I529: Machine Learning in Bioinformatics

Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington

Spring 2013

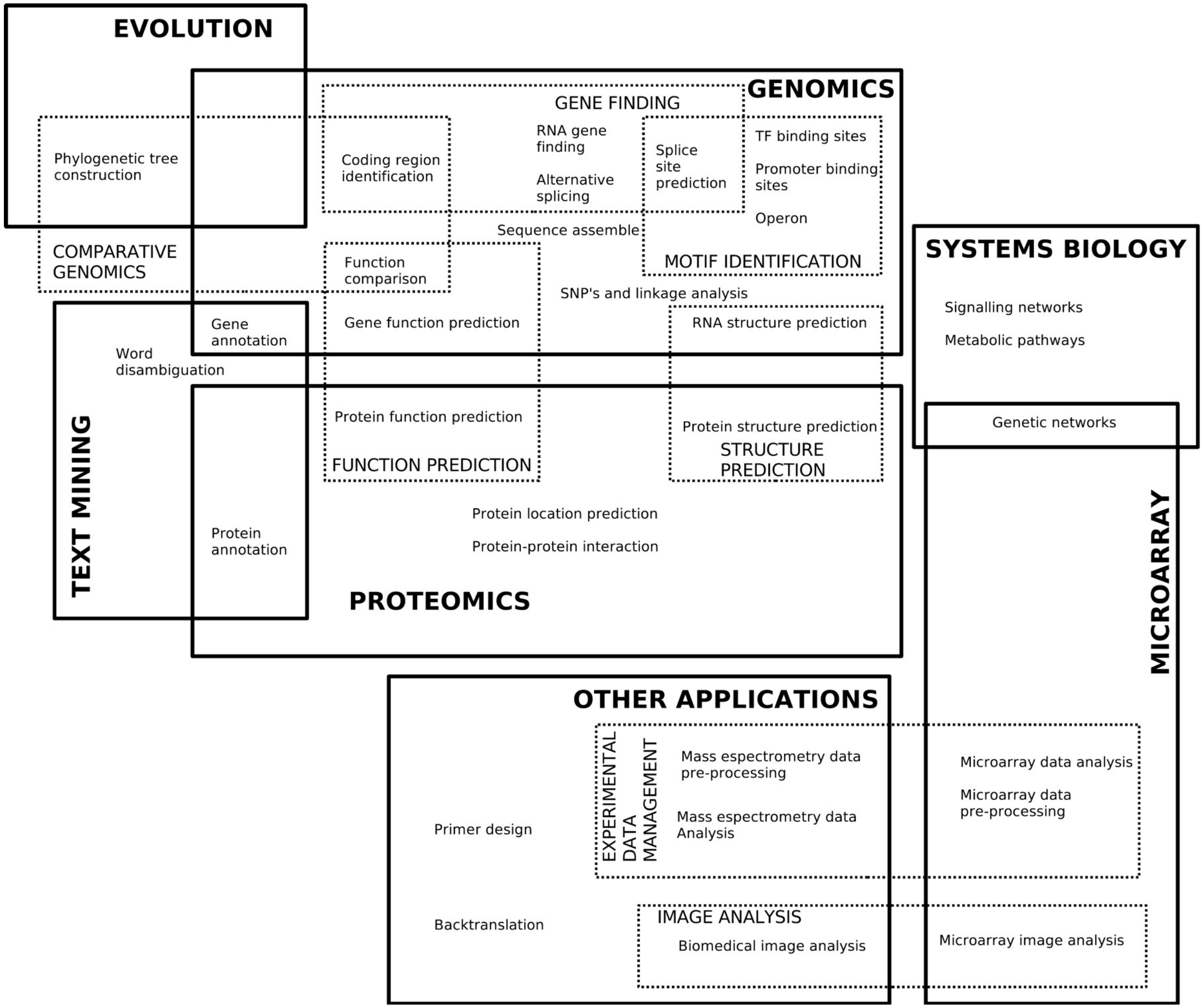
Overview

- **Prerequisites** I519 or equivalent knowledge in bioinformatics.
- **Grading:**
  - Combined assignments (30%), One mid-term exam (25%), Final exam (25%), Class Project (20%)
  - Attendance will be considered in borderline cases.
- **Required textbook**:
  - Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids , Cambridge University Press, 1999, (BSA)

Work with large amounts of biological data

- Efficient information storage and management
- Extraction of useful information from these data
  - development of tools and methods capable of transforming heterogeneous data into biological knowledge about the underlying mechanism.

Classification of the topics where machine learning methods are applied.



EVOLUTION

Phylogenetic tree construction

COMPARATIVE GENOMICS

GENOMICS

GENE FINDING

Coding region identification

RNA gene finding

Alternative splicing

Splice site prediction

TF binding sites

Promoter binding sites

Operon

MOTIF IDENTIFICATION

Sequence assemble

Function comparison

SNP's and linkage analysis

Gene annotation

Gene function prediction

RNA structure prediction

SYSTEMS BIOLOGY

Signalling networks

Metabolic pathways

Genetic networks

TEXT MINING

Word disambiguation

Protein function prediction

FUNCTION PREDICTION

Protein structure prediction

STRUCTURE PREDICTION

Protein annotation

Protein location prediction

Protein-protein interaction

PROTEOMICS

MICROARRAY

OTHER APPLICATIONS

EXPERIMENTAL DATA MANAGEMENT

Mass espectrometry data pre-processing

Mass espectrometry data Analysis

Microarray data analysis

Microarray data pre-processing

Primer design

Backtranslation

IMAGE ANALYSIS

Biomedical image analysis

Microarray image analysis

Larrañaga P et al. Brief Bioinform 2006;7:86-112

Briefings in Bioinformatics

A few definitions

- Machine learning
  - A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E* (T. Mitchell)
  - There are many different machine learning algorithms: supervised (classification) vs unsupervised (clustering); discriminative vs generative
- Probabilistic models
  - A model means a system that simulates the object under consideration
  - A probabilistic model is one that produces different outcomes with different probabilities (BSA)

An example

- Profile HMM
  - Each protein family is presented as a HMM
  - Given a new protein sequence, does it belong to one of known families (which model has the best chance of producing this sequence)?

