Introduction to Computational Biology

Instructor(s):
András Aszódi
Peter Sarkozy

Summary:
If the 20th century was the "Age of Physics", then without doubt the 21st century will be the "Age of Biology". This revolution in biomedical research has been driven by the introduction of high-throughput experimental technologies that generate enormous data sets. The ensuing "data deluge" opens up exciting new opportunities to computer scientists because the scientific potential of the data can be exploited only if they are analyzed with sophisticated computational methodologies. In the not too distant future the majority of biomedical researchers are expected to be data analysts rather than experimental biologists or chemists. This is why pharmaceutical companies and biotech enterprises already offer very attractive career paths to computer scientists with strong experience in bioinformatics, data mining/biostatistics, AI/machine learning and related fields.

The main aim of the "Introduction to Computational Biology" (ICB) course is to teach Computer Science majors how to apply their mathematical and computational knowledge to practical data analysis problems in the pharma and biotech industries. Only high-school level biology is required; we rather focus on the essential "soft skill" of how to communicate with biologists. In short, the course provides the intellectual foundations for those students who may later decide to join the "biotech revolution", contributing to new discoveries of great relevance to society.

Synergies with other courses:
The ICB course is strongly recommended to those taking the "Data mining", "Structure and dynamics of complex networks" or the "Deep learning" courses as these methodologies are immediately applicable to biological data analysis tasks.

The course consists of two parts: Biocybernetics and High-throughput data analysis. Biocybernetics (A. Aszódi) addresses information processing in living systems, in particular the analysis of gene expression regulation and metabolic control. High-throughput data analysis (P. Sarkozy) focuses on the analysis of large-scale genomic sequencing experiments, with special emphasis on medical decision support.

Motivation and perspective:
You will gain specific skills in how to understand and model biological phenomena, and learn how to collaborate efficiently with biologists and chemists working in pharmaceutical research. In addition, you will learn scientific methodologies applicable to model complex phenomena in any discipline, further enhancing your employability.

Prerequisites:
- Good mathematical skills. We will teach you how to use basic calculus (differential equations), linear algebra (vectors, matrices) and basic probability theory.
- Python 3 programming skills.
- High school biology. We will teach you all the necessary knowledge beyond that level.

Textbooks:
Topics in detail

Part I: Biocybernetics (A. Aszódi)

Introduction to biocybernetics:

First-order kinetics:

Biochemical kinetics:

Principles of genomic regulation:
The molecules of life: DNA, RNAs, proteins. The "central dogma": transcription and translation. The differences between prokaryotic and eukaryotic organisms. Prokaryotic genome regulation: the E. coli lac operon, the lambda phage bistable switch, the "Repressilator" synthetic genetic network. Eukaryotic genome organisation: chromosome structure, splicing, epigenetic regulation. The self-regulatory Hes1 network.

Oscillations and chaos:

Stochastic biochemical kinetics:

Systems modelling:

The theory of evolution:

Computing with biomolecules:
Fourier analysis:

Regulation in spacetime:

Artificial life:

Part II: High-throughput data analysis (P. Sarkozy)

Introduction to molecular genetics:
Role and characteristics of DNA in organisms. Mutation types, population genetics, linkage disequilibrium, transcription and translation of DNA to proteins, gene expression, epigenetic modifications, the Human Genome Project and the path to personalized medicine.

Overview of DNA sequencing technologies:
Sanger sequencing to single-molecule real-time DNA sequencing, in vitro diagnostics, high-throughput measurement methods, partial genetic association studies, genome-wide association studies, single-molecule real-time DNA sequencing.

High-throughput measurements:
Quality control, filtering, common failure modes and platform-specific error profiles of common measurement methods, sample multiplexing and study design.

Strings in bioinformatics:
Naïve exact matching, Z algorithm, naïve approximate matching, radix sorting, suffix indices, longest common prefix, Burrows-Wheeler transformation.

