Spring Semester 2003

10.555 Bioinformatics: Principles, Methods and Applications

Instructors
Gregory Stephanopoulos[1] and Isidore Rigoutsos[2]

9 units (H), Class meets Tuesdays 2-5 pm, Room 56-154

This course provides an introduction to *Bioinformatics*. We define this field as the principles and computational methods aiming at the *upgrade* of the information content of the large volume of biological data generated by genome sequencing, as well as cell-wide measurements of gene expression (DNA microarrays), protein profiles (proteomics), metabolites and metabolic fluxes. Additionally, bioinformatics is concerned with whole organism data, especially human physiological variable measurements including organ function assessments, hormone levels, blood flow, neuronal activity etc., that characterize normal and pathophysiology. The overall goal of this data upgrade process is to elucidate cell function and physiology from a comprehensive set of measurements as opposed to using single markers of cellular function. Fundamentals from systems theory will be presented to define modeling philosophies and simulation methodologies for the integration of genomic and physiological data in the analysis of complex biological processes, e.g. genetic regulatory networks and metabolic pathways. Various computational methods will address a broad spectrum of problems in functional genomics and cell physiology, including; analysis of sequences, (alignment, homology discovery, gene annotation), gene clustering, pattern recognition/discovery in large-scale expression data, elucidation of genetic regulatory circuits, analysis of metabolic networks and signal transduction pathways. Applications of bioinformatics to metabolic engineering, drug design, and biotechnology will be also discussed.

COURSE OUTLINE

Part I: INTRODUCTION, DEFINITIONS, PRIMERS

Lecture 1: February 4

- Historical perspectives, definitions
- Impact of genomics on problems in molecular and cellular biology; need for integration and quantification, contributions of engineering
- Overview of problems to be reviewed in class: Sequence driven and data driven problems
- Overview of course methods
- Integrating cell-wide data at the cellular level
- Connection with broader issues of physiology

[1] Department of Chemical Engineering, Room 56-469, gregstep@mit.edu, 253-4583
[2] Manager, Bioinformatics & Pattern Discovery, Computational Biology Center, IBM Thomas J Watson Research Center, rigoutso@us.ibm.com

Lecture 2: February 11 (Assignment 1, due February 25)

- Primer on probabilities, inference, estimation, Bayes theorem
- Dynamic programming. Application to sequence alignment
- Markov Chains
- Hidden Markov Models

Part II: SEQUENCE DRIVEN PROBLEMS

No class on February 18 (Monday schedule-President's day)

Lecture 3: February 25 (Assignment 2, due on March 4)

- Primer on Biology (the units, the code, the process, transcription, translation, central dogma, genes, gene expression and control, replication, recombination and repair)
- Data generation and storage
- Schemes for gene finding in prokaryotes/eukaryotes
- Primer on databases on the web
- Primer on web engines

Lecture 4: March 4 (Assignment 3, due March 11)

- Some useful computer science (notation, recursion, essential algorithms on sets, trees and graphs, computational complexity)
- Physical mapping algorithms
- Fragment assembly algorithms

Lecture 5: March 11 (Assignment 4, due on March 18)

- Comparison of two sequences
- Dynamic programming revisited
- Building and using scoring matrices
- Popular algorithms: Smith-Waterman, Blast, Psi-blast, Fasta
- Multiple sequence alignment
- Functional annotation of sequences

Lecture 6: March 18 (Assignment 5, due April 1)

- Pattern discovery in biological sequences
- Protein motifs, profiles, family representations, tandem repeats, multiple sequence alignment and sequence comparison through pattern discovery
- Functional annotation

No class on March 25: Spring Break

PART III: UPGRADING EXPRESSION AND METABOLIC DATA

Lecture 7: April 1 (Assignment 6, due on April 22)

- Primer on cell physiology. Definition at the macroscopic, organism level

- Molecular cell physiology
- Interactions of pathways, cells, organs
- Measurements: molecular, cellular, clinical
- Integration of measurements, importance of kinetics
- Distribution of kinetic control among pathway steps
- Rudiments of Metabolic Control Analysis (MCA)

Lecture 8: April 8
- MCA continued
- Analysis of metabolic pathways
- Metabolic fluxes: *The metabolic phenotype*
- Methods for metabolic flux determination

