BIOL 313 – Introduction to Bioinformatics and Computational Biology

Background
This is an introductory course to the field of Bioinformatics and Computational Biology. This is an emerging interdisciplinary field that deals with the development and application of computing algorithms and methods in biology. Many applications of computing in biology focus on processing the vast amounts of data generated by high-throughput experimental techniques, like DNA sequencing. Nonetheless, the range of applications extends also to simulations of biological processes, drug design or artificial life systems. Similarly, Bioinformatics relies mostly on algorithm design and programming, but other subfields of computer sciences are also strongly involved. These include parallel processing systems, database management, machine learning systems and hardware design. In addition, Bioinformatics draws substantially from numerical methods and statistical theory and is thus a true, hydra-headed hybrid of biology, computer sciences and mathematics.

Target audience
The course is intended for a multi-disciplinary audience. BIOL, CMSC, BINF, MATH, STAT and IS majors are especially invited, but the course is open to any undergraduates with basic training in the Sciences (i.e. fundamental math skills). The course will provide the necessary introduction to the underlying biology and computer science methods covered in class, even though enrolled students should have basic training (BIOL 141, CMSC 104) in at least one of the two underlying fields.

Course description
This is a question-driven course. Rather than explaining a series of methods and their applications in the biological sciences, the course will analyze some of the questions posed by current biology and will then analyze some of the proposed solutions in Bioinformatics. The course will start with a brief introduction to algorithms, molecular biology and evolutionary theory in order to provide a common ground and lexicon for the remainder of the course. It will then explore some of the problems faced by biology in the last decades, such as DNA sequencing, genomics, sequence comparison or protein folding, and some of the methods and answers provided by the emerging field of Bioinformatics.

Basic implementation
Each week will typically consist of two lecture classes. In some weeks, one of the lecture classes might be a group discussion of a scientific paper*. The course will include some home-based computer labs with mandatory hand-in assignments and more conventional homework assignments dealing with the practical implementation of the theoretical methods seen in lectures. Computer labs require only user-level computing skills.

Course outline
The course will cover 4 main topics. Topic I will introduce students to the main contemporary concepts behind algorithms, molecular biology and evolution, in order to provide a common ground and lexicon that should allow students from different backgrounds to follow the remainder of the course. Students will be introduced to the concept of algorithm and the basic mainstream algorithmic techniques for optimization and problem solving. They will also be learn about the molecular building blocks of life (DNA, RNA and proteins) and their interplay through the Genetic Code in the Central Dogma of molecular biology. Students will also learn the basics of evolution by natural selection, the interlinked concepts of genotype and phenotype and the dynamic concept of an adaptive landscape. Through the concept of orthology, dot-plots and other classical methods, students will be made aware of the relevance of evolution to Bioinformatics. Topic II will address the advent and coming of age of Bioinformatics and Computational Biology as independent research fields, as a consequence of the increasing amounts of data generated by modern high-throughput biology. It will analyze the challenges posed by high-throughput data, the reinforcement cycle of bioinformatics and biology and the applications of and career opportunities in Bioinformatics. Topic III will cover the foundational basis of Bioinformatics: sequence and genome analysis. It will slowly introduce students to the concept of sequence comparison and alignment, analyzing in depth the methods of dynamic programming and its computational complexity. Genome assembly will be analyzed as a foundational topic of modern bioinformatics before proceeding to explore the problem of genome annotation, subdivided into gene finding and functional annotation by sequence comparison. The students will learn how gene-finding algorithms exploit genome-wide statistics and signal search to locate genes, as well as the basics behind Bioinformatics most famous algorithm (BLAST), the difference between similarity and homology and the

---

* The implementation of paper discussion sessions will depend on the number of students enrolled in the course.
relevance of statistics in assessing the reliability of results. From there they will move on to the alignment of multiple sequences and analyze how computational progressive methods cope with exponential complexity. Topic IV will cover several common advanced topics in Bioinformatics, but focusing more on the nature of the problem and the solutions proposed, rather than immersing students in the algorithmic detail of the methods studied. In this section, students will be exposed to the field of molecular phylogeny, how it can address biologically and socially important questions and how the basic methods approach the task of tree reconstruction. Finally, students will gain insight into the elucidation of protein and domain interactions, how these can be predicted and used to construct organism-wide maps of genetic interaction, and how all this information can be linked to relevant phenotypes, like inheritable diseases.

Course topics

Topic I – Introduction to molecular biology
- Life and information
  - Definitions of life. Information, autopoiesis and evolution. Life requirements.
- Algorithms
- Introduction to molecular biology
- Evolution
  - Basic aspects and requirements. Genotype and phenotype. The adaptive landscape. Relevance to bioinformatics.

Topic II – Introduction to bioinformatics
- Origins and etymology. Raisons d’être.
- High-throughput methods
- Biological databases.
- Careers and applications.

