (2) use gene ontology annotation for the characterization of a gene and its translated product, (3) predict the motifs, domains and post-translational modifications of a protein, and (4) predict a protein's cellular localization. In addition to the BLAST suite available on the database websites, the tools introduced here are SignalP [26], Phobius [27], TMHMM [28] and NetPhos [29]. To help students to become familiar with the databases and the tools, they have to complete an exercise where only a part of the information about a gene is given. Students have to fill the questions by exploring the databases and by using the presented tools.

Phylogenetic analysis and phylogenetic tree building. The ultimate goal of modern biology is to describe the phylogenetic relationships between all living organisms. Although we find both the ideas of phylogeny and phylogenetic tree in the prominent "On the Origin of Species" of Charles Darwin (1859), the term "phylogeny" was used for the first time by Ernst Haeckel in 1866 to define the branching order of animal and plant species. A phylogeny is usually represented in the form of a tree that is read from bottom to top. In recent years, the sequences of genes and proteins have progressively replaced morphological data to infer phylogenetic trees, and numerous algorithms and bioinformatics programs have been developed to this end. This section of the course deals with the theoretical and practical aspects of phylogenetic analyses. We describe the current methods employed for phylogenetic tree reconstruction and emphasize the importance of using an appropriate model of sequence evolution.

In this section, students are expected to use the correct terminology to define phylogenetic trees and to describe the steps of a phylogenetic analysis. They learn how to select the appropriate tree reconstruction methods for phylogenetic inferences and suitable evolutionary models of substitutions that best-fit nucleotide and amino acid data sets. In homework, students have the opportunity to construct a phylogenetic tree by the maximum likelihood method using their own protein data set that they selected for the *Sequence alignment* module homework. They also evaluate the robustness of the inferred tree by analysis of bootstrap replicates.

3D protein structure and homology modeling. The function of a protein is defined by its three-dimensional structure, and the structure of a protein is "encoded" in its primary sequence. Therefore, structural bioinformatics uses the structural properties of a protein as a link between the available protein primary sequence and its functional characteristics. Since protein structures are significantly more conserved than primary sequences, many protein studies include structural analyses. This includes using, analyzing and comparing experimental protein structures, predicting secondary and tertiary protein structures, predicting trans-membrane properties, predicting disorder and predicting function based on experimental or theoretical 3D structures.

We begin the module with a brief introduction of the role of structural bioinformatics and a review of protein's structural elements. Students learn to use and search the protein databank (PDB, www.rcsb.org). A portion of this module is dedicated to the use of the protein structural visualization software PyMOL [30]; this includes structure manipulation, atom and residue selections, using various types of representations, generation of high-quality protein structures figures, and superimposition of protein structure. We begin the second part of the module with a brief introduction to nuclear magnetic resonance spectroscopy and crystallography, the two experimental methods that feed the protein structure databases. This is followed by learning protein structure prediction methods; various secondary and tertiary protein structure predictions are covered (including homology modeling and *ab initio* methods), and discussing Critical Assessment Methods of protein Structure Prediction (CASP). The homework includes preparing three protein structure figures with specific requirements, carrying and

presenting secondary structure predictions from online servers, generation 3D models using MODELLER [31] on the bioinformatics server, and comparing template and predicted structures using PyMOL [30].

Introduction to molecular modeling and docking. Molecular modeling is now commonly used in several scientific disciplines, ranging from chemistry and biology to materials science. Molecular modeling techniques include several computer-based approaches used to model or mimic the behaviors of molecules. These techniques are used to investigate the structure, dynamics and thermodynamics of biological molecules.

The subjects of study include: protein folding and stability, protein design, protein-protein interactions, enzyme catalysis, conformational changes related to biological functions, protein dynamics and allosteric regulation, molecular recognition in the development of new drugs (drug design), structure and dynamics of biological membranes, and protein-membrane interactions.

In this module, students are introduced to the basics of molecular modeling, including the concepts of molecular force fields and the basis of popular modeling techniques such as molecular dynamics, Monte Carlo simulations, and docking simulations. Further, students become familiar with the different steps to perform a molecular docking simulation. The molecular docking uses Autodock Vina [32] and its PyMOL plugin [30]. The different steps, presented in a series of clips, can be done by students on the bioinformatics server. Finally, each team of 2 or 3 students performs a docking

simulation on a different protein-ligand complex from the PDB database, and discusses the results in a report.

Assessments

According to a course evaluation filled out by students at the end of the Fall 2015 semester, the general assessment of the course (including the content, the administration of the course, and the teachers) was low. It ranked at a satisfaction level of 73.7%, while department courses have an average of 88.1% in student rankings.

The major student concerns were the amount of work required to accomplish the homework, and the level of difficulty. These concerns were directly related to the low levels of student computer skills at the beginning of the course, as reported by a few students. It is worth noting that this is the only course of the curriculum that requires a minimum of computer skills for success, and for some students, the development of computer competencies was beyond their capacity.

