جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

# Course Syllabus: Application of AI in Bioinfomatics - CS 321

| | |
|---|---|
| **Division** | Computer, Electrical and Mathematical Sciences & Engineering |
| **Course Number** | CS 321 |
| **Course Title** | Application of AI in Bioinfomatics |
| **Academic Semester** | Fall |
| **Academic Year** | 2019/2020 |
| **Semester Start Date** | 08/25/2019 |
| **Semester End Date** | 12/10/2019 |
| **Class Schedule** (Days & Time) | 09:00 AM - 12:00 PM \| Sun |

## Instructor(s)

| Name | Email | Phone | Office Location | Office Hours |
|---|---|---|---|---|
| Vladimir Bajic | vladimir.bajic@kaust.edu.sa | +966128082386 | 4219, 3, Ibn Sina (bldg. 3) | Tuesday 10:00-11:00, Bld. 3, R4219 |

## Teaching Assistant(s)

| Name | Email |
|---|---|
| Dr. Christophe Van Neste Dr. Ashraf Kybria Dr. Adil Salhi | christophe.vanneste@kaust.edu.sa ashraf.kibriya@kaust.edu.sa adil.salhi@kaust.edu.sa |

## Course Information

| | |
|---|---|
| **Comprehensive Course Description** | **Summary**<br>The course consists of selected projects. The projects may change each year. These projects cover the application of artificial intelligence (AI) to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year, the targeted problems change. Students get assigned one project, and they work either alone or in groups of two. Students, in the interactive discussions with the whole class and the instructor, attempt to solve the project problems. Students regularly present their progress and defend their approach and result in front of the entire class. During one semester, several types of topics are dealt with (e.g., data integration; knowledge-, text- and data-mining of big biomedical data, genomic signal recognition, etc.). Students get direct experience in research methodology, report writing, presentations, and, most importantly, different ways of approaching solving AI applications for different bioinformatics problems.<br>**Projects for the 2019 year**<br>**Project 1: Genomic signal recognition**<br>There are several types of marker genomic signals that are useful in demarcating genes and their models/structure. In this project, there are four main signals we will target. In all four cases, the goal is to predict as accurately as possible each of the signals. Any AI/machine learning model can be used. Then, the models for the one type of signals will be compared regarding accuracy on genomes of several eukaryotic organisms. The signals are a) Translation Initiation Sites (TIS) in eukaryotic organisms; b) Poly(A) signals in mammalian organisms; c/ Donor splice sites in eukaryotic organisms; d) Acceptor splice sites in eukaryotic organisms.<br>**Project 2: Synonymous mutations and splice sites**<br>Sometimes there are so-called synonymous mutations in the genome, which do not alter the protein-encoding but introduce previously non-existing genomic signals. These signals may disrupt normal genome functioning, potentially leading to diseases. The goal of this project is to identify such synonymous mutations that introduce novel splice sites using the 1000 genome project data.<br>**Project 3: Applications of word2vec type methods for molecular function prediction**<br>Data in many types of problems are described by a very large number of descriptors, so-called features. This very large number of features causes the problem in the downstream analysis of the data. There are many ways how to compress information that encodes the data. One of the latest approaches is based on word2vec and similar types of methods. These methods are capable of compressing information coming from several tens of thousands of descriptors into several hundred new descriptors captured in the new feature vectors that describe original data items. Such dimensionality reduction allows for significantly more efficient analysis of the original data that can be subjected further to various machine learning and AI processing.<br>**Project 3a. Apply above to the analysis of the function of different transcripts.**<br>**Project 3b. Apply above to the analysis of different data describing cancers.**<br>**Project 4: Deep Learning Networks for analysis of free-text data**<br>Develop methods for classifying or clustering or summarizing sentences and abstracts of biomedical scientific documents. Utilize data from the PubMed resource. Apply different embedding techniques, e.g., sentence embedding, and different deep neural networks in the solution. For example, using Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Hierarchical Attention Network (HAN), etc.<br>**Project 5: Deep Learning Networks for identification of entity association**<br>Many problems across different fields relate to identifying and extracting correct relationships between different entities from free text. Develop methods for identifying the correct relationship between specific biomedical entities. Apply different embedding techniques and deep neural networks in the solution.<br>Possible examples could be associations between genes/proteins and diseases; mutations and diseases; mutations and pathways; mutations and Gene Ontology categories; drugs and diseases; drugs and pathways; etc.<br>Data will be provided.<br>**Project 6: Ranking edges in a biological relationship network**<br>Biological networks represent networks of relationships between various biomedical entities. When a subnetwork of such large networks is selected, develop a method to rank the edges of such subnetwork so that the edges with a higher likelihood of representing the correct relationship between the linked nodes (entities) are ranked higher than those with lower likelihood.<br>**Project 7: Enrichment of topic-specific dictionaries**<br>Develop methods that can enrich topic-specific dictionaries by adding terms that are not included but fall into the same topic. Use semantic similarities and deep neural networks to develop a solution.<br>Data will be provided. |
| **Course Description from Program Guide** | These projects cover application of AI to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year problems change. Students get assigned one (1) project and they work either alone or in groups of 2. Students in the interactive discussions with the whole class and the instructor solve the project problems. Students regularly present their progress and defend their approach and results in front of the whole class. During one (1) semester several types of topics are dealt with. Students get direct experience in research methodology, report writing, presentations and, most importantly, different ways of approaching solving AI problems |
| **Goals and Objectives** | The course consists of selected projects. The projects may change each year. These projects cover the application of artificial intelligence (AI) to some of the relevant problems of analysis of large biological data and generally deal with complex information. Each year, the targeted problems change. Students get assigned one project, and they work either alone or in groups of two. Students, in the interactive discussions with the whole class and the instructor, attempt to solve the project problems. Students regularly present their progress and defend their approach and result in front of the entire class. During one semester, several types of topics are dealt with (e.g., data integration; knowledge-, text- and data-mining of big biomedical data, genomic signal recognition, etc.). Students get direct experience in research methodology, report writing, presentations, and, most importantly, different ways of approaching solving AI applications for different bioinformatics problems. |

