Workshop in BioInformatics (0382.3102)

Prof. Benny Chor
benny@cs.tau.ac.il

Tel-Aviv University

Spring Semester, 2005
Preliminaries

The course is a required course for students of the bioinformatics track, and is offered to 3rd year Computer Science students as well. Students from other disciplines should consult the instructor.
Preliminaries

- The course is a required course for students of the bioinformatics track, and is offered to 3rd year Computer Science students as well. Students from other disciplines should consult the instructor.

- The course was also given in the fall semester (by another instructor), and is expected (but not guaranteed) to be given next fall semester.
Preliminaries

- The course is a required course for students of the **bioinformatics track**, and is offered to 3rd year Computer Science students as well. Students from other disciplines should consult the instructor.

- The course was also given in the fall semester (by another instructor), and is expected (but not guaranteed) to be given next fall semester.

- Some biological background knowledge is useful, but is not absolutely necessary.
AdministraTrivia

Grade is based on a software project implementation (55-65% of total), and on an outline (10 min.) and a presentation (30 min.) of the project (35-45%).
AdministraTrivia

- Grade is based on a software project implementation (55-65% of total), and on an outline (10 min.) and a presentation (30 min.) of the project (35-45%).

- Projects to be done in groups of size 1 or 2.
AdministraTrivia

- Grade is based on a software project implementation (55-65% of total), and on an outline (10 min.) and a presentation (30 min.) of the project (35-45%).
- Projects to be done in groups of size 1 or 2.
- Outlines presented on March 28.
AdministraTrivia

- Grade is based on a software project implementation (55-65% of total), and on an outline (10 min.) and a presentation (30 min.) of the project (35-45%).
- Projects to be done in groups of size 1 or 2.
- Outlines presented on March 28.
- Projects’ presentations will take place during the last 2-3 weeks of the semester.
Presentations and outlines should use computerized tools (prosper LaTeX, power-point, or any other software of your choice).
AdministraTrivia (2)

- Presentations and outlines should use computerized tools (prosper LaTeX, power-point, or any other software of your choice).

- In addition, there will be two-three lectures on various relevant topics in software engineering, given by the Computer Science system staff.
Presentations and outlines should use computerized tools (prosper LaTeX, power-point, or any other software of your choice).

In addition, there will be two-three lectures on various relevant topics in software engineering, given by the Computer Science system staff.

Physical attendance in all presentations and lectures is mandatory.
Projects’ Descriptions

Analysis, design and implementation of combinatorial optimization algorithms with bioinformatics relevance.
Projects’ Descriptions

- Analysis, design and implementation of combinatorial optimization algorithms with bioinformatics relevance.

- Some contemporary problems in comparative genomics, DNA chips analysis, phylogenetic analysis, regulatory motives finding, and more.
Projects’ Descriptions

- Analysis, design and implementation of combinatorial optimization algorithms with bioinformatics relevance.
- Some contemporary problems in comparative genomics, DNA chips analysis, phylogenetic analysis, regulatory motives finding, and more.
- Getting acquainted with publicly available Bioinformatics databases and using them.
Projects’ Descriptions

- Analysis, design and implementation of combinatorial optimization algorithms with bioinformatics relevance.
- Some contemporary problems in comparative genomics, DNA chips analysis, phylogenetic analysis, regulatory motives finding, and more.
- Getting acquainted with publicly available Bioinformatics databases and using them.
- Conducting supervised research in computational biology.
Projects’ Descriptions

- Analysis, design and implementation of combinatorial optimization algorithms with bioinformatics relevance.

- Some contemporary problems in comparative genomics, DNA chips analysis, phylogenetic analysis, regulatory motives finding, and more.

- Getting acquainted with publicly available Bioinformatics databases and using them.

- Conducting supervised research in computational biology.

- Efficient implementation of algorithms in C, C++, Java or Matlab (if you insist, we will also consider cobol or even scheme).
Projects’ Requirements

Projects are **individual** per group.

- They require studying a problem in depth (typically based on research publications);
Projects’ Requirements

Projects are *individual* per group.

- They require studying a problem in depth (typically based on research publications);
- 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 research publications);
- 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.
Tentative TimeTable

Specification released February 28th.
Tentative TimeTable

- Specification released February 28th.
- Choices sent to staff by March 6th.
Tentative TimeTable

- Specification released February 28th.
- Choices sent to staff by March 6th.
- Two page written summary of intended project sent to staff no later than March 20th.
Tentative TimeTable

- Specification released **February 28th**.
- Choices sent to staff by **March 6th**.
- Two page written summary of intended project sent to staff no later than **March 20th**.
- Short interviews with each group held week of **March 20th to 27th**.
Tentative TimeTable

- Specification released February 28th.
- Choices sent to staff by March 6th.
- Two page written summary of intended project sent to staff no later than March 20th.
- Short interviews with each group held week of March 20th to 27th.
- Outlines presented on March 28.
Tentative TimeTable

- Specification released **February 28th**.
- Choices sent to staff by **March 6th**.
- Two page written summary of intended project sent to staff no later than **March 20th**.
- Short interviews with each group held week of **March 20th to 27th**.
- Outlines presented on **March 28**.
- Projects’ presentations during the **last 2-3 weeks** of the semester.
Tentative TimeTable