No class on April 15: Patriots Day

Lecture 9: April 22 (Assignment 7, due on May 6)
- Importance of metabolic fluxes in deciphering metabolic controls
- Linking the metabolic and expression phenotypes
- Use of isotopic tracers for flux determination
- Isotopic Spectral Analysis (ISA)

Lecture 10: April 29
- Monitoring gene expression levels. Gene chips, DNA microarrays
- Data collection, error analysis, normalization and filtering
- Other novel applications of DNA microarrays
- Analysis of gene expression data
- Clustering methods: Identification of coordinated gene expression
- Identification of discriminatory genes
- Determination of gene expression patterns. Use in diagnosis
- Data visualization
- Reconstruction of gene regulatory networks

Lecture 11: May 6
- Signaling and signal transduction pathways
- Measurements in signaling networks
- Integrated analysis of signal transduction networks

Lecture 12: May 13
- Putting it all together
- Project presentations

HOMEWORKS

There will be 6-7 Problem Sets on the methodologies and computational algorithms covered in the course, as follows:

Problem Set – 1: Material of Lectures 1 and 2
Problem Set – 2: Material of Lecture 3
Problem Set – 3: Material of Lecture 4
Problem Set – 4: Material of Lecture 5
Problem Set – 5: Material of Lecture 6
Problem Set – 6: Material of lectures 7,8
Problem Set – 7: Material of lectures 9,10

PROJECTS

The students, in groups of 2-3, will carry out a project on a course-related subject of their own choosing, or from a list of suggested topics. The groups must be formed and topics selected by April 8, 2003. An oral presentation of the project by the group members will take place on May 13, 2003, at which time the final report on the project will be also due.

GRADE

There will be no mid-term or final exams. The grade in the course will be based on the homeworks, the group project, and the oral presentation, with the following weights: Homeworks (40 %); Written project report (35 %); Oral presentation (25 %)

CLASS NOTES and REFERENCES

Copies of the lecture notes will be placed on the web. Additionally, the course material will draw from published papers and the following books, recommended as references:

1. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, ISBN: 0521585198
2. *Fundamental Concepts of Bioinformatics,* D.E. Krane and M.L. Raymer, Benjamin Cummings, ISBN: 0-8053-4633-3 (2003)
3. *Introduction to Probability,* D.P. Bertsekas, and J.N. Tsitsiklis, Athena Scientific, ISBN: 1-886529-40-X (2002)
4. *Genetics, a Molecular Approach*, T.A.Brown, Chapman & Hall, ISBN: 0412447304
5. *Introduction to Computational Molecular Biology*, J.Setubal and J.Meidanis, PWS Publishing Company, ISBN: 0534952623
6. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, A.D.Baxevanis and B.F.F.Ouellette, Wiley-Interscience, ISBN: 0471191965
7. *Bioinformatics: The Machine Learning Approach*, P. Baldi and S. Brunal, MIT Press, ISBN: 0-262-02442-X
8. *Introduction to Computational Biology: Maps, Sequences, Genomes*, M.S.Waterman, Chapman & Hall, ISBN: 0412993910
9. *Biological Sequence Analysis: Probabilistic Models of proteins and Nucleic Acids*, R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Cambridge University Press, ISBN: 0-521-62041

10. *Bioinformatics: Methods and Protocols*, S. Misener and S.A. Krawetz (editors), Humana Press, ISBN: 0-89603-732-0
11. *Bioinformatics Basics: Applications in Biological Science and Medicine*, H.H. Rashidi and L.K. Buehler, CRC Press, ISBN: 0-8493-2375-4
12. Introduction to Protein Structure?, C.Branden and J.Tooze, Garland Publishing Inc., ISBN: 0815302703
13. *Molecular Biotechnology: Principles and Applications of Recombinant DNA*, B.R.Glick and J.JPasternak, ASM Press, ISBN: 1555811361
14. *Introduction to Proteins and Protein Engineering*, B.Robson and J.Garnier, Elsevier Science Publishers, ISBN: 0444810471
15. *Computational Molecular Biology: An algorithmic approach,* Pavel Pevzner, MIT Press, ISBN: 0262161974
16. *Metabolic Engineering: Principles and Methodologies,* G. Stephanopoulos, A. Aristidou and J. Nielsen, Academic Press, ISBN: 0-12-666260-6

Additional references of web-based material will be distributed to the students during the course.