Topic III – Sequence and genome analysis
- A jigsaw puzzle: Genome sequencing and genome assembly
  - Conventional sequencing. Shotgun sequencing. Genome assembly.
- Memory: Sequence analysis
  - Sequence comparison and the alignment problem.
    - Local and global alignment. Dynamic programming. Pair-wise alignment.
- Where is Waldo? Genome analysis:
- Guess who? Genome annotation
  - Databases. Search by content. Homology and similarity. BLAST. E-values.
- Jenga: Multiple sequence alignment
  - Alignment and complexity. Exact and progressive methods. Tradeoffs, greediness and errors.
- Join-the-dots: Molecular phylogeny

Topic IV – Advanced topics in bioinformatics
- Lego: Protein domains, disease and the interactome
Course schedule*

The following is a tentative schedule for the course:

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Thursday, January 27, 2011</td>
<td>Introduction</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 2</td>
<td>Tuesday, February 01, 2011</td>
<td>Life &amp; information systems</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 3</td>
<td>Tuesday, February 08, 2011</td>
<td>Molecular biology</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 4</td>
<td>Tuesday, February 15, 2011</td>
<td>Evolution</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 5</td>
<td>Tuesday, February 22, 2011</td>
<td>Biological Databases</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 6</td>
<td>Thursday, February 24, 2011</td>
<td>Midterm (1)</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 7</td>
<td>Tuesday, March 01, 2011</td>
<td>Genome assembly</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 8</td>
<td>Thursday, March 03, 2011</td>
<td>Sequence alignment</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 9</td>
<td>Tuesday, March 08, 2011</td>
<td>Dynamic programming</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 10</td>
<td>Thursday, March 10, 2011</td>
<td>Genome analysis</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 11</td>
<td>Tuesday, March 15, 2011</td>
<td>Genome analysis</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 12</td>
<td>Tuesday, March 22, 2011</td>
<td>Spring Break</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 13</td>
<td>Thursday, March 29, 2011</td>
<td>Genome annotation: search by content</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 14</td>
<td>Thursday, March 31, 2011</td>
<td>BLAST</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 15</td>
<td>Tuesday, April 05, 2011</td>
<td>BLAST and e-values</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 16</td>
<td>Thursday, April 07, 2011</td>
<td>Multiple sequence alignment</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 17</td>
<td>Tuesday, April 12, 2011</td>
<td>Multiple sequence alignment</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 18</td>
<td>Thursday, April 14, 2011</td>
<td>Phylogeny</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 19</td>
<td>Tuesday, April 19, 2011</td>
<td>Midterm (2)</td>
<td>Ivan Erill</td>
</tr>
<tr>
<td>Week 20</td>
<td>Thursday, April 21, 2011</td>
<td>Scoring functions</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 21</td>
<td>Tuesday, April 26, 2011</td>
<td>PSI-BLAST</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 22</td>
<td>Thursday, April 28, 2011</td>
<td>Protein domains</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 23</td>
<td>Tuesday, May 03, 2011</td>
<td>CDD and PSSM</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 24</td>
<td>Thursday, May 05, 2011</td>
<td>Biomedical applications of bioinformatics</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 25</td>
<td>Tuesday, May 10, 2011</td>
<td>Career options</td>
<td>Diane Crump-Fogle</td>
</tr>
<tr>
<td>Week 26</td>
<td>Thursday, May 12, 2011</td>
<td>Biomedical applications of bioinformatics</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 27</td>
<td>Tuesday, May 17, 2011</td>
<td>Finals week</td>
<td>Maricel Kann</td>
</tr>
<tr>
<td>Week 28</td>
<td>Thursday, May 19, 2011</td>
<td>Finals week</td>
<td>Maricel Kann</td>
</tr>
</tbody>
</table>

Textbook

The course will not follow strictly a particular book, so students are encouraged to attend the lecture sessions. It will, however, rely mostly on the following as reference textbooks:

- David W. Mount, “Bioinformatics: Sequence and Genome Analysis”

Additional books recommended as accompanying introductory books are:

- Marketa Zvelebil and Jeremy O. Baum, “Understanding Bioinformatics”
- Arthur M. Lesk, “Introduction to Bioinformatics”

---

* The details of the schedule (what is taught each week) are for orientation purposes only. This does not apply Midterm and Final exam dates, which should not change without prior notice.
Grading
Grading will be based on the following items:

