CHEN 4900 Topics in Chemical Engineering Fall 2022

Computer Simulation in Biology: Machine Learning & Physical Methods

CHEN 4900

Instructor

Professor Ben O'Shaughnessy
3 Points

Prerequisites

Basic thermodynamics

Textbook

No required textbook

Recommended background reading

Understanding Molecular Simulation: From Algorithms to Applications
Daan Frenkel and Berend Smit (Academic Press, 2002)

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition
Aurélien Géron (O'Reilly)

Molecular Biology of the Cell, 6th Edition
B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (Garland Press, New York).

The Fokker-Planck Equation: Methods of Solution and Applications
H. Risken (Springer, 1996)

Course Requirements and Grading

Participation in class discussion. Weekly assignments: reading or coding (Jupyter Notebook). Final exam: take home, written report on an assigned course topic. Final grade based on participation in class discussion, weekly assignments and final take-home exam.

Rationale for course and course summary

A vast research effort in science, engineering, medicine and biotechnology revolves around the machineries of life at the cellular and multicellular organism level. Computer simulation is central to these spheres of biological research, whether the goal is to unlock the secrets of life, to prevent or treat disease, to develop new drugs or to evolve novel biotechnologies with societal and commercial impact. Computational methods are needed because the machineries and systems of life have high complexity. In individual cells, vast numbers of molecules and multimolecular structures coordinate for essential functions, such as synthesis of nutrients, secretion of signaling molecules, migration, cell division, gene expression, chromosome replication, defense against pathogens and countless others. At the supracellular level, in humans and other organisms multiple cells coordinate in complex systems to execute tasks such as tissue reshaping during embryogenesis, specialized organ function and synchronized firing of neurons in complex neural circuits for brain function. Due to this complexity and increasingly sophisticated experimental methods, vast amounts of data have recently become available. A major current challenge is to integrate computer simulation with machine learning methods to extract information from big data and

improve and accelerate simulations.

The subject of this course is the integration of physical and machine learning methods for computer simulation of multimolecular and multicellular living systems. We discuss how simulation can unravel mechanisms, help combat disease and aid biotechnology development. Introduction to biological principles and jargon, with cellular and multicellular examples. Introduction to principles and results of statistical mechanics. Methods of many-molecule molecular dynamics (MD) simulation, from isolated deterministic MD to non-isolated stochastic MD (temperature, fluctuation-dissipation theorem). Simulations of biological molecular, subcellular and multicellular systems at different scales: from atomistic to mildly coarse-grained MARTINI MD of protein and viral systems, to ultracoarse-grained MD of cellular and multicellular systems. Introduction to neural networks (NN) and machine learning (ML). Using ML and NN to accelerate MD simulations by enhanced sampling, to determine simulation parameters from big data, and to extract complex simulation behaviors including dimensional reduction. Using ML and NN for systematic coarse-graining. Throughout, methods will be illustrated by review and class discussion of recently published research papers. Students will gain hands-on exposure to simulation code through use of Jupyter Notebook.

Transcript title (30 char max):
Computer Simulation Biosystems

Syllabus

Topic 1. Overview of biological systems. Molecules in cells including proteins and Protein Data Bank (PDB) database for protein structures. Subcellular structures: cytoskeleton, actomyosin rings and networks for cell division and shape regulation, membrane-enclosed organelles, neurotransmitter release machinery in neurons. Exocytosis, release of insulin and other hormones. Misregulated cell division and cancer. Tissue reshaping during embryogenesis, supracellular force networks. Neurotransmission, neural circuitry in the brain. Viruses: CoV-2 spike protein cell entry mechanism, influenza. Antiviral vaccines and drugs.

Topic 2. Review of Statistical Mechanics. Entropy, molecular origin of the 2nd law. Boltzmann distribution, formulae for entropy, free energy.

Topic 3. Molecular dynamics (MD) computational methods. Deterministic MD simulations for isolated systems. Coding up isolated hard core or interacting gas/liquid MD simulations. Introduction to Jupyter Notebook (web-based interactive computational environment). Jupyter Notebook documents for live code, numerical simulation, data visualization. Emergence of temperature concept. Velocity distributions, fluctuations.

Topic 4. Stochastic MD simulations (non-isolated systems). Drag forces, random forces to impose temperature. Fluctuation-dissipation theorem. Coding and running a stochastic MD simulation using Jupyter Notebook.

Topic 5. All-atom molecular dynamics (MD) simulation methods. Using PDB files as input to MD simulations. Creating, analyzing MD simulation trajectories: GROMACS and other MD simulation packages. All-atom force fields (e.g. GROMOS-96, AMBER, CHARMM.)

Topic 6. Coarse-grained (CG) force fields: Martini. Martinizing (CG procedure for a protein). Molecular dynamics Martini simulations. All-atom and Martini simulations of viral fusion proteins. Ultra coarse-grained (UCG) force fields. HOOMD-blue MD simulation package. UCG molecular dynamics of extended membranes and cellular membrane fusion machinery.

Topic 7. Introduction to machine learning and neural networks.

Topic 8. Enhanced sampling to accelerate MD simulations. Using neural networks to identify collective coordinates and to impose enhanced sampling.

Topic 9. Using machine learning and large experimental datasets to optimize simulation parameters and to systematically coarse grain molecular representations.

Topic 10. Machine learning and neural networks to extract complex behaviors from big simulation output datasets. Neural networks to identify dimensional reduction.