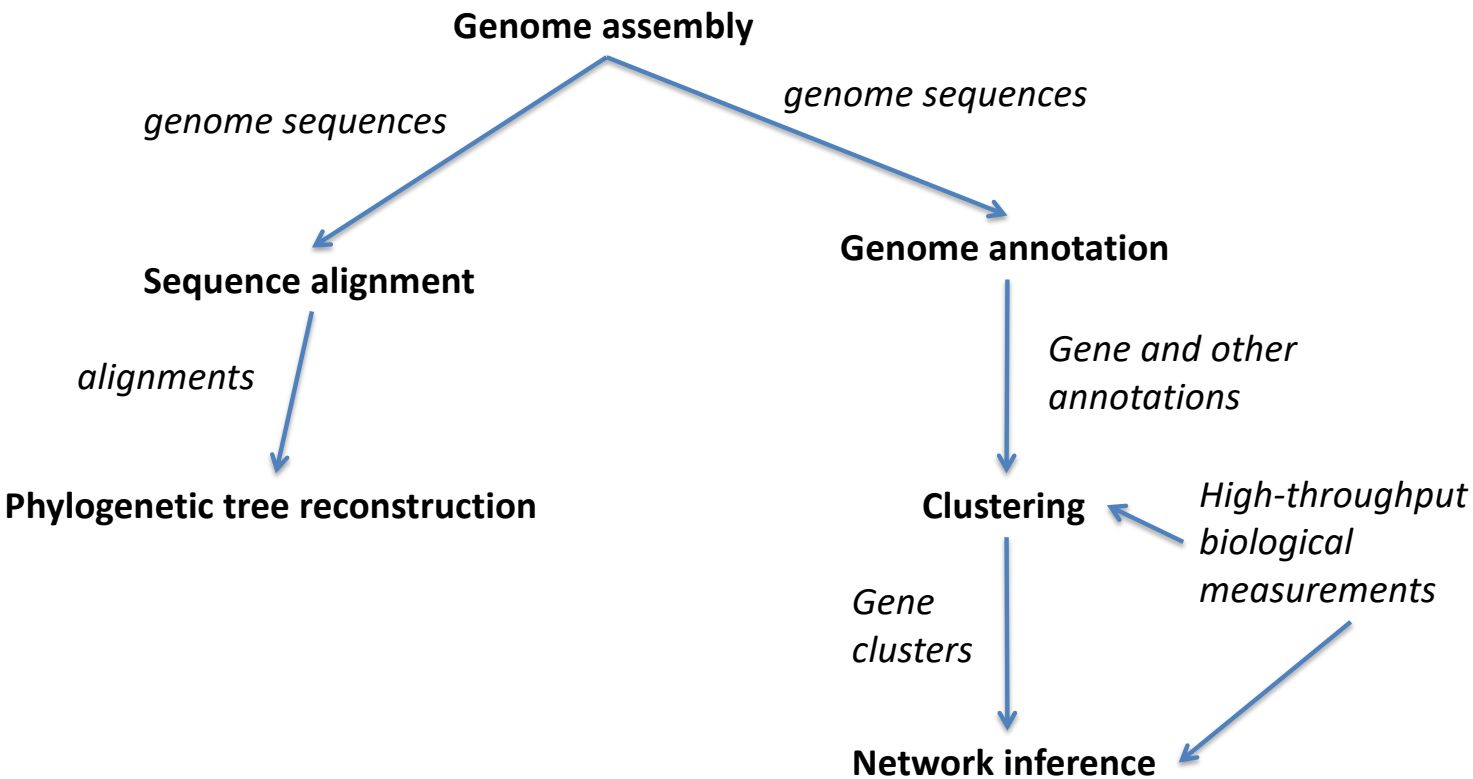
# Introduction to Bioinformatics – Final Thoughts

