Computational Genomics (0382.3102)

http://www.cs.tau.ac.il/~bchor/comp-genom.html

Prof. Benny Chor

benny@cs.tau.ac.il

Tel-Aviv University

Fall Semester, 2002-2003

AdministraTrivia

- Students' Affiliations:

AdministraTrivia

- Students' Affiliations:
  - The course is a 75% required course for 3rd year students of the bioinformatics track.

AdministraTrivia

- Students' Affiliations:
  - The course is a 75% required course for 3rd year students of the bioinformatics track.
  - It is opened to 3rd year and M.Sc. Computer Science students.

AdministraTrivia

- Students' Affiliations:
  - The course is a 75% required course for 3rd year students of the bioinformatics track.
  - It is opened to 3rd year and M.Sc. Computer Science students.
  - Students from other disciplines are encouraged to contact the instructor or the teaching assistant.

AdministraTrivia

- Students' Affiliations:
  - The course is a 75% required course for 3rd year students of the bioinformatics track.
  - It is opened to 3rd year and M.Sc. Computer Science students.
  - Students from other disciplines are encouraged to contact the instructor or the teaching assistant.
- Students who took Prof. Ron Shamir's course *Algorithms in Molecular Biology* (0368.4020.01) cannot take this course for credit.

AdministraTrivia II

- Prerequisites: A good background in algorithms, probability, and programming is required from all students.

AdministraTrivia II

- Prerequisites: A good background in algorithms, probability, and programming is required from all students.
- Grade is based on five problem sets (55-65% of total) and a project (35-45%).

AdministraTrivia II

- Prerequisites: A good background in algorithms, probability, and programming is required from all students.
- Grade is based on five problem sets (55-65% of total) and a project (35-45%).
- Projects and solutions to the problem sets to be submited in singles or pairs (no triplets, quartets, quintets etc.).

Bonuses

- Unspecified bonuses will be given for high quality contributions that will help improve future versions of this course. For example:

Bonuses

- Unspecified bonuses will be given for high quality contributions that will help improve future versions of this course. For example:
  - An original, relevant and interesting problem with a worked out solution.

Bonuses

- Unspecified bonuses will be given for high quality contributions that will help improve future versions of this course. For example:
  - An original, relevant and interesting problem with a worked out solution.
  - Original portions of lectures (written in `Prosper` LaTeX).

Bonuses

- Unspecified bonuses will be given for high quality contributions that will help improve future versions of this course. For example:
  - An original, relevant and interesting problem with a worked out solution.
  - Original portions of lectures (written in `Prosper` LaTeX).
  - Original, beautiful solutions to problems.

Problem Sets

- A total of 5 problem sets, consisting of both "dry" assignments and "wet" ones.

Problem Sets

- A total of 5 problem sets, consisting of both "dry" assignments and "wet" ones.
- The "wet" parts will require understanding and running existing software, but not writing any code.

Problem Sets

- A total of 5 problem sets, consisting of both "dry" assignments and "wet" ones.
- The "wet" parts will require understanding and running existing software, but not writing any code.
- Homework should be solved independently. External sources (books, journal articles, web pages) can be used but should be clearly quoted.

Projects Requirements

Projects are individual per group.

- They require studying a problem in depth (typically based on a research publication);

Projects Requirements

Projects are individual per group.

- They require studying a problem in depth (typically based on a research publication);
- Understanding a solution (or devising a new one), and implementing it.

Projects Requirements

Projects are individual per group.

- They require studying a problem in depth (typically based on a research publication);
- Understanding a solution (or devising a new one), and implementing it.
- Implementation will require coding a fairly large program, testing it on simulated and actual biological data, and analysing the results.

Projects Timetable

- Specification released Nov. 1st.

Projects Timetable

- Specification released Nov. 1st.
- Two page written summary of intended project – December 2nd.

Projects Timetable

- Specification released Nov. 1st.
- Two page written summary of intended project – December 2nd.
- Short interviews with each group held week of December 2nd to 9th.

Projects Timetable

- Specification released Nov. 1st.
- Two page written summary of intended project – December 2nd.
- Short interviews with each group held week of December 2nd to 9th.
- A written report, accompanied by working and documented software, due on January 5, 2003.

Staff Contact Info

- Instructor: Prof. Benny Chor,
  benny@cs.tau.ac.il,
  Office: Schreiber 223, tel.: 640-5977 .
  Office hours: By e-appointment.
- Teaching assistant: Mr. Amos Tanay,
  amos@post.tau.ac.il,
  Office: Schreiber 011, tel.: 640-5394.
  Office hours: By e-appointment.

