CSCI-693 Syllabus

• **Instructor:** Dr. Elena Harris
• **E-mail:** eyharris@csuchico.edu
• **Office:** OCNL 221
• **Office Hours:**
  MW 1:00pm-2:00pm and 3:00pm-3:45pm
  TR 12:30pm-1:30pm
  Or by an appointment
CSCI-693 Syllabus

• Grade Evaluation:
  • Literature Review 10%
  • Critical Review 10%
  • Research Project 60%
    – Weekly Progress Reports (20%)
    – Software (20%)
    – Research Report (20%)
  • Oral Presentation 10%
  • Class Discussions 10%
CSCI-693 Syllabus

Research Project, 60% of the final grade
Each student will do a research project on a topic given by the instructor

Motivation:
1. Learning by doing
2. Real-life project that will contribute to scientific research
3. The results will be most likely published in a scientific journal
4. Real-life experience in one of the fast growing fields in CS, bioinformatics
CSCI-693 Syllabus

Research Project, 60%

- Weekly Progress Reports (20%)
- Software (20%)
- Research Report (20%)
Research Project, 60%
  – Weekly Progress Reports (20%)
1 hour outside of class weekly meeting with the instructor

Motivation:
1. Real-life projects require to report weekly results (collaboration with biologists)
2. Helps to organize and structure work (helps with procrastination)
3. Timely feedback
4. Provides a lot of help from the instructor
CSCI-693 Syllabus

Research Project, 60%
  – Weekly Progress Reports (20%)

1 hour outside of class weekly meeting with the instructor

Meeting time available:
MW 11:00am-12:00pm
MW 12:00pm-1:00pm
TR  4:00pm-5:00pm
TR  5:00pm-6:00pm
F   12:00pm-4:00pm
Research Project, 60%

– Software (20%)

1. High-quality tested code submission
2. Manual for using the software
3. Software will be a part of open-source tools available for biologists to conduct research
4. Software is a part of publication results
CSCI-693 Syllabus

Research Project, 60%
   – Research Report (20%)

1. Gives experience of writing technical papers
2. Clearly describes methods used in developing software and/or analysis of bio-data
3. Provides the basis for a publishable paper
4. Will be a part of a research paper in a journal
CSCI-693 Syllabus

- Literature Review 10%
- Critical Review 10%
- Oral Presentation 10%
- Class Discussions 10%
CSCI-693 Syllabus

Lecture attendance is MANDATORY.
Failure to attend at least 80% of the lectures will result in an F in this class.
CSCI-693 Syllabus

Plan to spend at least 6 hours weekly working on the research project (not including weekly meetings with the instructor)
<table>
<thead>
<tr>
<th>Scale (inclusive)</th>
<th>Letter Grade</th>
<th>University Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>93-100</td>
<td>A</td>
<td>Superior Work</td>
</tr>
<tr>
<td>90-92</td>
<td>A-</td>
<td></td>
</tr>
<tr>
<td>87-89</td>
<td>B+</td>
<td></td>
</tr>
<tr>
<td>83-86</td>
<td>B</td>
<td>Very Good Work</td>
</tr>
<tr>
<td>80-82</td>
<td>B-</td>
<td></td>
</tr>
<tr>
<td>77-79</td>
<td>C+</td>
<td></td>
</tr>
<tr>
<td>73-76</td>
<td>C</td>
<td>Adequate Work</td>
</tr>
<tr>
<td>70-72</td>
<td>C-</td>
<td></td>
</tr>
<tr>
<td>67-69</td>
<td>D+</td>
<td></td>
</tr>
<tr>
<td>60-66</td>
<td>D</td>
<td>Minimally Acceptable Work</td>
</tr>
<tr>
<td>0-59</td>
<td>F</td>
<td>Unacceptable</td>
</tr>
</tbody>
</table>
Research Projects

1. Analysis of Differential Gene Expression using real RNA-seq data and existing tools (2 people)
2. Associative analysis of micro-RNA abundance and gene expression using real RNA-seq data
3. Build a tool for RNA-seq alignment including splicing identification from RNA-seq data
4. Build a tool for isoform assembly and quantification from RNA-seq data
5. Build a tool for differential gene expression from RNA-seq data
6. Discovering novel genes using real RNA-seq data
Research Projects


8. Improvement of mapping tool BRAT-BW (indels, local alignment, RRBS)

9. Identifying Hypomethylated and Differential-methylated regions given methylome(s)

10. Speeding up alignment of reads using perfect repeats

11. Tools for associative analysis of HMRs/DMRs and gene expression

12. Identifying Allelic Differentiate Regions, AMRs
Biological Background

• To understand research projects
• It is important to know enough biology to solve a problem correctly
• DNA carries genetic information
• DNA is a double helix of two complementary strands formed by four nucleotides (bases): **Adenine, Cytosine, Guanine and Thymine**
Next generation sequencing (NGS) technology revolutionized genomic research.