Types of data sets

- Record data
  - For the most basic form of record data, there is no explicit relationship among records or data fields, and every record (object) has the same set of attributes
- Graph-based data
  - Data with relationships among objects: the data objects are the nodes, and the relationships among objects are captured by the links between objects.
  - Data with objects that are graphs: Protein structures; small molecules
- Ordered data
  - Sequential data
  - Sequence data
  - Time series data
  - Spatial data

Data compression

- String compression
  - It is becoming a new challenging problem in Bioinformatics, because of the rapid development of sequencing techniques!!
  - Need developments of bioinformatics tools that work on compressed data
    - Most current bioinformatics tools only work on only uncompressed sequence data
    - So need to expand the data before using

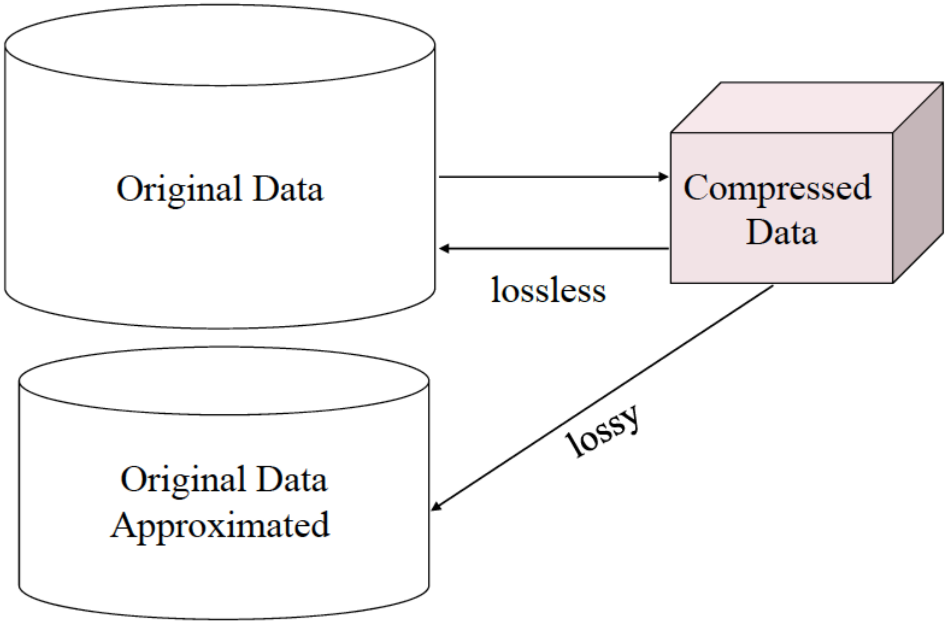
"Algorithms that compute directly on compressed genomic data allow analyses to keep pace with data generation"

Compressive Genomics, Nature Biotechnology 30, 627–630, 2012

Lossy or lossless compression

Lossless compression reduces bits by identifying and eliminating statistical redundancy. No information is lost in lossless compression.

Lossy compression reduces bits by identifying marginally important information and removing it.



Knowing your data

- Descriptive data summarization
  - Boxplot
  - Histogram
  - Quantile & Quantile-quantile plot
  - Scatter plot & Loess curve (add a smooth curve to scatter plot)
- Data preprocess/cleaning
  - Data can be incomplete, noisy, and inconsistent
  - No quality data, no quality mining
- Handle missing data
  - Exclude records with missing feature
  - Fill in manually
  - Fill in automatically (e.g., “UNK”, mean, most probable value).

Genomics and beyond

- Genomics
  - Study of the genomes of organisms
  - Problems: Gene finding; CpG island; motif finding
- Epigenomics
  - Study of the epigenomes; the epigenome of an organisms is the complete set of epigenetic modifications on its genetic material
  - "The cells in a multicellular organism have nominally identical DNA sequences.., yet maintain different terminal phenotypes. This **nongenetic** cellular memory, which records developmental and environmental cues.., is the basis of epi-(above)–genetics. " (Science, *330 no. 6004 p. 611, 2010*)
  - Problems: chromotin state decoding; identification of combinatorial epigenetic regulation patterns; identification of DNA methylation states
- Metagenomics
  - Study of the genomic sequences of an entire microbial community
  - Problems: classification of 16S rRNA reads and shotgun sequences
- Systems biology
  - Systems biology is the study of systems of biological components, which may be molecules, cells, organisms or entire species.
  - Problems: integration of inhomogeneous data for function prediction and for the inference of cellular pathways

Focus of I529

- Probabilistic models (Markov models, Hidden Markov models, and Bayesian networks) for biological sequence analysis and systems biology.
- Applications in genomics and beyond (e.g., metagenomics, and epigenomics)
- Other machine learning approaches will only be covered briefly.