Students overall felt that the exams added a substantial amount of work that did not contribute to fulfilling the course objectives. The course content was adjusted for the second year (Fall 2016) to include BASH practical exercises, which facilitated homework completion. Along with the withdrawal of the two in-class exams to the profit of additional practical homework, this measure contributed to significantly increase student satisfaction with the course (general assessment of 86.4%, compared to an average of

86.9% for the courses of the department). Further, we added a few hours of in-class support a few days before each assessment deadline, to help with interactive computer debugging that generated student frustration and anxiety. Although many students did not attend the in-class support, this resource was appreciated by the most "computer-disoriented" students.

The list of the student's assessments is included in Table 1 along with their respective percentage of the final grade and the average marks obtained by the students for the semesters of Fall 2015 and Fall 2016. The marks indicate that the students have achieved the CLOs at a satisfactory level. Noticeably, the workload required to complete the last two homework was increased as of Fall 2016 following the withdrawal of the two exams, which is reflected in the lower average marks obtained by the students (96.3% vs 83.3%, and 93.6% vs 88.9%, respectively). However, the final grades were higher for Fall 2016 as of the average scores of the Fall 2015 exams were rather low.

To a lesser extent, students were irritated by the lack of uniformity in the presentation of course material between the modules. As indicated in the introduction section, bioinformatics is a multidisciplinary scientific discipline; it was consequently obvious to have a specialist in each domain to teach the different modules. However, this causes an inherent discrepancy between teaching styles that seems to have been difficult for students.

Conclusion

With the development of high-throughput technologies, bioinformatics has become a fundamental cornerstone for life sciences. Bioinformatics involves multidisciplinary expertise from biology, physical chemistry, mathematics, computer science, and statistics. Hence, the new introductory course included all of those, but was adapted to the curriculum of life science students. As distance learning is becoming more prevalent, we developed an online bioinformatics introductory course which encompasses both theoretical aspects and practical applications.

The course is divided into three parts that each covers a different aspect of bioinformatics (Fig. 1: informatics, biological sequences, and structural bioinformatics). The teaching is supervised by several professors, each specialized in one of the topics covered by the course.

One of the main challenges for this kind of course is the relatively low interest and competency in informatics of some life science students. Substantially expanding the portion of practical training contributed to better student satisfaction with the course.

Conflict of Interest

The authors declare that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgments

The authors want to thank the students who gave useful comments to improve the course and also the Faculty of Science and Engineering at Université Laval for the computing resources.

References

[1] Dayhoff, M.O. and Ledley, R.S. (1962) Comprotein: A Computer Program to Aid Primary Protein Structure Determination. Proc December 4-6, 1962, Fall Jt Comput Conf. ACM, New York, NY, USA. p. 262–74.

[2] Hogeweg, P. (2011) The roots of bioinformatics in theoretical biology. PLoS Comput Biol. 7, e1002021.

[3] Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J. and Altman, R.B. (2011) Bioinformatics challenges for personalized medicine. Bioinformatics. p. 1741–1748.

[4] Luciani, F., Bull, R.A. and Lloyd, A.R. (2012) Next generation deep sequencing and vaccine design: Today and tomorrow. Trends Biotechnol. p. 443–452.

[5] Katara, P. (2013) Role of bioinformatics and pharmacogenomics in drug discovery and development process. Netw Model Anal Heal Informatics Bioinforma. 2, 225–230.

[6] Wattam, A.R., Davis, J.J., Assaf, R., Boisvert, S., Brettin, T., Bun, C. et al. (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 45, D535–542.

[7] Xiong, J. (2006) Introduction. Essent Bioinforma. Cambridge University Press, Cambridge. p. 3–9.

[8] Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B. et al. (2014) Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. PLoS Comput Biol. 10. e1003496

[9] Vincent, A.T. and Charette, S.J. (2015) Who qualifies to be a bioinformatician? Front Genet. 6, 164.

[10] Kwok, R. (2013) Out of the hood. Nature. 504, 319–321.

[11] Pevzner, P. and Shamir, R. (2009) Computing Has Changed Biology—Biology Education Must Catch Up. Science. 325, 541 LP – 542.

[12] Luscombe, N.M., Greenbaum, D. and Gerstein, M. (2001) What is bioinformatics? A proposed definition and overview of the field. Methods Inf Med. 40, 346–358.

[13] Maloney, M., Parker, J., Leblanc, M., Woodard, C.T., Glackin, M. and Hanrahan, M. (2010) Bioinformatics and the undergraduate curriculum. CBE Life Sci Educ. 9, 172–174.

[14] Cohen, J. (2003) Guidelines for Establishing Undergraduate Bioinformatics Courses. J Sci Educ Technol. 12, 449–456.

[15] Ditty, J.L., Kvaal, C.A., Goodner, B., Freyermuth, S.K., Bailey, C., Britton, R.A. et al. (2010) Incorporating genomics and bioinformatics across the life sciences curriculum. PLoS Biol. 8, 17–18.