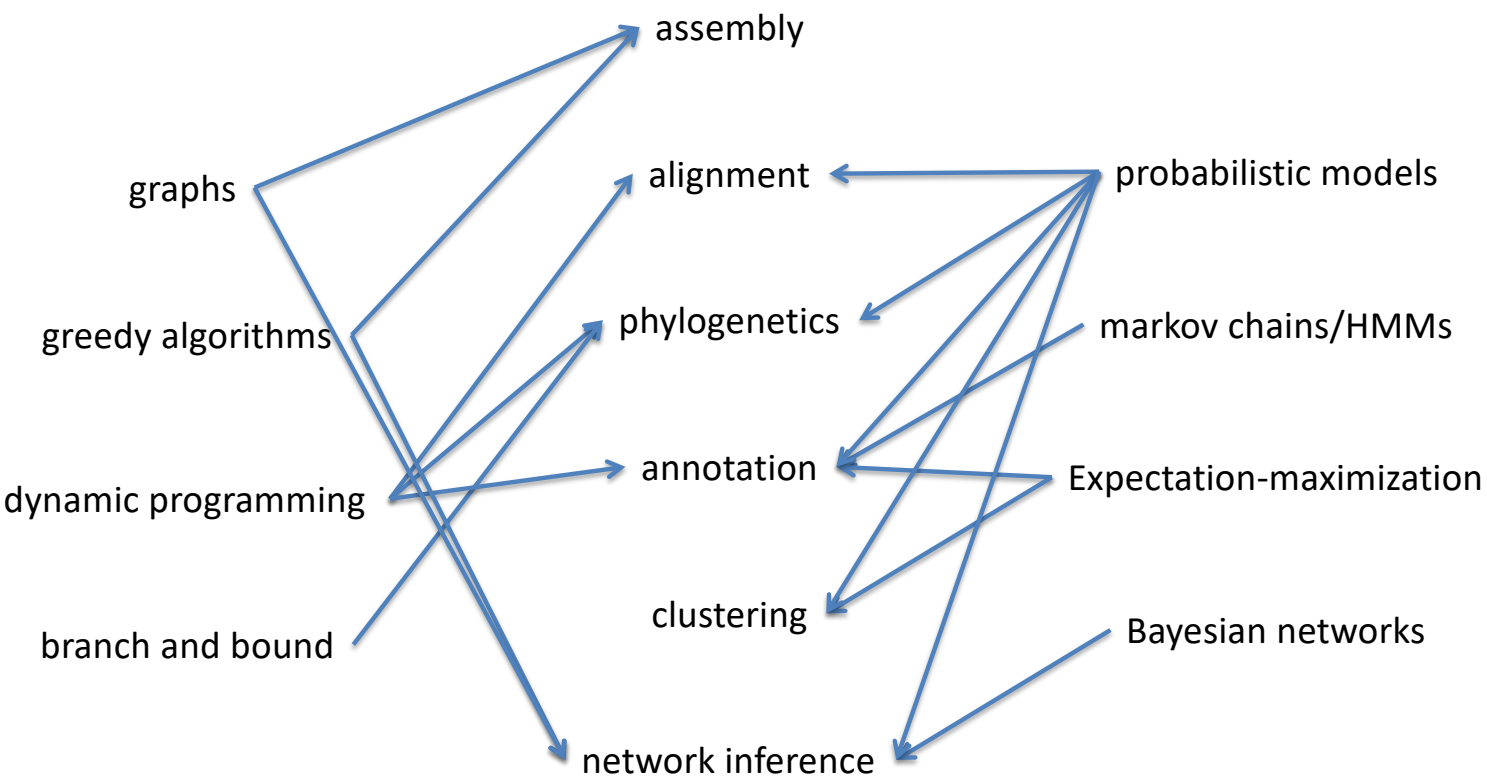
**Colin Dewey**

**Fall 2019**

# Where we've been



# Concept map


assembly
graphs
alignment
probabilistic models
greedy algorithms
phylogenetics
markov chains/HMMs
dynamic programming
annotation
Expectation-maximization
clustering
branch and bound
Bayesian networks
network inference