- Specification released February 28th.
- Choices sent to staff by March 6th.
- Two page written summary of intended project sent to staff no later than March 20th.
- Short interviews with each group held week of March 20th to 27th.
- Outlines presented on March 28.
- Projects’ presentations during the last 2-3 weeks of the semester.
- A written report, accompanied by working and documented software, due on June 1st, 2005.
Important Remark

- Projects intended at mini research.
- As such, they may chart some unexplored territories.
- Last year, one of the projects has led to a paper accepted to a prestigious and highly competitive conference.
- This year, we’d like to see more such results.
- However, research has risks – not every attempt will succeed.
- You may still get a high grade in the workshop even if your project was not a success (research wise)!
The BioTechnology Revolution

- Biological sciences have gone through a revolution in the last decade.
The BioTechnology Revolution

- Biological sciences have gone through a revolution in the last decade.
- To a large extent, this revolution was driven by advances in biotechnology.
The BioTechnology Revolution

- Biological sciences have gone through a revolution in the last decade.
- To a large extent, this revolution was driven by advances in biotechnology.
- One of the better known results of this revolution is the sequencing of the human genome.
The BioTechnology Revolution

- Biological sciences have gone through a revolution in the last decade.
- To a large extent, this revolution was driven by advances in biotechnology.
- One of the better known results of this revolution is the sequencing of the human genome.
- Also sequenced: Genomes of about two hundred other organisms (mouse, rice, fruit fly – Drosophila, worm – C. Elegans, mosquito – Anopheles, malaria, bacteria – E. Coli, ...), and thousands viruses.
The BioTechnology Revolution

Robotics’ sequencers
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
- mass spectrometry
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
- mass spectrometry
- SNPs genotyping
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
- mass spectrometry
- SNPs genotyping
- and very many other biotechnologies
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
- mass spectrometry
- SNPs genotyping
- and very many other biotechnologies
- produce massive amounts of data.
The BioTechnology Revolution

- Robotics’ sequencers
- DNA microarrays
- 2D gels
- mass spectrometry
- SNPs genotyping
- and very many other biotechnologies
- produce massive amounts of data.
- The task of analyzing, interpreting, and understanding this data is where bioinformatics comes in.
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.
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.
Important BioInfo Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.
Important BioInfo Topics

Algorithms and heuristics motivated by problems originating from molecular biology.

- Sequence comparison and alignment.
- Constructing phylogenetic (evolutionary) trees from sequence data.
Important BioInfo 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**.
Important BioInfo 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.
Genomics and Proteomics

Attempts to understand function and interactions between families of genes and proteins within the cell, tissue, or organism.
Genomics and Proteomics

- Attempts to understand function and interactions between families of genes and proteins within the cell, tissue, or organism.

- Quite recently, new rolls of small non-coding RNAs were discovered (Science magazine *discovery of the year*, 2002).
Molecular Biology Background

- Two important linear molecules: DNA and Proteins: Strings over 4- and 20-letter alphabets, respectively.
Molecular Biology Background

- Two important linear molecules: DNA and Proteins: 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: 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: 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 further discovery
And Now

To a short tour of some specific topics and problems.
Suggested Topics

Employing string operators, influenced by information theoretic tools, for gene finding.
Suggested Topics

- Employing string operators, influenced by information theoretic tools, for gene finding.

- “System Biology": Employing linear and probabilistic models to infer genetic networks, based on gene expression datasets.
Suggested Topics

- Employing string operators, influenced by information theoretic tools, for gene finding.
- “System Biology": Employing linear and probabilistic models to infer genetic networks, based on gene expression datasets.
- Finding highly conserved segments among pairs and triplets of genome sequences.
Suggested Topics

- Employing string operators, influenced by information theoretic tools, for gene finding.
- “System Biology": Employing linear and probabilistic models to infer genetic networks, based on gene expression datasets.
- Finding highly conserved segments among pairs and triplets of genome sequences.
- Finding common and separating properties of regulatory and metabolic networks over different species.
Suggested Topics

- Employing string operators, influenced by information theoretic tools, for gene finding.

- “System Biology": Employing linear and probabilistic models to infer genetic networks, based on gene expression datasets.

- Finding highly conserved segments among pairs and triplets of genome sequences.

- Finding common and separating properties of regulatory and metabolic networks over different species.

- Testing the hypothesis that there is a correlation between proximity of genes (on the chromosome) and their interaction.
Suggested Topics (cont.)

Finding regulatory motifs, employing an existing system (MeX) together with gene expression datasets and regions conserved across different genomes.
Suggested Topics (cont.)

- Finding regulatory motifs, employing an existing system (MeX) together with gene expression datasets and regions conserved across different genomes.

- Finding associations among cancer related genes, using an existing tool for detecting linear separability of genes, and gene expression datasets.
Suggested Topics (cont.)

- Finding regulatory motifs, employing an existing system (MeX) together with gene expression datasets and regions conserved across different genomes.

- Finding associations among cancer related genes, using an existing tool for detecting linear separability of genes, and gene expression datasets.

- Exploring properties of the likelihood function in phylogenetic (evolutionary) trees, for simulated and real sequences.
Suggested Topics (cont.)

- Finding **regulatory motifs**, employing an existing system (MeX) together with gene expression datasets and regions conserved across different genomes.


- Exploring properties of the **likelihood function** in phylogenetic (evolutionary) trees, for simulated and real sequences.

- Other topics **you** would like to explore (after discussing them with us and getting our approval).