[16] Pham, D.Q.D., Higgs, D.C., Statham, A. and Schleiter, M.K. (2008) Implementation and assessment of a molecular biology and bioinformatics undergraduate degree program. Biochem Mol Biol Educ. 36, 106–115.

[17] Vincent, A.T. and Charette, S.J. (2014) Freedom in bioinformatics. Front Genet. 5, 259.

[18] Mardis, E.R. (2016) The challenges of big data. Dis Model Mech. 9, 483 LP – 485.

[19] Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L. et al. (2013) Using tablet for visual exploration of second-generation sequencing data. Brief Bioinform. 14, 193–202.

[20] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol. 215, 403–410.

[21] Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

[22] Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., Mcgettigan, P.A., McWilliam, H. et al. (2007) Clustal W and Clustal X version 2.0. Bioinfomatics. 23, 2947–2948.

[23] Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. and Barton, G.J. (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics. 25, 1189–1191.

[24] Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T. et al. (2012) Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res. 40. D700-705

[25] Gaudet, P., Fey, P., Basu, S., Bushmanova, Y.A., Dodson, R., Sheppard, K.A. et al. (2011) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. Nucleic Acids Res. 39, D620–624.

[26] Petersen, T.N., Brunak, S., von Heijne, G. and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8, 785–786.

[27] Käll, L., Krogh, A. and Sonnhammer, E.L.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. 35. W429-432

[28] Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 305, 567–580.

[29] Blom, N., Gammeltoft, S. and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 294, 1351–1362.

[30] Schrödinger, L. (2010) The PyMOL Molecular Graphics System, Version 1.3r1.

[31] Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 234, 779–815.

[32] Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 31, 455–461.

Table 1. Overview of the course

Module	Duration (weeks)	CLO[a]	Student assessments		
			Assessment type (% of grade[b])	Average marks	
				2015	2016
Part 1: Informatics					
GNU/Linux and software	1	1	Homework (15%)	N/A[c]	94.8%
Part 2: Biological sequences					
Biological databanks	1	2	Formative quiz (0%)	N/A[d]	N/A
Sequencing and sequence assembly techniques	1	3.1	Formative quiz (0%)	N/A	N/A
Sequence alignment and database search	2	3.2	Homework (10%) and a quiz (10%)	94.5% and 81.6%	93.7% and 81.8%
Model organism databases and prediction of protein motifs and functions	1	3.2, 4	Quiz (15%)	94.6%	94.9%
Phylogenetic analysis, and phylogenetic tree building	2	3.3	Homework (15%)	96.3%	93.3%
Part 3: Structural bioinformatics					
3D protein structure and homology modeling	2	3.4, 5	Homework (15%)	98.2%	83.3%
Introduction to molecular modeling and docking	2	3.5, 6	Homework (20%)	93.6%	88.9%

a) The numbers refer to the CLOs presented in the section *Course Learning Outcomes (CLOs).*
b) The percentage of grade is as of Fall 2016, following the withdrawal of the two in-class exams. The average scores of the mid-term and final exams of Fall 2015 are 69.2% and 84.0%, respectively.
c) This homework was added in Fall 2016.
d) N/A: Not-applicable

FIGURES

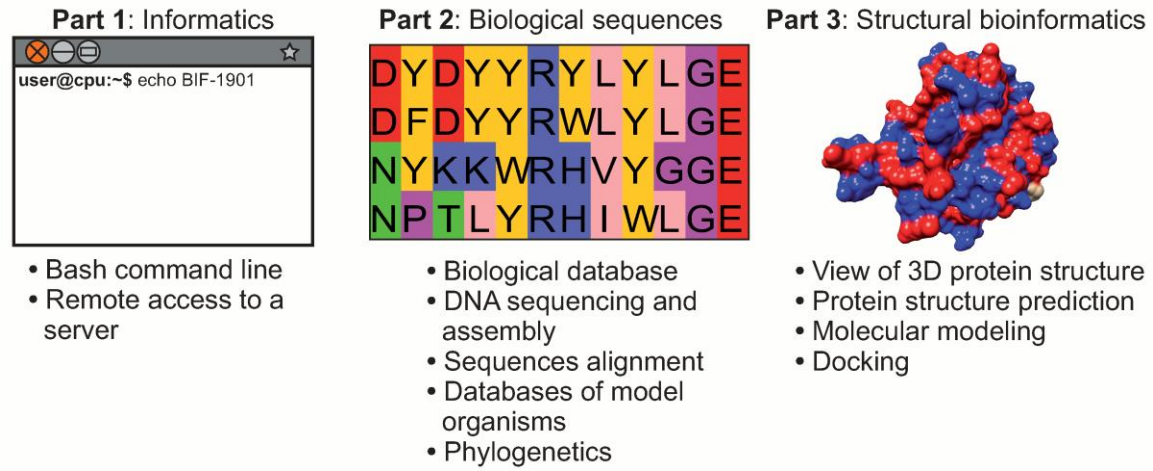

Figure 1. Overview of the BIF-1901 course.