

CS2220: Intro to Computational Biology

Course Briefing

Wong Limsoon



Recommended “pre-requisites”



- **Data Structures and Algorithms**
- **Basic statistics**
- **Biochemistry of Biomolecules**
- **Molecular Genetics**

You may also find the following to be a useful hands-on complement to CS2220:

- **LSM2241 Introductory Bioinformatics**

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, sequence homology interpretation, phylogeny analysis, etc.**

Contents of course overview



- **Time table**
- **Course syllabus**
- **Course homepage**
- **Teaching style**
- **Project, assignments, exams**
- **Readings**
- **Assessment**

Time table

- **Lecture**
 - Wednesday 3pm – 6pm, by Zoom
- **Tutorial**
 - Integrated into each lecture
- **Email**
 - wongls@comp.nus.edu.sg
- **Consultations**
 - Any time; just make appt

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

- **Phylogenetic Trees**

- Phylogeny reconstruction method basics
- origin of Polynesians & Europeans
- Large-scale sequencing basics

- **One or two other topics (drug-resistant mutation prediction, ortholog prediction, disease-causing mutations, etc.)**

Course homepage

- **LumiNUS**
 - <https://luminus.nus.edu.sg/modules/323d85da-ec20-43fb-95bb-3793b43940be>
- **Lecture Slides & etc.**
 - <http://www.comp.nus.edu.sg/~wongls/courses/cs220/2020>

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 simple programming required
- **Project (15% of marks)**
 - Based on material associated with self-learning
 - 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

- **References are provided in the lecture slides; please look them up**
- **You can also read these or other books for background in general**
 - Limsoon Wong, *The Practical Bioinformatician*, WSPC, 2004
 - Wing-Kin Sung, *Algorithms in Bioinformatics: A Practical Introduction*, CRC, 2010

What comes after CS2220

- **CS2220 Introduction to Computational Biology**
 - Understand bioinformatics problems; interpretational skills
- **CS4330 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
- Etc ...

Any questions?