Reference Books

- *Biological Sequence Analysis*, R. Durbin, S. Eddy A. Krogh, G. Mitchison , Cambridge University Press, 1998.

Reference Books

- *Biological Sequence Analysis*, R. Durbin, S. Eddy A. Krogh, G. Mitchison , Cambridge University Press, 1998.
- *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, 1997.

Reference Books

- *Biological Sequence Analysis*, R. Durbin, S. Eddy A. Krogh, G. Mitchison , Cambridge University Press, 1998.
- *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, 1997.
- *Post-Genome Informatics* , M. Kanehisa, Oxford University Press, 2000.

Reference Books

- *Biological Sequence Analysis*, R. Durbin, S. Eddy A. Krogh, G. Mitchison , Cambridge University Press, 1998.
- *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, 1997.
- *Post-Genome Informatics* , M. Kanehisa, Oxford University Press, 2000.
- *Introduction to Computational Molecular Biology*, J. Meidanis and J. Setubal, Brooks/Cole Pub Co., 1997.

Reference Books

- *Biological Sequence Analysis*, R. Durbin, S. Eddy A. Krogh, G. Mitchison , Cambridge University Press, 1998.
- *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, D. Gusfield, Cambridge University Press, 1997.
- *Post-Genome Informatics* , M. Kanehisa, Oxford University Press, 2000.
- *Introduction to Computational Molecular Biology*, J. Meidanis and J. Setubal, Brooks/Cole Pub Co., 1997.
- *Computational Molecular Biology : An Algorithmic Approach*, P. Pevzner, MIT Press, 2000.

Course Material On-Line

- Nir Friedman's course slides: Computatinal Molecular Biology, the Hebrew University of Jerusalem, Fall 1999.

Course Material On-Line

- Nir Friedman's course slides: Computatinal Molecular Biology, the Hebrew University of Jerusalem, Fall 1999.
- Richard Karp, Larry Ruzzo and Martin Tompa's course notes: Algorithms in Molecular Biology, University of Washington, Seattle, Winter 1998.

Course Material On-Line

- Nir Friedman's course slides: Computatinal Molecular Biology, the Hebrew University of Jerusalem, Fall 1999.
- Richard Karp, Larry Ruzzo and Martin Tompa's course notes: Algorithms in Molecular Biology, University of Washington, Seattle, Winter 1998.
- Ron Shamir's course notes: Algorithms in Molecular Biology, Tel-Aviv University, Fall 2001/2.

Course Material On-Line

- Nir Friedman's course slides: Computatinal Molecular Biology, the Hebrew University of Jerusalem, Fall 1999.
- Richard Karp, Larry Ruzzo and Martin Tompa's course notes: Algorithms in Molecular Biology, University of Washington, Seattle, Winter 1998.
- Ron Shamir's course notes: Algorithms in Molecular Biology, Tel-Aviv University, Fall 2001/2.
- Martin Tompa's course notes: Computational Biology, (CSE 527), University of Washington, Seattle, Winter 2000.

Course Material On-Line

- Nir Friedman's course slides: Computatinal Molecular Biology, the Hebrew University of Jerusalem, Fall 1999.
- Richard Karp, Larry Ruzzo and Martin Tompa's course notes: Algorithms in Molecular Biology, University of Washington, Seattle, Winter 1998.
- Ron Shamir's course notes: Algorithms in Molecular Biology, Tel-Aviv University, Fall 2001/2.
- Martin Tompa's course notes: Computational Biology, (CSE 527), University of Washington, Seattle, Winter 2000.
- And numerous other courses.

What is this course about ?

- A relatively new, multidisciplinary, and fast growing scientific area.

What is this course about ?

- A relatively new, multidisciplinary, and fast growing scientific area.
- While Computational Genomics is not in wide usage, the terms Computational Biology and BioInformatics are widely used.

What is this course about ?

- A relatively new, multidisciplinary, and fast growing scientific area.
- While Computational Genomics is not in wide usage, the terms Computational Biology and BioInformatics are widely used.
- Will attempt to *define* these terms.

Definition (take 1)

Working definitions from NIH (US National Institute of Health):

- Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Definition (take 1)

Working definitions from NIH (US National Institute of Health):

- Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.
- Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Definition (take 2)

Definitions from Hwa A. Lim page (1994):

- Bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;

Definition (take 2)

Definitions from Hwa A. Lim page (1994):

- Bioinformatics refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;
- Computational Biology encompasses the use of algorithmic tools to facilitate biological analyses.