- Midterm & final exams (15% + 15% + 15% + 25%)
  - Midterm and final exams will consist of a main multiple-choice section and a short-question section. Answers to short questions will be graded according to their accuracy at answering the question and the grammatical/orthographical correctness of the written answer.
- Lab & homework assignments (25%)
  - Students must hand in a report for lab or homework assignments. Each assignment will also be given a specific deadline for completion, indicating a specific class at which homework should be handed in as hard copy (paper printout). Assignments cannot be turned in by email unless this is explicitly required or properly justified. Every day that an assignment is late drops that assignment grade by one letter. All lab and homework assignments will be graded individually. The relative contribution of each homework/lab assignment to the final grade (20%) will be the same and proportional to the final number of homework/lab assignments (i.e. each lab/homework counts the same [1/(#labs+#homework)] and they all add up equally to the 20% grade.
- Participation (5%)
  - Participation will be assessed in class thorough the course. Students will earn participation points by attending class, answering to questions and contributing to discussions. Students are strongly encouraged to participate in class.
  - Attending class
    - Attending all classes will earn students three fifths of their participation grade (3%). Missing classes without proper justification will drop a student’s participation grade proportionally. Students are responsible to sign in at the start of each class session. Repeatedly turning up late for class (more than 15 minutes late) without justification will also drop participation grades (3 lates count as one missing).
  - Class extensions
    - Students are welcome to prepare class extensions. These are explanatory summaries of specific material covered in a previous class or of additional material not covered in, but related to a previous class. In the former case, the student has to come up with a convincingly different way of explaining the material already covered in class. The student will have 5-10 min at the beginning/end of the class to present the material using 2-3 slides and answer questions. Preparing a class extension, delivering it and managing the following discussion properly will earn students additional participation grade (up to an additional 5%). Class extensions are limited to one per student.
- Final grading
  - Grading will be computed as follows:
    - Final grade= Midterm1 (15%) + Midterm2 (15%) + Midterm3 (15%) + Final (25%) + Assignments (25%) + Participation (5%).
  - Class extensions will add up 5% to the final grade (up to a maximum score of 100%).

Cell and laptop policy
No cell phone use is allowed in the class. No laptop use is allowed either, unless properly justified (e.g. handicapped individuals) and restricted specifically to taking class notes. This policy will be enforced strictly.

Plagiarism and cheating policy
Given their diversity of skills and backgrounds, students in this course are encouraged to talk to one another, learn from one another and share their own understanding and insights. However, all work submitted for grading must be the students’ original work. Plagiarized (either from another class member, or from an external source such as the internet) content in assignments will result in that assignment be given an instant “F”. Furthermore, and depending on the degree of plagiarism, the entire semester grade may become “F”. Cheating in assignments or in exams will elicit similar penalties. Repeat offenders will automatically be given an “F” in the course. This is in accordance with UMBC policies on academic integrity that you can find at: http://www.umbc.edu/provost/integrity.html and http://www.umbc.edu/undergrad_ed/ai/documents/ACC2011.pdf. Following ACC instructions, all instances of academic misconduct (cheating, fabrication, plagiarism and dishonesty) will be reported to the Academic Misconduct Reporting Database. The ACC will determine whether additional institutional penalties are applicable (these include specific notations on your transcript or inability to enroll in further courses at UMBC).

You will be asked to sign a document stating that you understand what academic misconduct is and that you are aware of the academic and institutional penalties associated with it.
Additional comments on plagiarism
The difference between plagiarism and (i) helping one another or (ii) using appropriate sources for your work is simple:

(i) Helping one another is discussing a problem with source code or with analysis or interpretation, so that you leave the conversation with an understanding that you lacked previously. Plagiarism is copying code, an analysis, or a piece of text from another student or from an external source without making it clear that this was copied as a quote. Copying refers to copy-pasting or to simply writing out what someone else has written.

(ii) Using appropriate sources for your work means identifying an appropriate piece of information from an external source (e.g., a web page on HTML development, an example of ANOVA analysis or a paragraph about molecular biology) and using it to reach your desired goal (to produce a web page, perform ANOVA on sequence data or explain the mechanism of gene transcription). Using appropriate sources for your work means that if you find text, code or analysis from an external source that is so relevant to what you want to do that you intend to use them, then you must make it absolutely clear that this is not your own, original work and you state clearly where exactly it comes from. In addition, mosaicism (repeatedly making use of slightly modified sentences without quotes or quoting portions of text/data/graphs/analysis so large (and/or so many) that these end up constituting a significant percentage of your work) is not an appropriate use of sources, even if all sources are properly cited.

Learning objectives
At the end of the semester, students should have:

- Acquired basic notions and vocabulary of molecular biology, computing and evolutionary theory that should enable them to interact with other class members regardless of their declared majors, read related bibliography and to attend related courses in the BIOL, CMSC and BINF majors.
- Gained an understanding of the problems faced by 21st-century biology and how biology has changed with the advent of new experimental techniques and the increasing reliance on data processing.
- Understood the basic computational methods developed to tackle these biological problems and be aware of the availability of alternative methods to address those same problems.
- Grasped the multi-disciplinary nature of Bioinformatics and Computational Biology and the importance of interdisciplinary collaborations to address most of biology current problems.
- Learned to read and interpret a popular science paper in the context of Bioinformatics*.
- Mastered basic computing skills to complete the lab assignments. These include navigation through bioinformatics web resources and application of some of the studied bioinformatics methods.

Contact information
Main contact:
Dr. Ivan Erill
- Office: BS117
- Phone: 410-455-2470
- Email: erill@umbc.edu

Other instructors:
Dr. Maricel Kann
- Office: BS 116
- Phone: 410-455-2258
- Email: mkann@umbc.edu

Teaching assistant:
Ron O’Keefe
- Office: BS 105 (Thursdays, 5:00-7:00 PM)
- Phone: 5-3087
- su57618@umbc.edu

*The implementation of paper discussion sessions will depend on the number of students enrolled in the course.