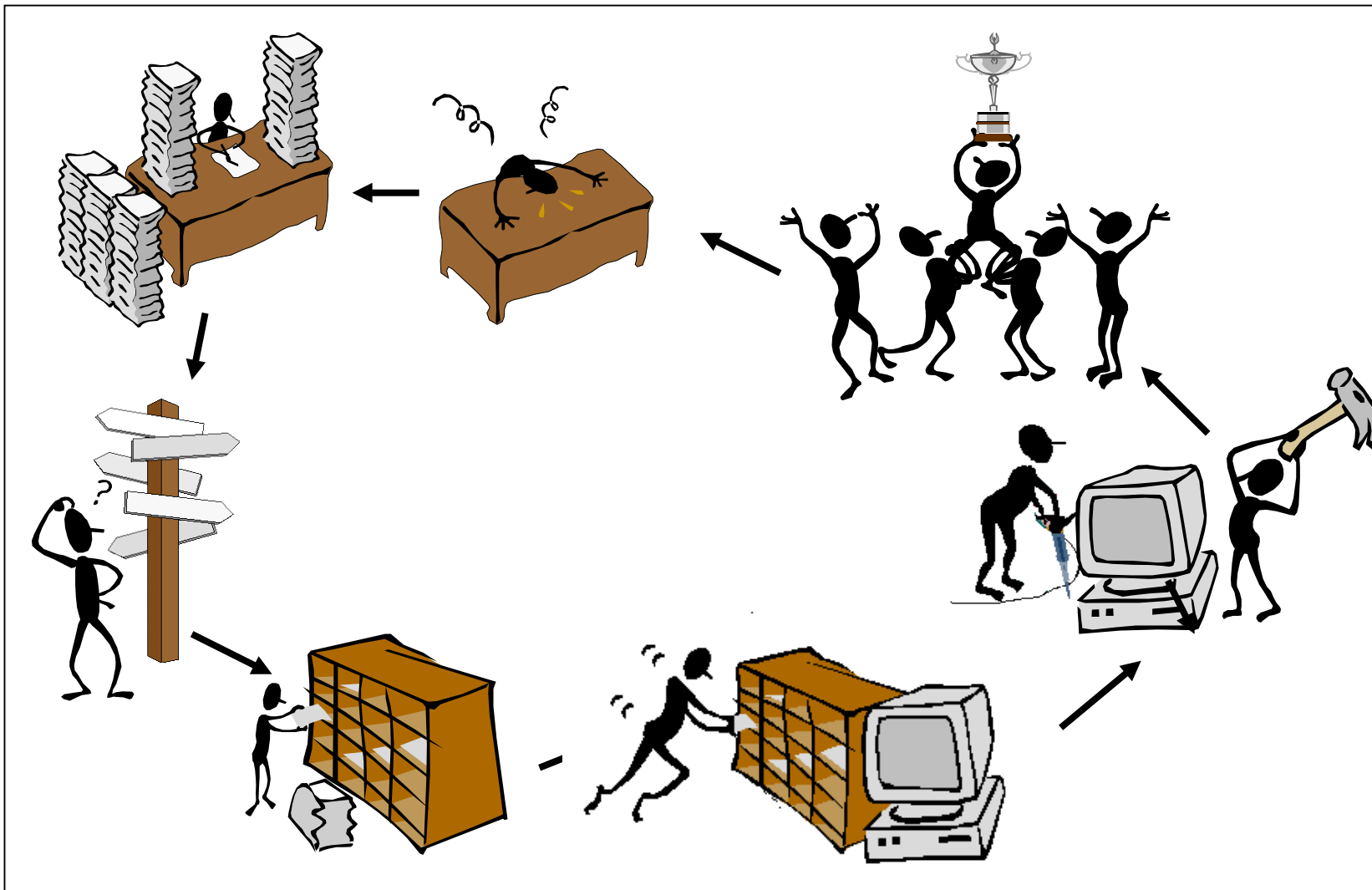


I hope you will enjoy this class 😊

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

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

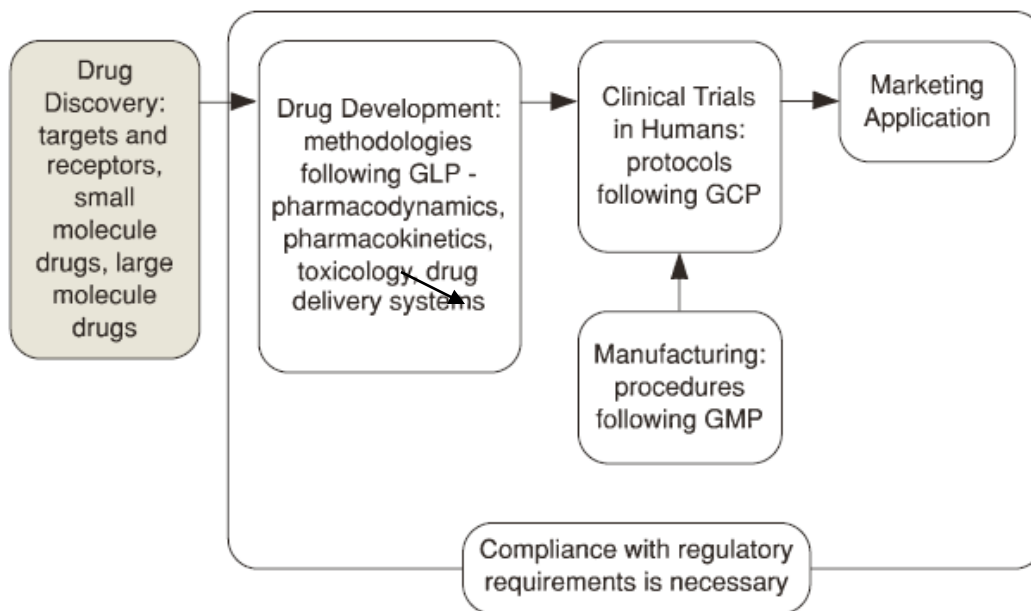


Figure from Rick Ng, *Drugs: From Discovery to Approval*

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

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., protein identification from mass-spectra
- **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**
- **Protein function prediction**
- **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.