Mapping and assembly of large, complex genomes:
Alignment scoring schemes, de-novo assembly, reference mapping, the Smith-Waterman algorithm, the Needleman-Wunsch algorithm, understanding and correcting alignment bias in DNA sequencing, local and global alignment, paired-end sequencing.

Phylogenetics and metagenomics:
Multiple sequence alignment, clustering approaches, distance metric, phylogenetic tree construction, metagenomic population studies, molecular clock hypothesis.

Interpretation of results:
Identifying variants, detecting somatic mutations, heterogeneous population sequencing, construction of local phylogenetic trees for cancer evolution, resolving haplotypes, copy number variations, large-scale genomic rearrangements.

DNA editing:
DNA repair mechanisms, homology directed repair, DNA editing in-vivo and in-vitro, CRISPR-CAS9 system, zinc finger nuclease technology, in vivo delivery methods, RNA interference, system biological
approach to diseases.

Models in Computational Genomics:

Bayesian frameworks in bioinformatics:
Frequentist vs. Bayesian approaches, naïve Bayes classifiers, Bayesian networks, probabilistic classifiers, network structure learning, semantic technologies for computational biology.

The co-evolution of man and machine:
Brain-computer interfaces, the challenges of biosynthetic organisms, the extension of mankind with weak artificial intelligence, challenges posed by strong artificial intelligence.

Homework assignments

Part I:
Data analysis and simulation tasks (reaction kinetics of biochemical pathways, predator-prey models, biological pattern formation models etc.) using the CoCalc on-line computer algebra system.

Part II:
Data analysis of an in-silico genetic association study using various open-source software packages. Simulation of a DNA sequencing experiment, analysis of the results.

Exam

Part I:
The students prepare an essay (no less than 12000 characters in length without spaces) on biological information processing from a topic list provided by the lecturer. Those who propose a topic on their own will get an extra half grade (i.e. B+ instead of B). Instead of writing an essay it is also possible to write small programs to simulate biological regulation phenomena.

Part II:
The students write a research report that summarizes their homework assignments and provide an objective overview of their results in no less than 12000 characters in length (without spaces). An additional grade (e.g. B to A, B+ to A+) will be given for the use of publicly available real measurement data instead of simulated data.

Grading will be based on the following criteria:-

- Essay and research report (70%): the students must demonstrate that they have understood the principles discussed in the lectures and can apply their knowledge in a practical context. Originality and a critical approach is especially important.
- Course activity (20%): students are required to ask questions and challenge the lecturer and each other.
- Homework assignments (10%): timely completion of the tasks with correct results is required.

Instructors' bio:

András Aszódi (born 1964) studied chemistry at Eötvös Loránd University in Budapest where he graduated in 1988. He then studied molecular neurobiology at the University of Oxford, supported by a Soros
scholarship. He received his Ph.D. in 1991 on the kinetic models of simple learning processes. From 1992 to 1996 he developed protein structure prediction methods at the National Institute for Medical Research in London. In 1996 he joined the Novartis Research Institute in Vienna as a computational modeller. He built up the In Silico Sciences unit that provided bioinformatics and computational chemistry tools to researchers. In 2006 he joined the Research Institute of Molecular Pathology in Vienna where he was developing data analysis tools and databases for high-throughput sequencing projects. He is currently teaching scientific programming and biostatistics to PhD students and postdocs. He has over 35 scientific publications, including a book with W.R. Taylor on protein structure prediction.

Peter Sarkozy (born 1984) received his degree in Computer Science from the Budapest University of Technology and Economics in 2009, and continued his graduate studies at the Department of Measurement and Information systems. During his graduate studies from 2009 to 2012 he participated in multiple projects together with the Department of Genetics, Cell and Immunobiology at the Semmelweis University. His areas of interest include the measurement and error characteristics of next-generation DNA sequencing technologies. He is the first person in Hungary to apply Oxford Nanopore Technologies’ single molecule real-time sequencing technology. He is currently working as a research assistant at the Department of Measurement and Information Systems at BUTE.