Bioinformatics and Computational Biology

Professor Mohammed El-Kebir

Computer Science • University of Illinois at Urbana-Champaign

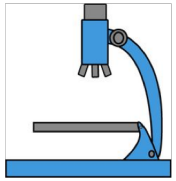
What is Computational Biology/Bioinformatics?

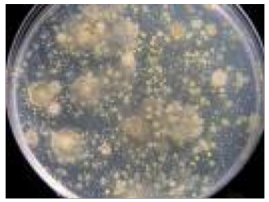
Computational biology and **bioinformatics** is an interdisciplinary field that develops and applies **computational methods** to analyze large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or **discover new biology**.

https://www.nature.com/subjects/computational-biology-and-bioinformatics

Technology and Bioinformatics are Transforming Biology

Until late 20th Century





Hypothesis Generation and Validation

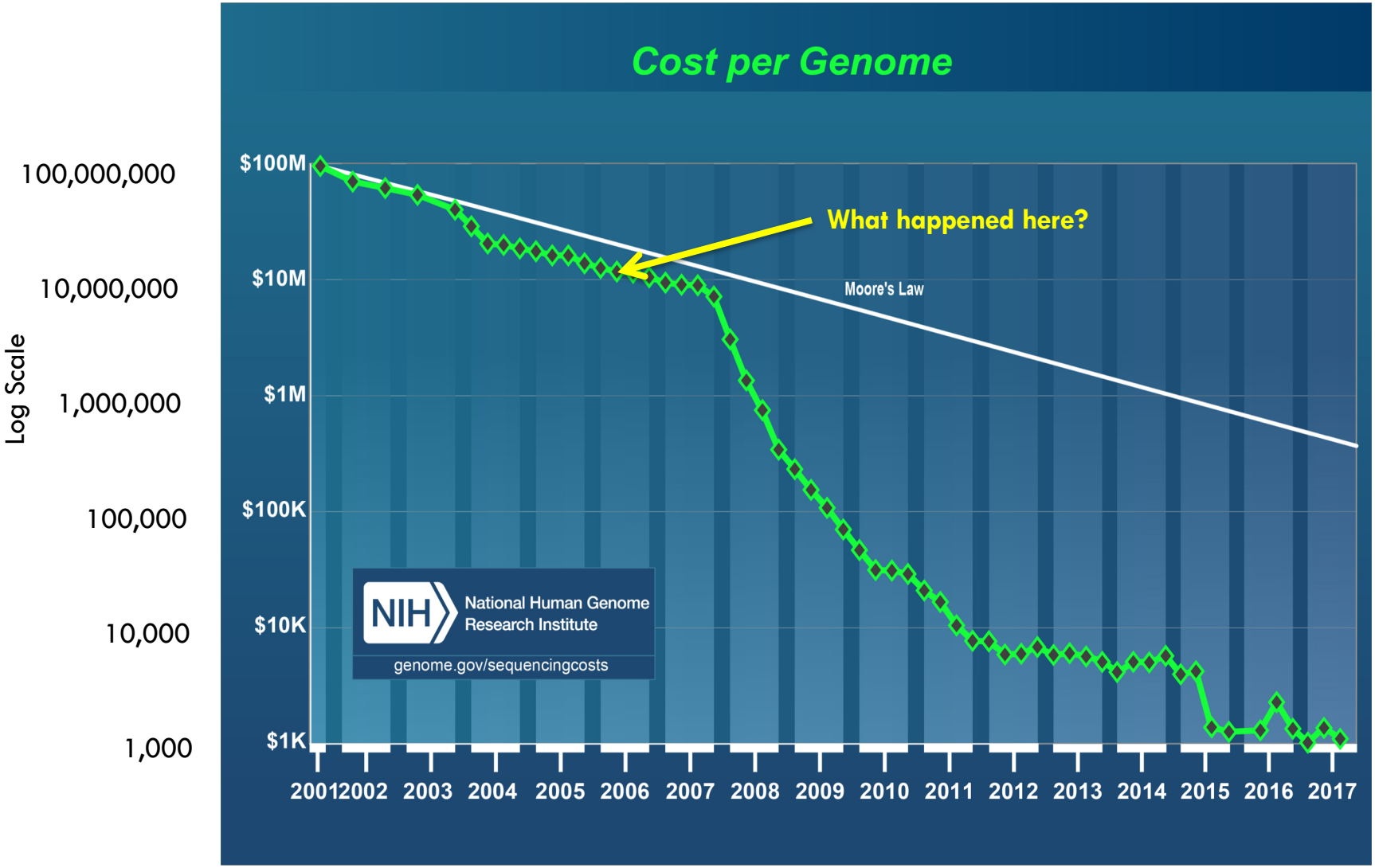
21th Century and Beyond

Algorithms

Hypothesis Generation and Validation

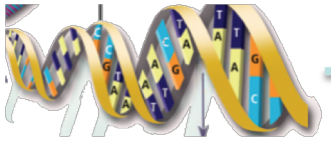
High throughput technologies

A Deluge of Data



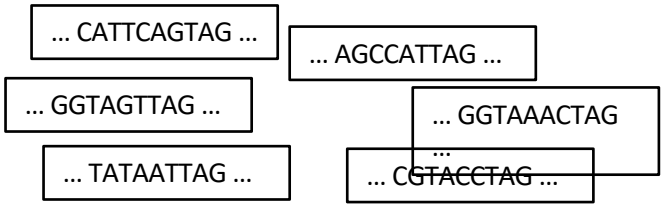
Question: What does it mean that we can sequence a genome?

No technology exists that can sequence a complete (human) genome from end to end!



Genome
Millions -billions nucleotides

Next-generation DNA sequencing



10-100's million noisy ***reads***
R*eads:* 30-1000 nucleotides

Making sense of this data absolutely requires the use and development of **algorithms**!