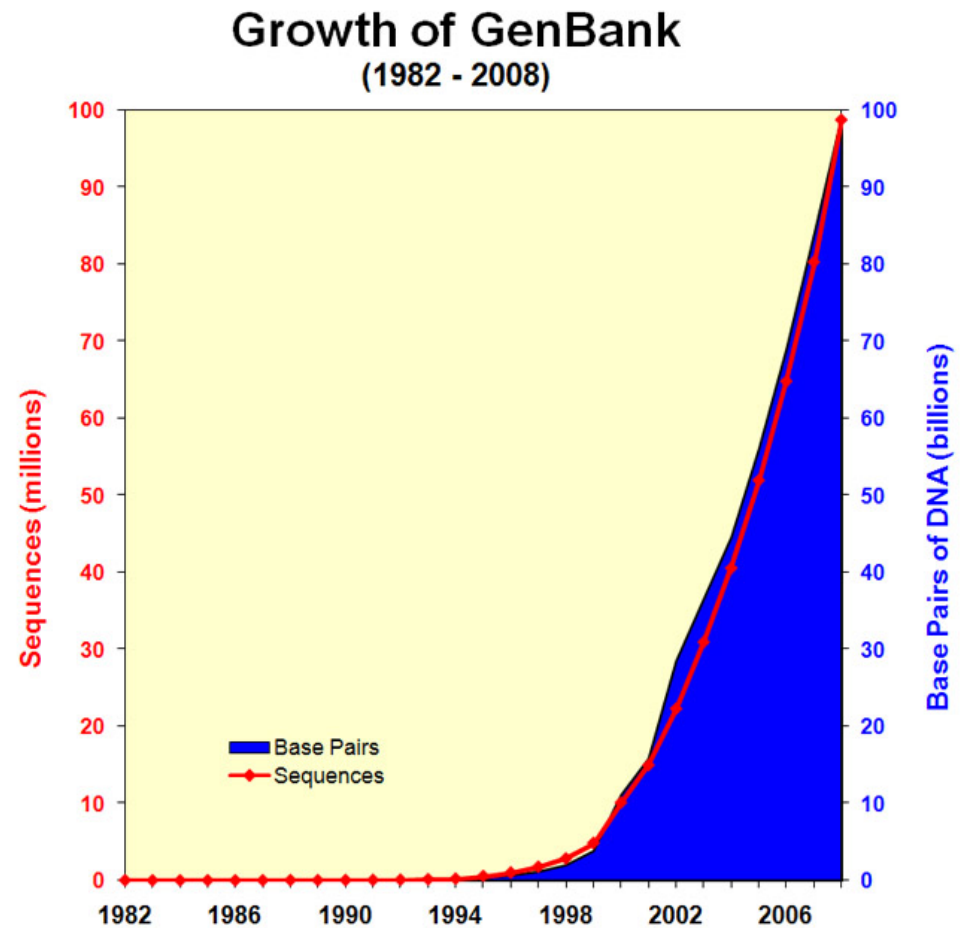


Image credit: NCBI

Cis-regulatory DNAs