| | |
|---|---|
| **Required Knowledge** | It is desirable that students have proficiency in programming.<br>Preferred knowledge of as many as possible of C/C++, Java, Python, R, Matlab, HPC (parallel computing).<br>The previous exposure to machine learning or data analytics is needed. |
| **Reference Texts** | doi: 10.1186/s12864-017-4033-7.<br>Word2Vec, http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf<br>doi: 10.1093/bioinformatics/bts638.<br>doi: 10.1016/j.jbi.2009.09.004.<br>doi: 10.1093/nar/gkp479.<br>https://www.aclweb.org/anthology/D17-1024<br>https://medium.com/jatana/report-on-text-classification-using-cnn-rnn-han-f0e887214d5f<br>https://medium.com/jatana/unsupervised-text-summarization-using-sentence-embeddings-adb15ce83db1<br>https://arxiv.org/pdf/1706.03960.pdf<br>doi: 10.1186/1471-2105-9-S12-S8.<br>doi: 10.1371/journal.pone.0102039<br>https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2<br>https://www.microsoft.com/developerblog/2016/09/13/training-a-classifier-for-relation-extraction-from-medical-literature/ |
| **Method of evaluation** | **20.00**% - Attendance and Participation<br>**15.00**% - Written report<br>**10.00**% - Presentation<br>**5.00**% - Oral presentation<br>**50.00**% - Course Project(s) |
| **Nature of the assignments** | Research project as defined in the course description complemented by the presentation of results, discussions on the methods of solution, written mid-term and final report. |
| **Course Policies** | Student absence of more than three times without justifiable reason will lead to failing the course. |
| **Additional Information** | Assessment of students is continuous. |

| Tentative Course Schedule<br>*(Time, topic/emphasis & resources)* | | |
|---|---|---|
| **Week** | **Lectures** | **Topic** |
| 1 | Sun 08/25/2019 | Course introduction<br>Projects introduction |
| 2 | Sun 09/01/2019 | Different topics from data analysis and machine learning |
| 3 | Sun 09/08/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 4 | Sun 09/15/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 5 | Sun 09/22/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 6 | Sun 09/29/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 7 | Sun 10/06/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 8 | Sun 10/13/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 9 | Sun 10/20/2019 | Writing reports |
| 10 | Sun 10/27/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 11 | Sun 11/03/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 12 | Sun 11/10/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 13 | Sun 11/17/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 14 | Sun 11/24/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 15 | Sun 12/01/2019 | Students present progress in their projects.<br>Discussions by all students. |
| 16 | Sun 12/08/2019 | Final reports discussion |

**Note**

The instructor reserves the right to make changes to this syllabus as necessary.