CS-150 Selected Topics: Computational biology

Description:
CS-010, Cr. 4.

The course offers a range of topics in structural bioinformatics, including: an introduction to protein structures, protein structure representations, structure geometry and symmetry, computational analysis of protein sequences and structures, mutations, sequence structure relationships, comparison of structures, visualization and modeling, structure prediction, protein fold classification using machine learning, protein-protein interactions, coarse-grain modelling of proteins, and molecular dynamics simulations.

Prerequisite: CS-010 or consent from the instructor. No prior background in biology required.

Course meeting hours and location:
10-11:50AM TR,
Location: TBD

Course Instructor:
Guang Song
Office: Winter 304
Office hours: TBD, or by appointment
Email: gsong@westmont.edu

Textbook:

Reference Texts:
1. Protein Structure and function, Petsko, 2004
2. Exploring bioinformatics, A project-based approach, C. Clair and J. Visick, 2015
4. Introduction to Protein Structure, Branden and Tooze, 1998
6. Textbook of Structural Biology, Liljas et al, 2010

Commented [GS1]: The course covers a range of topics on proteins: protein sequence, structure, dynamics, and function, etc.

Commented [GS2]: The textbook, Protein Actions, is awarded for the best new textbook in the life sciences in 2018. The book takes mostly a computational approach to life science.
Learning Outcomes and GE requirements:

We have many goals for students taking courses in the Computer Science Program. Some of them are specific to particular courses, but almost all are examples of our Program Learning Outcomes (PLOs).

These are the overarching Computer Science Program Learning Objectives:

1. **Programming**: Students will be able to write computer programs well. They will effectively use computer languages to efficiently implement automated abstractions.
2. **Analysis**: Students will be able to develop and understand new solutions to algorithmic problems independent of implementation language and be able to accurately analyze and improve existing algorithms.
3. **Context**: Students will be able to interrogate real-world problems, creatively, in response to existing, discovered and unknown data so that they can make good design decisions.
4. **Christian Connection**: Students will be able to assess, evaluate, choose and defend a position regarding the social impacts of computational artifacts on individuals and society in context of general and specific Christian revelation.

Some of the things that you will learn map to the above learning objectives as follows:

- Learning to develop computational methods to solve problems in computational biology. - PLO1
- Learning how to read programming specifications - PLO2
- Figuring out how to think computationally - PLO1, PLO3
- Understanding the ethical implications of what we are creating - PLO4

Attendance:

Attendance is required.

Grading and Studying:

Grades are managed in Canvas. You should be able to see your current grade at any given time based on the assignments that have been turned in and evaluated.

This class has several components:

1. (10%) Attendance and **reading assignments**
2. (40%) **programming homework assignments**

**Commented [G53]**: The reading assignments expose students to various computational (and experimental) approaches to life sciences.

**Commented [G54]**: Most homework assignments require programming and computation.
Tentative topics:
- protein structure and geometry, protein sequence alignment and analysis,
- homology modeling and alpha fold, PDB files, visualization, protein
  geometry calculations, protein fold classification using machine learning,
- elastic network models for protein dynamics modelling, molecular
  dynamics,
- PDB files, visualization,
- protein geometry calculations,
- protein fold classification using machine learning,
- elastic network models for protein dynamics modelling, molecular
  dynamics

Homework policy:
- For assignments the late penalty is 1% per hour after the moment
  that it is due. Due dates are specified with a date and time in local
  time. For example, if an assignment's due date is 11:59pm on
  Monday night and it is turned in on 11:59pm Tuesday night it is 24
  hours late will be result in a 24% penalty.
- Homework re-grade requests must be submitted in the written form
  within one week after a grade is received.

3. (50%) Final project and presentation
   What is expected: summarize and review one research area, propose a new
   approach, implement the proposed approach, write a final report and give a
   presentation on it.

   Project mentoring: Once the topics have been agreed upon with the instructor,
   students will start to meet with instructor, individually or in small groups
   regularly to talk about project progress and to receive feedback.

   Possible project areas/topics:
   - protein structure alignment
   - protein geometry
   - identification of domains
   - modeling
   - tertiary structure prediction
   - protein-protein interaction
   - molecular dynamics
   - statistical potentials
   - using big data
   etc., etc.

   Students may choose a different subject area than those listed above.

   Project proposal should include: a definition of the problem, why the problem is
   significant, the current state (what is known about the problem), description of the
   proposed approach, justification of the proposed approach, how the results are to
   be evaluated, expected finding of proposal work, significance of the results to be
   obtained.
Project presentation and report should address each of the above issues in more details, and in addition, present results and discussions.

As the class progresses I reserve the option to alter the percentages to reflect the way the class unfolds.