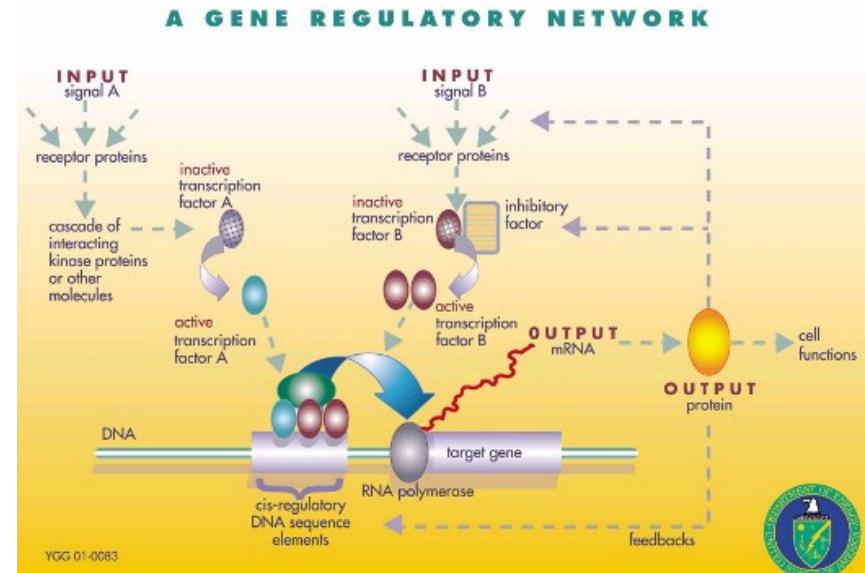
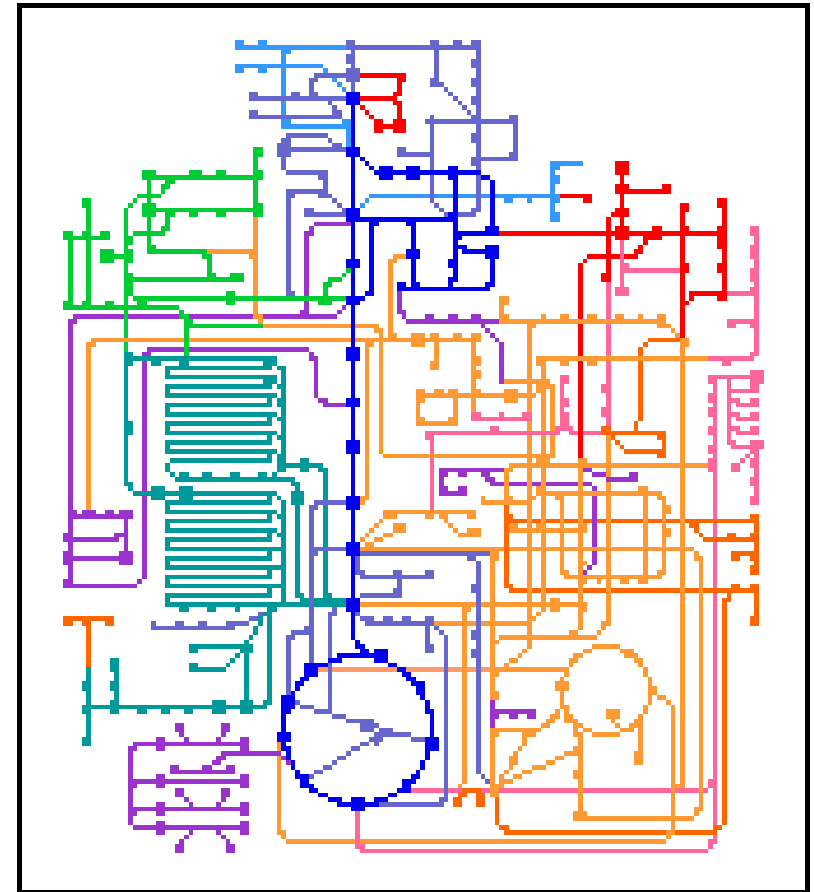


