

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

Copyright 2011 © Limsoon Wong.

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.

Copyright 2011 © Limsoon Wong.

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

Copyright 2011 © Limsoon Wong.

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

Copyright 2011 © Limsoon Wong.

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
 - 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 Analysis**
 - Microarray basics
 - Gene expression profile normalization
 - Classification of gene expression profiles
 - Clustering of gene expression profiles
 - Molecular network reconstruction
- **Essence of Seq Comparison**
 - Dynamic programming basics
 - Sequence comparison and alignment basics
 - Needleman-Wunsh global alignment algorithm
 - Smith-Waterman local alignment algorithm
- **Seq Homology Interpretation**
 - protein function prediction by sequence alignment
 - protein function prediction by phylogenetic profiling
 - active site and domain prediction
 - key mutation sites prediction
- **Gene Finding**
 - Overview of gene finding
 - GRAIL
 - Handling of frame shifts and in-dels
- **Phylogenetic Trees**
 - Phylogeny reconstruction method basics
 - origin of Polynesians & Europeans
 - Large-scale sequencing basics
- **Some hot current topics like PPI, miRNA, etc.**

Copyright 2011 © Limsoon Wong.

7

Course Homepage



- **IVLE**
 - https://ivle.nus.edu.sg/lms/public/list_course_public.aspx?code=cs2220&acadyear=2011%2f2012
- **Lecture Slides & etc**
 - <http://www.comp.nus.edu.sg/~wongls/courses/cs2220/2011>

Copyright 2011 © Limsoon Wong.

8

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**

Copyright 2011 © Limsoon Wong.

9

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

Copyright 2011 © Limsoon Wong.

10

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

Copyright 2011 © Limsoon Wong.

11

Background Readings




- Limsoon Wong, *The Practical Bioinformatician*, WSPC, 2004
- Wing-Kin Sung, *Algorithms in Bioinformatics: A Practical Introduction*, CRC, 2010
- Marketa Zvelebil and Jeremy Baum, *Understanding Bioinformatics*, Garland, 2007

Copyright 2011 © Limsoon Wong.

12

What comes after CS2220



<ul style="list-style-type: none"> • CS2220 Introduction to Computational Biology <ul style="list-style-type: none"> – Understand bioinformatics problems; interpretational skills • CS3225 Combinatorial Methods in Bioinformatics • CS4220 Knowledge Discovery Methods in Bioinformatics <ul style="list-style-type: none"> – Clustering; classification; association rules; SVM; HMM; Mining of seq, trees, & graphs 	<ul style="list-style-type: none"> • CS5238 Advanced Combinatorial Methods in Bioinformatics <ul style="list-style-type: none"> – Seq alignment, whole-genome alignment, suffix tree, seq indexing, motif finding, RNA structure prediction, phylogeny reconstruction • CS6280 Computational Systems Biology <ul style="list-style-type: none"> – Dynamics of biochemical and signaling networks; modeling, simulating, & analyzing them • Etc ...
---	---

Copyright 2011 © Limsoon Wong.

13

Any questions?

I hope you will enjoy this class ☺

NUS

Copyright 2011 © Limsoon Wong.

Themes and Applications of
Bioinformatics

NUS
National University
of Singapore

15

What is Bioinformatics?

NUS

Copyright 2011 © Limsoon Wong.

16

Themes of Bioinformatics
Themes of This Course

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

Knowledge Discovery =
Statistics + Algorithms + Databases

NUS

Copyright 2011 © Limsoon Wong.

17

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

NUS

Copyright 2011 © Limsoon Wong.

18

Fulfilling the
Promise via
Drugs

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

- **Bioinformatics is applicable to drug development**
- **Drug discovery: Design small molecules that bind target proteins**
 - Which proteins?
 - What should binding accomplish?
- **Biomarkers**

NUS

Copyright 2011 © Limsoon Wong.

19

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!**

Copyright 2011 © Limsoon Wong.

20

Some Bioinformatics Problems

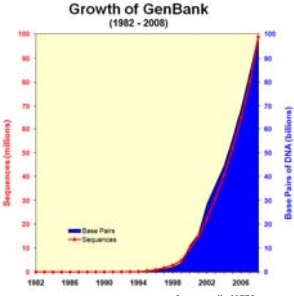
- **Biological Data Searching**
- **Biological Data Integration**
- **Gene/Promoter finding**
- **Cis-regulatory DNA**
- **Gene/Protein Network**
- **Protein/RNA Structure Prediction**
- **Evolutionary Tree reconstruction**
- **Infer Protein Function**
- **Disease Diagnosis**
- **Disease Prognosis**
- **Disease Treatment Optimization, ...**

Copyright 2011 © Limsoon Wong.

21

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.



Growth of GenBank (1982 - 2006)

The graph shows two exponential growth curves from 1982 to 2006. The left y-axis is 'Sequences (millions)' and the right y-axis is 'Base Pairs of DNA (billions)'. The x-axis shows years from 1982 to 2006. A legend indicates 'Base Pairs' (blue area) and 'Sequences' (red line).

Image credit: NCBI

Copyright 2011 © Limsoon Wong.

22

Cis-Regulatory DNAs

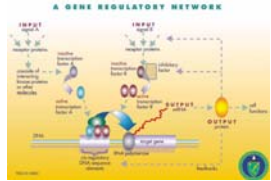


Image credit: US DOE

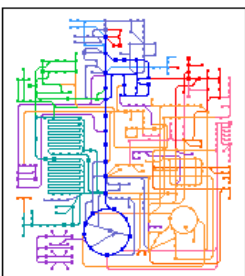
- **Cis-regulatory DNAs control whether genes should express or not**
- **Cis-regulatory DNAs may locate in promoter region, intron, or exon**
- **Finding & understanding cis-regulatory DNAs is one of the key problem in coming years**

Copyright 2011 © Limsoon Wong.

23

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**



Copyright 2011 © Limsoon Wong.

24

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**

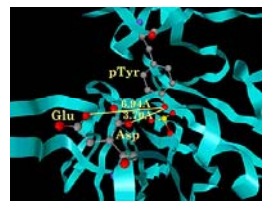


Image credit: Kollatkar

Copyright 2011 © Limsoon Wong.

25

Evolutionary Tree Reconstruction

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

Image credit: Sykes
Copyright 2011 © Limsoon Wong.

26

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

Image credit: Veer
Copyright 2011 © Limsoon Wong.

Commonly Used Data Sources

Copyright 2011 © Limsoon Wong.

28

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 2011 © Limsoon Wong.

29

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
- Transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes

Exercise: Name a transcriptome database

Copyright 2011 © Limsoon Wong.

30

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.

31

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 of 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

Copyright 2011 © Limsoon Wong.

32

Protein-Protein Interaction Databases



- **Proteins are true workhorses**
 - Lots of cell's activities are performed thru PPI, e.g., message passing, gene regulation, etc.
- **Methods for generating PPI db**
 - biochemical purifications, Y2H, synthetic lethals, in silico predictions, mRNA-co-expression
- **Function of a protein depends on proteins it interacts with**
- **Contain many false positives & false negatives**

Exercise: Name a PPI database


Copyright 2011 © Limsoon Wong.

Any Question?



34

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**

Copyright 2011 © Limsoon Wong.

35

References



- S.K. Ng, "Molecular Biology for the Practical Bioinformatician", *The Practical Bioinformatician*, Chapter 1, pages 1-30, WSPC, 2004
- DOE HGP Primer, http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/index.shtml
- Lots of useful videos, http://www.as.wvu.edu/~dray/Bio_219.html

Copyright 2011 © Limsoon Wong.