NGS technology allows fast and inexpensive DNA sequencing producing hundreds of millions of DNA sequenced reads:

- Sequence DNA fragments
- Map reads back to a reference genome
- Analyze the data
Next-generation Sequencing

- Mapping DNA reads back to a reference genome is the first step in the data analysis.
- Mapping short sequenced reads back to a reference genome is a string search problem: given a text and a query, find all (approximate) occurrences of the query in the text.

```
ATATGTTAGTCAAGTTAAGACCTATGTTAG
```

TCAAG
• **Gene expression** is the process by which DNA is transcribed into mRNA (eventually translated into proteins)
Transcription is the first step of gene expression, in which a particular segment of DNA (gene) is copied into RNA by the enzyme RNA polymerase.
Transcription in Eukaryotes:

1. A gene is copied into a primary RNA, mRNA precursor.
2. Introns are removed in a process called **splicing**.
3. Poly-tail of A’s is attached at 3’ end of mature mRNA.
**Splicing** is a process of removal of introns from the primary RNA.
**Translation** is the process of synthesizing a protein from mRNA

**Genetic code:** Triples of mRNA bases (codons) associate with corresponding amino acids
**Figure 7.1** All the triplet codons have meaning: 61 represent amino acids, and 3 cause termination (STOP).

Table: Genetic Code

<table>
<thead>
<tr>
<th>First base</th>
<th>Second base</th>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>UUC</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UUA</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UUG</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UCU</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UCC</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UCA</td>
<td>Ser</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UCG</td>
<td>Ser</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>CUU</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUC</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUA</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CUG</td>
<td>Leu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCU</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCC</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCA</td>
<td>Pro</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CGA</td>
<td>Pro</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>AUU</td>
<td>Ile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUC</td>
<td>Ile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUA</td>
<td>Ile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUG</td>
<td>Met</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACU</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACC</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACA</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ACG</td>
<td>Thr</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCU</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCC</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCA</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCG</td>
<td>Val</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>GUU</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUC</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUA</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUG</td>
<td>Val</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCU</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCU</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCA</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCG</td>
<td>Ala</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGU</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGU</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCC</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAC</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAG</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGG</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAU</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAA</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GAG</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGC</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGG</td>
<td>Asp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GGU</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCC</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCA</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCG</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUA</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AUG</td>
<td>Gly</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GUG</td>
<td>Gly</td>
</tr>
</tbody>
</table>

[YouTube Video Link: https://www.youtube.com/watch?v=OEWOZS_JTgk]
Alternative splicing

**Example (mouse gene)**

α-TM EXON GENE ORGANIZATION

5'UT 1-20 20-00 30-00 81-125 126-154 155-200 201-254 255-304 3'UT

α-TM mRNA TRANSCRIPTS

- Striated muscle
- Striated muscle
- Myoblast
- Smooth muscle
- Nonmuscle/fibroblast
- Hepatoma
- Brain
"RNA-Seq-alignment" by Rgocs (talk) - Transferred from en.wikipedia to Commons..
Licensed under Attribution via Wikimedia Commons -
http://commons.wikimedia.org/wiki/File:RNA-Seq-alignment.png#mediaviewer/File:RNA-Seq-alignment.png
Mapping RNA-seq Data

Difference between this read and the reference sequence

Consensus sequence

Reference sequence

Read sequence
Analysis of Differential Gene Expression using real RNA-seq data and existing tools (2 people)

Analysis is done using existing tools for RNA-seq

1. Given a set of RNA-seq reads from $k$ different samples, genome, and genes annotation, map RNA-seq reads back to the reference genome

2. Count how many reads are mapped to each gene

3. Using statistics, identify which sets of genes were differentially expressed in the samples
Discovering novel genes using real RNA-seq data