Image credit: US DOE

- **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**

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

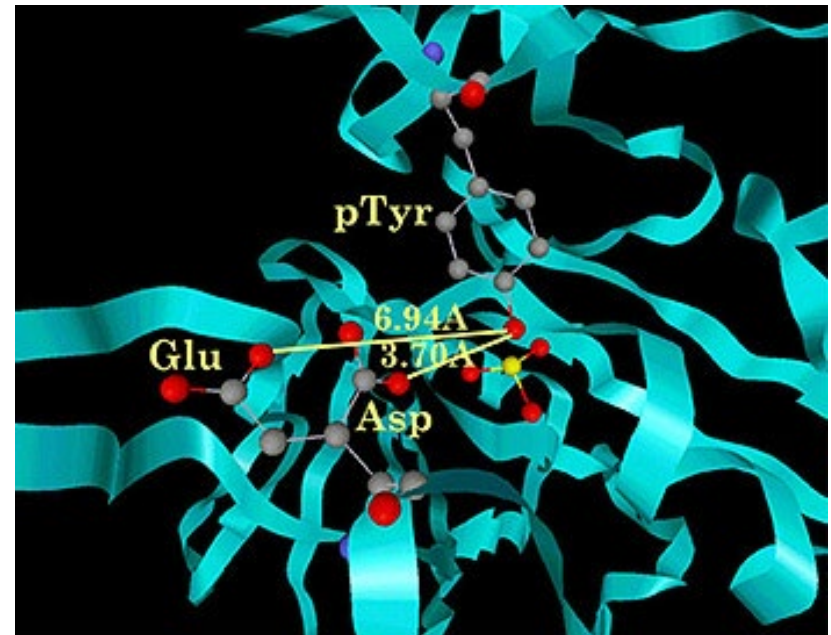


Image credit: Kolatkar

Evolutionary tree reconstruction

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

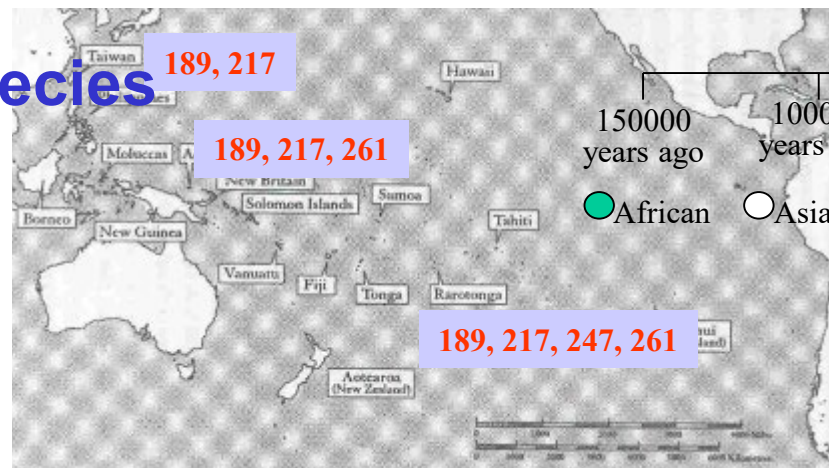
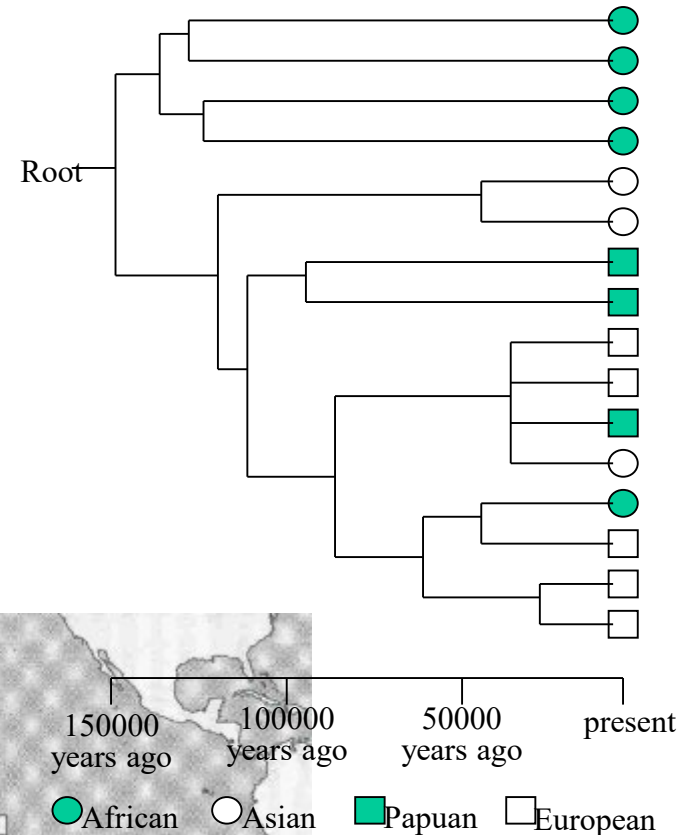