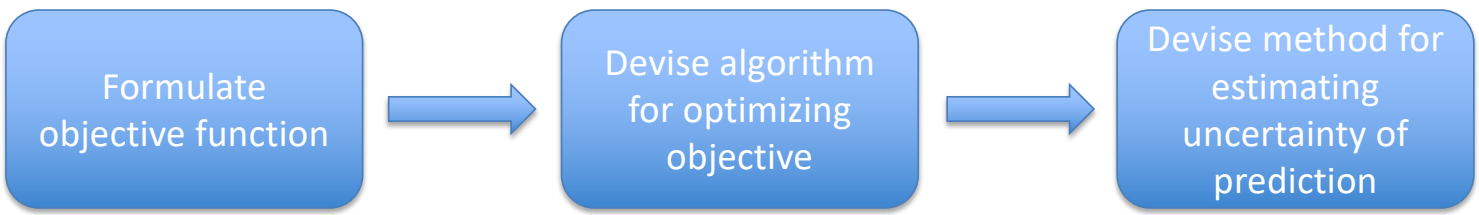
# The many facets of bioinformatics

- Molecular biology -> data and tasks
- Statistics -> models and objective functions
- Mathematics -> structural insight
- Computer Science -> efficient computation

# Science as unsupervised learning

- We do not know the “truth”
- There are only **data** and **models**
- Generally, the goal is to find the model that maximizes P(model | data)

# General strategy



# Objective functions

- Assembly: shortest superstring
- Alignment: sum of substitution scores + gap penalty
- HMMs: maximum likelihood
- Clustering: within-cluster scatter/maximum likelihood
- Networks: P(graph | data)

Formulate objective function

Objective function: should be based on the biology and optionally on computational complexity

# Devising algorithms

Devise algorithm for optimizing objective

- Graph-based algorithms
- Greedy algorithms
- Dynamic programming
- Branch and bound
- Is NP-hard?
  - use heuristics to obtain fast (but perhaps not optimal) result
  - Use exact algorithm but cut down on search space (e.g., branch and bound)

# Estimating uncertainty

Devise method for estimating uncertainty of prediction

- The bootstrap
- Equally-optimal solutions (high-road, low-road alignments)
- Posterior probabilities
- Multiple initializations

