CS2220: Intro to Computational Biology
Course Briefing
Limsoon Wong

Recommended “Pre-requisites”

• CS1020 Data Structures and Algorithms I
• CS2020 Data Structures and Algorithms II
• LSM1101 Biochemistry and Biomolecules
• LSM1102 Molecular Genetics

Objectives

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

To achieve goals above, we expose students to 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 9am – 11am, SR@LT19
• Tutorial
  – Monday 1pm – 2pm, COM1-217
• Email
  – wongls@comp.nus.edu.sg
• Consultations
  – Any time; just make appt to make sure I am in
  – Pls email my PA, tanps@comp.nus.edu.sg

Course Syllabus

• Intro to Bioinformatics
  – molecular biology basics
  – tools and instruments for molecular biology
  – themes and applications of bioinformatics

• Essence of Knowledge Discovery
  – Classification performance measures
  – Feature selection techniques
  – Support vector 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 Analysis
  – Microarray analysis
  – Clustering of gene expression profiles
  – Change of gene expression profiles

• Gene Finding
  – Overview of gene finding
  – G R A I L
  – Handling of frame shifts and in-dels

• Phylogenetic Trees
  – Phylogeny reconstruction
  – Origin of Polynesians & Europeans
  – Large-scale sequencing basics

• Some hot current topics like PPI, miRNA, etc.
Course Homepage

- IVLE
- Lecture Slides & etc

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 (35% of marks)
  - 3 assignments
  - Some are simple programming assignments
- Project (15% of marks)
  - Based on a case study in the class
  - 8-10 pages of report / ppt slides expected
- Exam (50% of marks)
  - 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

Background Readings

- Limsoon Wong, The Practical Bioinformatician, WSPC, 2004
- Marketa Zvelebil and Jeremy Baum, Understanding Bioinformatics, Garland, 2007

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 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?

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 drug development
- Drug discovery: Design small molecules that bind target proteins
  - Which proteins?
  - What should binding accomplish?
- Biomarkers
Pervasiveness of Bioinformatics

• Bioinformatics is mandatory for large-scale biology
  – e.g., High-throughput, massively-parallel measurements, or “lab on a chip” miniaturization
• Computational data analysis is mandatory for indirect experimental methods
  – e.g., reconstruction based on phase contrast or wave diffraction.
• What about the rest of biology (and medicine) ?
  • Limitless opportunities!

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.

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 & understanding cis-regulatory DNAs is one of the key problem in coming years

Gene Networks

• Cell is a complex system
• Expression of one gene depends on expression of another gene
• Such interactions can be form gene network
• Understanding such networks helps identify association betw genes & diseases

Protein/RNA Structure Prediction

• Structure of Protein / RNA is essential to its functionality
• Impt to predict structure of a protein / RNA given its seq
• Problem is considered a “grand challenge” problem in bioinformatics
Evolutionary Tree Reconstruction

- Protein /RNA / DNA mutates
- Evolutionary tree studies evolutionary relationship among set of protein / RNA / DNAs
- Origin of species

Copyright 2011 © Limsoon Wong

Breast Cancer Outcome Prediction

- Van't Veer et al., Nature 415:530-536, 2002
- Training set contains 78 patient samples
  - 34 patients develop distant 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

Copyright 2011 © Limsoon Wong

Type of Biological Databases

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

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

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

Copyright 2011 © Limsoon Wong

Commonly Used Data Sources

- **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

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

Copyright 2011 © Limsoon Wong

Transcriptome Database

- Complete collection of all possible mRNAs (including splice variants) of an organism
- Regions of an organism’s genome that get transcribed into messenger RNA

Exercise: Name a transcriptome database

Copyright 2011 © Limsoon Wong

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

Copyright 2011 © Limsoon Wong
Metabolites Database

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

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

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.

Protein-Protein Interaction Databases

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

Function of a protein depends on proteins it interacts with
- Methods for generating PPI db
  - Biochemical purifications, Y2H, synthetic lethals, in silico predictions, mRNA-co-expression

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

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