Image credit: Sykes

Breast cancer outcome prediction

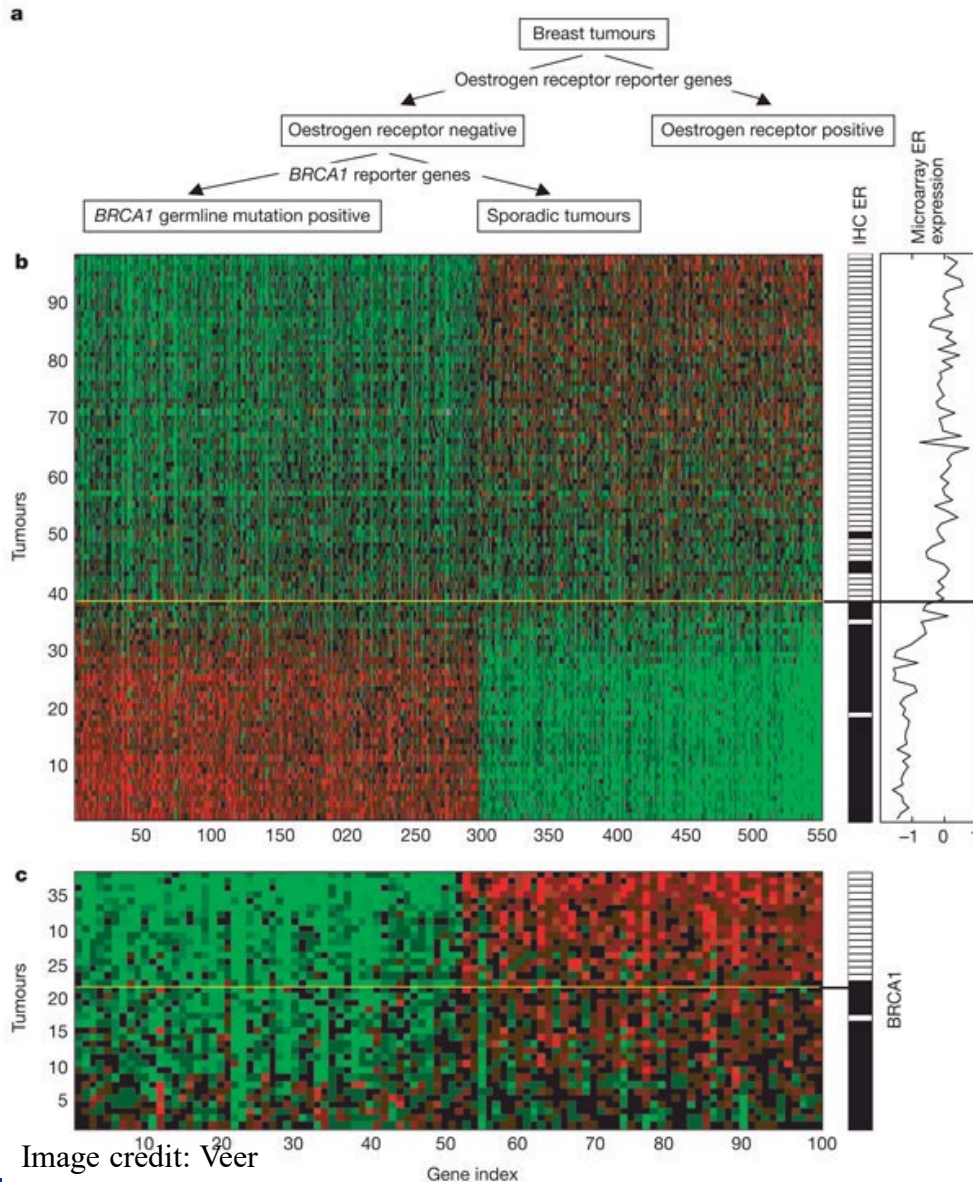


Image credit: Veer

- **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

- **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

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

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 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

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

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