CS2220: Introduction to Computational Biology
2010
Course Briefing

Lisa Tucker-Kellogg
Co-teaching with Limsoon Wong

Recommended “Pre-requisites”

- CS1102: Data Structures and Algorithms
- LSM1102: Molecular Genetics
Objectives

• Develop flexible and logical problem solving skill
• Understand bioinformatics problems
• Appreciate techniques and approaches to bioinformatics

To achieve the goals above, we expose students to a series of case studies spanning gene feature recognition, gene expression and proteomic analysis, gene finding, sequence homology interpretation, phylogeny analysis, etc.

Contents of Course Overview

• Time Table
• Course Syllabus
• Course Homepage
• Teaching Style
• Project, Assignments, Exams
• Readings
• Assessment

• Quick Overview of Themes and Applications of Bioinformatics
Time Table

- **Lecture**
  - Thursday 2:00pm – 4:00pm, COM1-211
- **Tutorial**
  - Thursday 4:00pm – 5:00pm, COM1-207
- **Email**
  - wongls@comp.nus.edu.sg
  - tucker@comp.nus.edu.sg
- **Consultations**
  - Available for walk-in at COM1, Level 3.
    - Room 34 (Prof. Wong)
    - Room 24 (Dr. Lisa) -- Call 6516-2865 to see if I’m there.

Course Syllabus

- **Intro to Bioinformatics**
- **Essence of Knowledge Discovery**
  - Classification performance measures
  - Feature selection techniques
  - Supervised & unsupervised machine learning techniques
- **Gene Feature Recognition from Genomic DNA**
  - Feature generation, selection, & integration
  - Translation initiation site (TIS) recognition
  - Transcription start site (TSS) recognition
- **Gene Expression and Proteome Analysis**
  - Microarray and mass-spec basics
  - Classification of gene expression profiles
  - Classification of proteomic profiles
  - Clustering of gene expression profiles
  - Molecular network reconstruction
- **Essence of Seq Comparison**
  - Dynamic programming basics
  - Sequence comparison and alignment basics
  - Needleman-Wunsch global alignment algorithm
  - Smith-Waterman local alignment algorithm
- **Protein Seq, Structure, and Applications of Optimization**
  - Homology modeling and molecular modeling
  - Essence of optimization
  - Active site and domain prediction
- **Seq Interpretation**
  - Key mutation sites prediction
  - Protein function prediction by sequence alignment
  - Protein function prediction by phylogenetic profiling
- **Gene Finding**
  - Overview of gene finding
  - GRAIL
  - Handling of frame shifts and in-dels
- **Some hot current topics like PPI, miRNA, etc.**
Course Homepage

• IVLE

• Lecture Slides & etc
  – http://www.comp.nus.edu.sg/~wongls/courses/cs2220/2009  (From Last Year)

Teaching Style

• Bioinformatics is a broad area

• Need to learn a lot of material by yourself
  – Reading books
  – Reading papers
  – Practice on the web

• Don’t expect to be told everything
Assignments, Project, & Exam

• Assignments
  – Probably 3-4 assignments
  – Some are simple programming assignments

• Project
  – Based on a case study in the class
  – 8-10 pages of report / ppt slides expected

• Exam
  – 1 final open-book exam

Be Honest

• Exam
  – Absence w/o good cause results in ZERO mark
  – Cheating results in ZERO mark

• Discussion on assignments is allowed
• Blatant plagiarism is not allowed
  – Offender gets ZERO mark for assignment or exam
  – Penalty applies to those who copied AND those who allowed their assignments to be copied

• Other cultures are far more punitive about cheating
Background Readings

• Limsoon Wong, *The Practical Bioinformatician*, WSPC, 2004
• Rick Ng, *Drugs: From Discovery to Approval* 2009

Assessment

• Continuous Assessment: 50%
• Final Exam: 50%
What comes after CS2220

- **CS2220 Introduction to Computational Biology**
  - Understand bioinformatics problems; interpretational skills

- **CS3225 Combinatorial Methods in Bioinformatics**

- **CS4220 Knowledge Discovery Methods in Bioinformatics**
  - Clustering; classification; association rules; SVM; HMM; Mining of seq. trees, & graphs

- **CS5238 Advanced Combinatorial Methods in Bioinformatics**
  - Seq alignment, whole-genome alignment, suffix tree, seq indexing, motif finding, RNA sec struct prediction, phylogeny reconstruction

- **CS6280 Computational Systems Biology**
  - Dynamics of biochemical and signaling networks; modeling, simulating, & analyzing them
  - Etc ...

Any questions?

I hope you will enjoy this class 😊
Themes and Applications of Bioinformatics

What is Bioinformatics?

Copyright 2010 © Limsoon Wong, Lisa Tucker-Kellogg
Themes of Bioinformatics
Themes of This Course

Bioinformatics involves
  Data Mgmt +
  Knowledge Discovery +
  Sequence Analysis +
  Physical Modeling + ...

Knowledge Discovery =
  Statistics + Algorithms + Databases

The Promises of Bioinformatics

To the patient:
  Better drug, better treatment

To the pharma:
  Save time, save cost, make more $

To the scientist:
  Better science
Fulfilling the Promise via Drugs

- Bioinformatics is applicable to many phases of drug development
- Drug discovery: Design small molecules that bind target proteins.
  - Which proteins? What should binding accomplish?
- Biomarkers
- Personalized Medicine (future)
- There have been some disappointments too.