Definition (take 3)

From Hwa A. Lim page (2001):

- Most lament that there are too many definitions of bioinformatics.

Definition (take 3)

From Hwa A. Lim page (2001):

- Most lament that there are too many definitions of bioinformatics.
- Bioinformatics will be defined differently depending on the domain of the person who is giving the definition. A computer scientist will give one definition, a biologist another, a biotechnologist yet another, and an individual from a pharmaceutical company will provide yet another definition.

Definition (take 3)

From Hwa A. Lim page (2001):

- Most lament that there are too many definitions of bioinformatics.
- Bioinformatics will be defined differently depending on the domain of the person who is giving the definition. A computer scientist will give one definition, a biologist another, a biotechnologist yet another, and an individual from a pharmaceutical company will provide yet another definition.
- Each definition is as good as the other. This is just the nature of the beast.

A Computer Scientist Perspective

- Recombinant DNA technology has created a revolution in Molecular Biology in the last decade.

A Computer Scientist Perspective

- Recombinant DNA technology has created a revolution in Molecular Biology in the last decade.
- New computational problems arise from large genome projects and novel high throughput biotechnologies.

A Computer Scientist Perspective

- Recombinant DNA technology has created a revolution in Molecular Biology in the last decade.
- New computational problems arise from large genome projects and novel high throughput biotechnologies.
- Problems involve collection, assembly, organization and interpretation of genetic sequence data.

A Computer Scientist Perspective

- Recombinant DNA technology has created a revolution in Molecular Biology in the last decade.
- New computational problems arise from large genome projects and novel high throughput biotechnologies.
- Problems involve collection, assembly, organization and interpretation of genetic sequence data.
- Novel algorithmic, mathematical and statistical tools are crucial for analyzing this flow of information and discovering new global structures in it.

Course Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.

Course Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.
- Constructing phylogenetic (evolutionary) trees from sequence data.

Course Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.
- Constructing phylogenetic (evolutionary) trees from sequence data.
- Probabilistic models for classification and analysis of sequence data, *e.g.* for gene finding.

Course Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.
- Constructing phylogenetic (evolutionary) trees from sequence data.
- Probabilistic models for classification and analysis of sequence data, *e.g.* for gene finding.
- Finding regulatory motifs in DNA sequences.

Structural BioInformatics

- Deals mainly with the interplay between proteins' 3-dimensional structure and function, and their relation to designing new medicines.

Structural BioInformatics

- Deals mainly with the interplay between proteins' 3-dimensional structure and function, and their relation to designing new medicines.
- Apply many tools from computer vision and computational geometry.

Structural BioInformatics

- Deals mainly with the interplay between proteins' 3-dimensional structure and function, and their relation to designing new medicines.
- Apply many tools from computer vision and computational geometry.
- Not covered at all in this course.

Structural BioInformatics

- Deals mainly with the interplay between proteins' 3-dimensional structure and function, and their relation to designing new medicines.
- Apply many tools from computer vision and computational geometry.
- Not covered at all in this course.
- Will be given by Prof. Haim Wolfson on the spring semester, 2003.

Molecular Biology Background

- Two important *linear* molecules: DNA and Proteins $\Longrightarrow$ strings over 4- and 20-letter alphabets respectively

Molecular Biology Background

- Two important *linear* molecules: DNA and Proteins $\Longrightarrow$ strings over 4- and 20-letter alphabets respectively
- Specific genes, substrings of DNA, code for specific proteins

Molecular Biology Background

- Two important *linear* molecules: DNA and Proteins $\Longrightarrow$ strings over 4- and 20-letter alphabets respectively
- Specific genes, substrings of DNA, code for specific proteins
- Protein sequence influences structure, which in turn determines its function

Molecular Biology Background

- Two important *linear* molecules: DNA and Proteins $\Longrightarrow$ strings over 4- and 20-letter alphabets respectively
- Specific genes, substrings of DNA, code for specific proteins
- Protein sequence influences structure, which in turn determines its function

Moral: Study of similarity in sequence, structure and function of biological strings gives clues to further discovery

Evolution

- Biological systems evolved over time from simpler to more complex organisms

Evolution

- Biological systems evolved over time from simpler to more complex organisms
- History of evolution gives key clues to important changes and improvements in biological function

Evolution

- Biological systems evolved over time from simpler to more complex organisms
- History of evolution gives key clues to important changes and improvements in biological function

Moral: Evolutionary history gives important leads to futher discovery

And Now

To a short tour of some relevant biology.