1. Given a set of RNA-seq reads from \( k \) different samples, genome, and genes annotation, map RNA-seq reads back to the reference genome
2. Identify reads mapped to un-annotated regions of the genome
3. Identify splice-junctions of the reads mapped to these regions
4. Assemble mRNA from the reads
5. Run mRNA consensus against existing data bases for coding proteins
6. (If time) Benchmark your tool against existing state-of-the-art tools
Build a tool for RNA-seq alignment including splicing identification from RNA-seq data

1. Given a set of RNA-seq reads, genome, and genes annotation, design a mapping tool for RNA-seq reads using BWT and Farragina-Manzini index (base of this tool is existing BRAT-BW tool)

2. Tool will identify splice-junctions of the reads mapped to the genome

3. Benchmark your tool against existing state-of-the-art tools
Build a tool for isoform assembly and quantification from RNA-seq data

1. Given a set of mapped RNA-seq reads to a given genome, and genes annotation, design a tool for isoform assembly and quantification
2. Tool will identify splice-junctions of the reads mapped to the genome
3. Benchmark your tool against existing state-of-the-art tools
Build a tool for differential gene expression from RNA-seq data

1. Given a set of mapped RNA-seq reads to a given genome, and genes annotation, design a tool for differential gene expression (for two or more samples)

2. Tool will use several existing statistical methods

3. Benchmark your tool against existing state-of-the-art tools
Micro-RNA Data

- Very short segments 20-25 bp
- Repeats in the genome
- Associative analysis with gene expression
Mapping miRNA reads
Mapping tool for micro-RNA data. Target prediction of miRNA and mRNA binding

1. Given a set of miRNA-seq reads and a genome, design a tool for mapping miRNA reads back to the genome
2. Adapter trimming
3. Mapping short reads with up to 2 mismatches (using BRAT-BW as a base)
4. Given miRNA and RNA-seq, predict binding sites
5. Calculate the abundance of miRNA from the mapped reads
6. (If time) Compare your tool against state-of-the-art tool(s)
Associative analysis of micro-RNA abundance and gene expression using real RNA-seq data

Analysis is done using existing tools

1. Given a set of mapped miRNA-seq reads
2. Given a set of mapped RNA-seq reads
3. Carry out associative analysis between miRNA abundance and mRNA enrichment using two or more samples
• DNA methylation is one of the mechanisms used by cells to control gene expression.
• DNA Methylation plays an important role in
  – Embryonic development
  – Cellular differentiation
  – Overall genome stability
• Errors in DNA methylation have been associated with human diseases
• Identification of methylated cytosines is the first step in the analysis of methylation
BS-seq Approach to Generate Methylomes

Original DNA

\[
\begin{align*}
5' & \quad \text{DS}^+ \quad A \quad A \quad C^m \quad G \quad T \quad C \quad G \\
3' & \quad \text{DS}^- \quad T \quad T \quad G \quad C^m \quad A \quad G \quad C
\end{align*}
\]

Bisulfite treatment, PCR

\[
\begin{align*}
5' & \quad \text{PCR}^+_1 \quad A \quad A \quad C \quad G \quad T \quad T \quad G \\
3' & \quad \text{PCR}^-_2 \quad T \quad T \quad G \quad C \quad A \quad G \quad T
\end{align*}
\]

Reference

\[
\text{ATGTTTTAAGCTCATAAATGTCAACATAAATAAA}
\]

Mapped reads

\[
\text{ATGTTTTAAGCTCATAAATGTCAACATAAATAAA}
\]

Methylation level

\[
\text{3/10} \quad 0/21 \quad 7/21
\]

Read coverage
Hypomethylated regions, HMR, appear as valleys of very low methylation level.
• Differentiated methylated regions, DMR, are the regions whose methylation is different between two conditions
Allelic methylated regions, AMR, appear as valleys of intermediate methylation level between very low and very high.
These projects will require benchmarking against existing tools

8. Improvement of mapping tool BRAT-BW (indels, local alignment, RRBS)

9. Identifying Hypomethalated and Differential-methylated regions given methylome(s)

10. Speeding up alignment of reads using perfect repeats

11. Tools for associative analysis of HMRs/DMRs and gene expression

12. Identifying Allelic Differentiate Regions, AMRs
Thank You