*Grade Scale:* the standard grading scale.
<table>
<thead>
<tr>
<th>Week</th>
<th>Section</th>
<th>Readings: Chapters</th>
<th>Topics</th>
<th>Homework and Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Introduction to proteins</td>
<td>1, 2</td>
<td>Go over the Syllabus</td>
<td>Mostly reading, in-class discussion, Presentation Seminar style (discuss chapter 1,2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What are atoms, chemical bonds, amino acids, what are protein functions, enzymes, how a protein functions, how proteins are synthesized, from DNA to RNA to protein, DNA: double helix, structure clearly reveals its functions. Proteins varies significantly in shape and have a vast array of different functions. Enzymes, different roles of proteins. What is life?</td>
<td></td>
</tr>
<tr>
<td>3-5</td>
<td>Protein geometry (2.5 weeks)</td>
<td>9</td>
<td>Visualization (pymol (watch a tutorial), rasmol, VMD), geometry, pdb format, Symmetry, ligand migration pathways in myoglobin, alpha-shape, How protein shape often dictates their dynamics and thus function Structure alignment (two chains of the same length) Structure matching: FATCAT (see structure matching slides)</td>
<td>HW1: a programming assignment to read pdb structures and perform geometric calculations. HW2a: computing Solvent accessible surface areas HW2b: find a largest clearance path to heme using Voronoi diagram</td>
</tr>
<tr>
<td>5-7</td>
<td>Protein sequence analysis (2 weeks)</td>
<td>8</td>
<td>Sequence alignment (chapter 8, pg. 181) String search Sequence evolution Sequence conservation: inference about protein stability and function. <strong>Dynamic programming</strong> (1 lecture of its own, see Vazironi’s book chapter) Substitution matrix Needleman-Wunsch (1970) (global alignment) Smith and Waterman (1981) (local alignment, see Exploring bioinformatics p. 57): it does NW first, then from the highest score cell, trace back to find the local alignment. Blastp (genome assembly)</td>
<td>HW Idea1: Implement the sequence alignment algorithm. HW Idea2: Perform a Blastp run and some sequence analysis</td>
</tr>
</tbody>
</table>
**Multiple Sequence Alignment (MSA)** (see slides)
- PSI-Blast or Clustral
- Co-evolve
- Constructing a Phylogenetic tree
  More in the textbook

<table>
<thead>
<tr>
<th>7-8</th>
<th>Protein structure prediction (1 week)</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mutagenesis using (FoldX), 1 lecture on homology modelling (Swiss-Model, FoldX)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>An example globin sequence: &gt;A0A4522K63_AEGTS/48 (taken from the downloaded all sequences (over 10K) of the Globin family)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>QEALVLSAWDAMKGDSAAIALKFFLRIFEIAPAAKPM FPFIRDAGEDAPLESHPKLAHAVTVFVMACESATQL GDVKVREATLRRLGATHVRAVGADAHFVVKTALL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Swiss Model did a great job. alphaFold, rosettaFold, (see homology_modeling slides)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 lecture on Docking (Hawkdock, Clustpro, Hdock); see docking slides. Covid-19 (RBD sequence): [MSA, how sequences evolve (trajectory), phylogenetic tree,] Wuhan variant, RBD domain and its binding to 7 antibodies, how the binding affinity change for Delta variant, and then for omicron. (applications of FoldX and Hawkdock).</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein-ligand binding (Haddock’s prodigy program, Kdeep) Statistical potential and threading (1 lecture)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In-class Discussion topics: what is the chapter</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8-10</th>
<th>Machine Learning (ML), Protein structure classification (2 weeks)</th>
<th>2 in-class labs: Learn the tools and apply them</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Structural matching (which may be used for assigning fold)! (DaliLite5, mTMalign, Tmalign, Deep-Align, And comparison with our ML-based approach</td>
<td>Lab1: we have several proteins, we have only their sequence, we want to know how they may interact. (homology model or alphaFold, and then docking)</td>
</tr>
<tr>
<td></td>
<td>ML (svm, NN, GNN), classification, folds, functions, SCOP, CATH</td>
<td>Lab2: RBD binding and docking</td>
</tr>
<tr>
<td></td>
<td>Features, build dataset, training, import to PyTorch. Backward propagation: watch A. Ng’s lectures (need a more accurate representation of protein graph)</td>
<td>Project ideas: Develop a ML program for protein structure classification</td>
</tr>
<tr>
<td>10-12</td>
<td>Simulations of protein dynamics (2 weeks) + Misc. (0.5 week)</td>
<td>10, 13</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>13-14</td>
<td>Protein modelling (2 weeks)</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Academic Honesty**

Each assignment or project is to be the product of your (or your team’s) own intellectual efforts.

Dishonesty of any kind may result in loss of credit for the work involved and the filing of a report with the Provost’s Office. Major or repeated infractions may result in dismissal from the course with a grade of F. Be familiar with the College’s plagiarism policy, found at: [http://www.westmont.edu/offices/provost/plagiarism/plagiarism_policy.html](http://www.westmont.edu/offices/provost/plagiarism/plagiarism_policy.html)

**Students with disability**

Students who have been diagnosed with a disability are strongly encouraged to contact the Office of Disability Services as early as possible to discuss appropriate accommodations for this course. Formal accommodations will only be granted for students whose disabilities have been verified by the Office of Disability Services. These accommodations may be necessary to ensure your equal access to this course. Please contact Sheri Noble, Director of Disability Services. (310A Voskuyl Library, 565-6186,
snoble@westmont.edu) or visit the website for more information:
http://www.westmont.edu/offices/disability

Emergency Procedure

In the event that an emergency occurs during instruction, it is important to be familiar with the practices in place for the classroom. Please review the document at https://integready.app.box.com/AnticipatingInClass and direct any questions or concerns to the Office of Institutional Resilience.