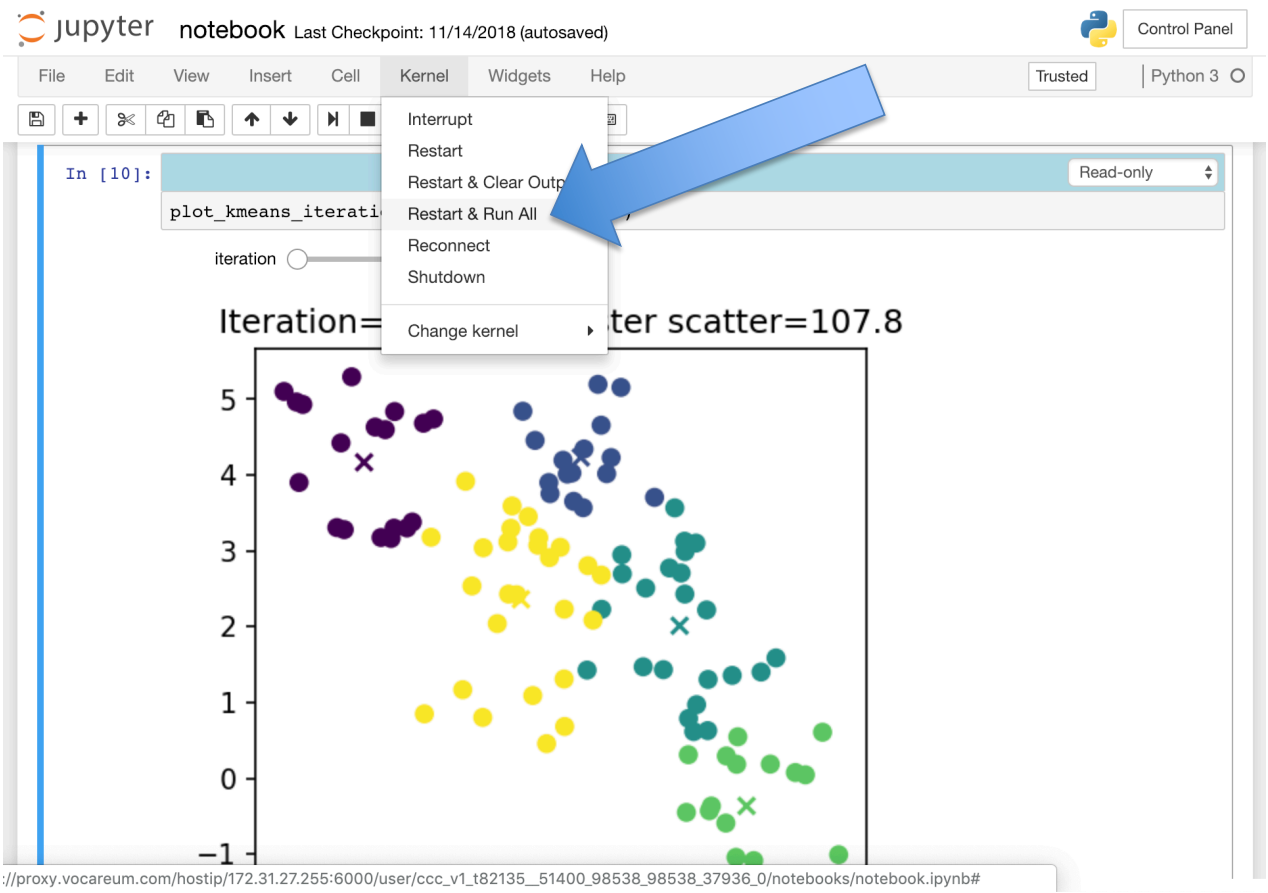
# “Magical” moments

- *Assembly*: Eulerian path formulation -> Polynomial time algorithm
- *Alignment*: Dynamic programming
- *Phylogenetic trees*: Neighbor joining “corrected” distances
- *Genome annotation*: Baum–Welch algorithm for unsupervised learning of HMMs
- *Clustering*: Soft clustering with Gaussian mixture models
- *Networks*: Closed-form expression for Bayesian network structure score

# Reproducible research

- "A minimal standard for data analysis and other scientific computations is that they be *reproducible*: that the code and data are assembled in a way so that another group can re-create all of the results (e.g., the figures in a paper)" – Dr. Karl Broman
- http://kbroman.org/Tools4RR/

# Jupyter notebooks: A valuable tool for reproducible research



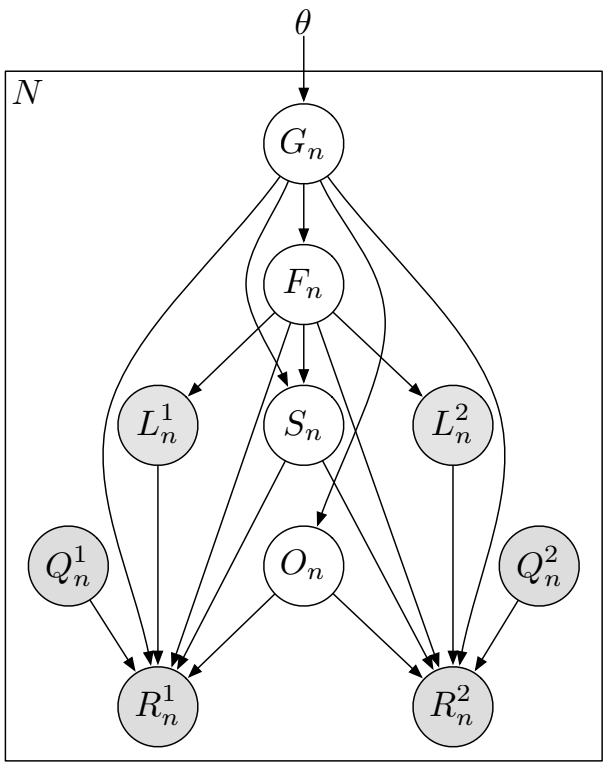
# To further develop your reproducible research skills