Why Study Computational Biology?

Interdisciplinary

- Biology
- Computer Science
- Mathematics
- Statistics

= FUN!

Why choose just 1?

Best Jobs	Worst Jobs
1. Actuary	200. Newspaper reporter
2. Audiologist	199. Lumberjack
3. Mathematician	198. Enlisted Military Personnel
4. Statistician	197. Cook
5. Biomedical Engineer	196. Broadcaster
6. Data Scientist	195. Photojournalist
7. Dental Hygienist	194. Corrections Officer
8. Software Engineer	193. Taxi Driver
9. Occupational Therapist	192. Firefighter
10. Computer Systems Analyst	191. Mail Carrier

http://www.careercast.com/jobs-rated/jobs-rated-report-2015-ranking-top-200-jobs

Donald Knuth
Professor emeritus of Computer Science at Stanford University
Turing Award winner
"father of the analysis of algorithms."

"I can't be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It's at that level."

Coursework for Bioinformatics Research

The usual computer science stuff, but especially

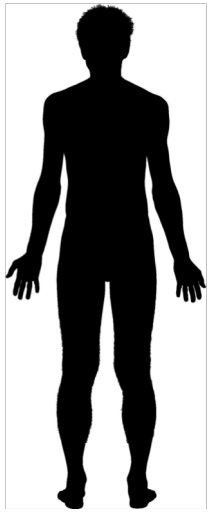
- CS 125 (programming)
- CS 173 (abstract thinking)
- CS 225 (data structures)
- CS 374 (algorithms and models for computation)

A bit of statistics is helpful (e.g., CS 361)

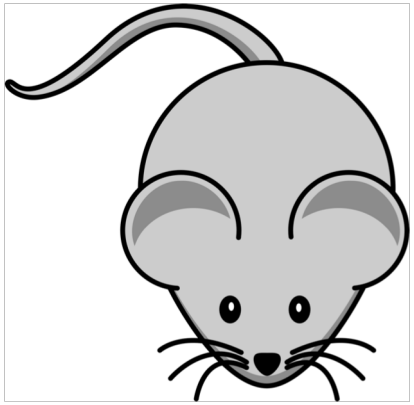
CS 466: Introduction to Bioinformatics! Good if you know some biology, but you can take CS 466, and learn it there!