Figure from Rick Ng, *Drugs: From Discovery to Approval*

Pervasiveness of Bioinformatics

For large-scale biology
e.g., High-throughput, massively-parallel measurements, or “lab on a chip” miniaturization.

Bioinformatics is mandatory

For indirect experimental methods
e.g., reconstruction based on phase contrast or wave diffraction.

Computational data analysis is mandatory

What about the rest of biology (and medicine)?

Limitless opportunities exist, but computational methods are in competition with non-computational methods, such as human intuition.
Some Bioinformatics Problems

• Biological Data Searching
• Biological Data Integration
• Gene/Promoter finding
• Cis-regulatory DNA
• Gene/Protein Network
• Protein/RNA Structure Prediction
• Evolutionary Tree reconstruction

• Infer Protein Function
• Disease Diagnosis
• Disease Prognosis
• Disease Treatment Optimization, ...

Biological Data Searching

• Biological Data is increasing rapidly

• Biologists need to locate required info

• Difficulties:
  – Too much
  – Too heterogeneous
  – Too distributed
  – Too many errors
  – Need approximate searches because of errors, mutations, etc.

Image credit: NCBI
Biological Data Integration

• In the “post-Genbank” era, most bioinformatics data is about experiments.
  • Some datasets are useless without reference points. (Measurements relative to control, or requiring normalization.)
  • Some datasets have little meaning without their biological context (Cell type, disease state, treatment conditions, what was selected and what was discarded.)
  • To benefit from data often requires a huge amount of “meta-data” about the samples and methods.

METADATA becomes mandatory

Semantics & Ontologies

“Artificial Intelligence meets Databases”

Cis-Regulatory DNAs

• Cis-regulatory DNAs control whether genes should express or not.
• Cis-regulatory DNAs may locate in promoter region, intron, or exon.
• Finding and understanding cis-regulatory DNAs is one of the key problem in coming years.
Gene Networks

- Inside a cell is a complex system
- Expression of one gene depends on expression of another gene
- Such interactions can be represented using gene network
- Understanding such networks helps identify association between genes and diseases

Protein/RNA structure prediction

- Structure of Protein/RNA is essential to its functionality
- Important to have some ways to predict the structure of a protein/RNA given its sequence
- This problem is important and it is always considered as a “grand challenge” problem in bioinformatics
**Evolutionary Tree Reconstruction**

- Protein/RNA/DNA mutates
- Evolutionary Tree studies evolutionary relationship among set of protein/RNA/DNAs
- Figures out origin of species

**Breast Cancer Outcome Prediction**

- Van't Veer et al., *Nature* 415:530-536, 2002
- Training set contains 78 patient samples
  - 34 patients develop distance metastases in 5 yrs
  - 44 patients remain healthy from the disease after initial diagnosis for >5 yrs
- Testing set contains 12 relapse & 7 non-relapse samples
Commonly Used Data Sources

Type of Biological Databases

- **Micro Level**
  - Contain info on the composition of DNA, RNA, Protein Sequences

- **Macro Level**
  - Contain info on interactions
    - Gene Expression
    - Metabolites
    - Protein-Protein Interaction
    - Biological Network

- **Metadata**
  - Ontology
  - Literature

Exercise: Name a protein seq db and a DNA seq db

Copyright 2010 © Limsoon Wong, Lisa Tucker-Kellogg
Transcriptome Database

- Complete collection of all possible mRNAs (including splice variants) of an organism
- Transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes
- Regions of an organism’s genome that get transcribed into messenger RNA

Exercise: Name a transcriptome database

Gene Expression Databases

- Detect what genes are being expressed or found in a cell of a tissue sample
- Single-gene analysis
  - Northern Blot
  - In Situ Hybridization
  - RT-PCR
- Many Genes: High Throughput Arrays
  - cDNA Microarray
  - Affymetrix GeneChip® Microarray

Exercise: Name a gene expression database
Metabolites Database

- A metabolite is an organic compound that is a starting material in, an intermediate in, or an end product of metabolism.

- Metabolites dataset are also generated from mass spectrometry which measure the mass the these simple molecules, thus allowing us to estimate what are the metabolites in a tissue.

- Starting metabolites:
  - Small, of simple structure, absorbed by the organism as food
  - E.g., vitamins and amino acids

- Intermediary metabolites:
  - The most common metabolites
  - May be synthesized from other metabolites, or broken down into simpler compounds, often with the release of chemical energy
  - E.g., glucose

- End products of metabolism:
  - Final result of the breakdown of other metabolites
  - Excreted from the organism without further change
  - E.g., urea, carbon dioxide

Protein-Protein Interaction Databases

- Proteins are true workhorses
  - Lots of the cell’s activities are performed thru PPI including message passing, gene regulation, etc.

- Function of a protein also depends on proteins it interact with

- Methods for generating PPI database include:
  - Biochemical purifications, yeast-two hydrid, synthetic lethals, in silico predictions, mRNA-co-expression

- Contain many false positives & false negatives

Exercise: Name a PPI database
Any Question?

Acknowledgements

• Most of the slides used in this lecture are based on original slides created by
  – Ken Sung
  – Anthony Tung
  – Limsoon Wong
• But you should blame me for any errors
References

• S.K. Ng, “Molecular Biology for the Practical Bioinformatician”, *The Practical Bioinformatician*, Chapter 1, pages 1—30, WSPC, 2004


• Lots of useful videos, http://www.as.wvu.edu/~dray/Bio_219.html