- Prof. Karl Broman's "Tools for Reproducible Research" course:
  - http://kbroman.org/Tools4RR/
- Other tools
  - git - version control
  - make (or similar) – workflow management
    - if you like Python: snakemake
  - R – rmarkdown/knitr

# Examples from the Dewey Lab: RSEM

- Given: RNA-seq data, gene annotations
- Outputs: estimates of gene expression values

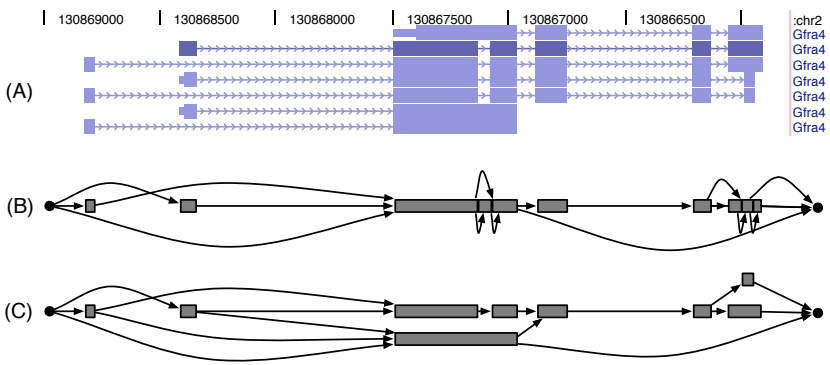
- Techniques:
  - Bayesian networks
  - Expectation–Maximization
  - Sequence alignment




B. Li, V. Ruotti, R. Stewart, J. Thomson, and C. Dewey (2010) **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 26(4): 493-500.

# Examples from the Dewey Lab: PSGInfer

- Given: RNA-seq data, gene splice graphs
- Outputs: estimates of splicing probabilities

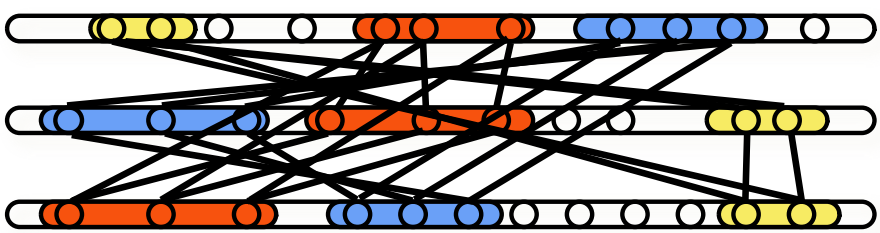



- Techniques:
  - Graph-based
  - Markov chains
  - Dynamic programming
  - Expectation–Maximization

L. LeGault and C. Dewey. (2013) **Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs**. *Bioinformatics*. 29(18):2300-2310.

# Examples from the Dewey Lab: Mercator

- Given: Multiple whole genome sequences
- Outputs: A whole genome multiple alignment



- Techniques:
  - Graph-based
  - Probabilistic graphical models
  - Sequence alignment

C. Dewey (2007) **Aligning multiple whole genomes with Mercator and MAVID**. In N. Bergman, editor, *Comparative Genomics*, volume 395 of *Methods in Molecular Biology*. Humana Press.

# Next steps

- BMI/CS 776: Advanced Bioinformatics
  - Taught in the spring (Prof. Daifeng Wang)
- Seminars
  - CIBM training program seminar
  - BMI departmental seminars
  - Genomics seminar
- Get involved in research

Thanks!