Course Topic #1: Sequence Alignment

Question: How do we compare two genes/genomes?



vs.



Human Genome:

...ACTCGACTGAGAGGATTTCGAGCATGA...

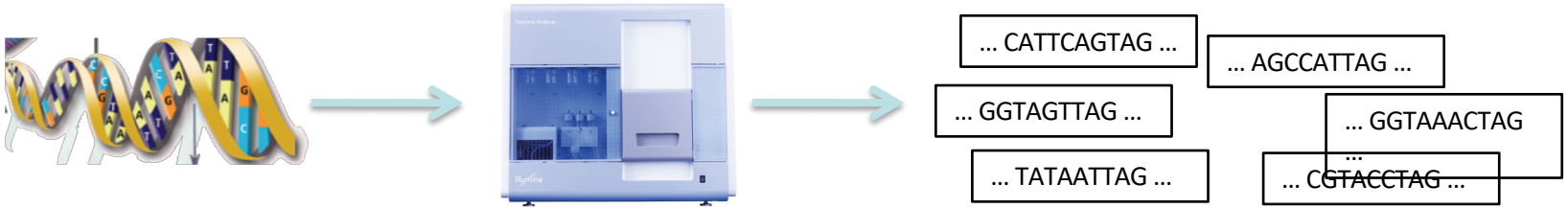
≈3.2 x 10^9 bp

Mouse Genome:

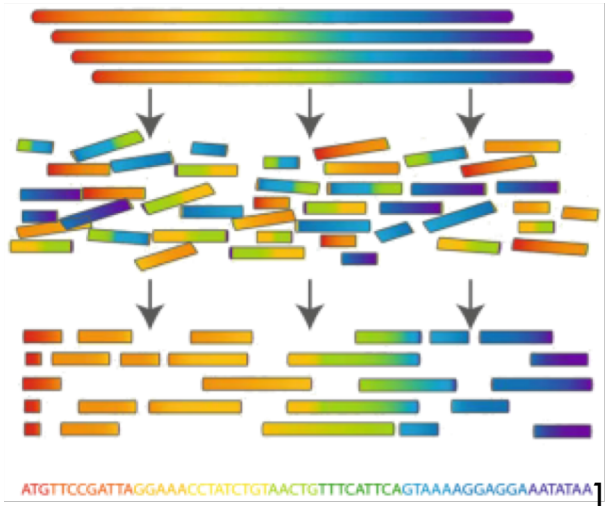
...ACTCAACTGAGATTCGAGCTTCAATGA...

≈2.8 x 10^9 bp

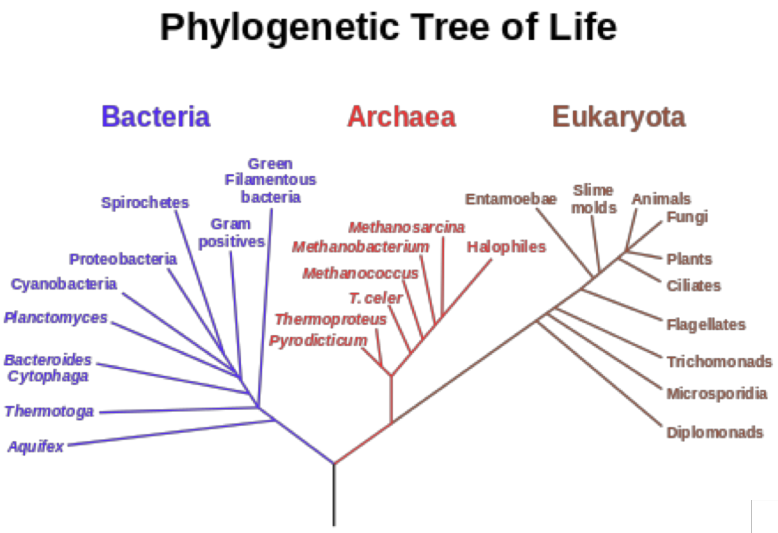
Course Topic #2: Genome Assembly



Question: How do we put all the pieces back together?



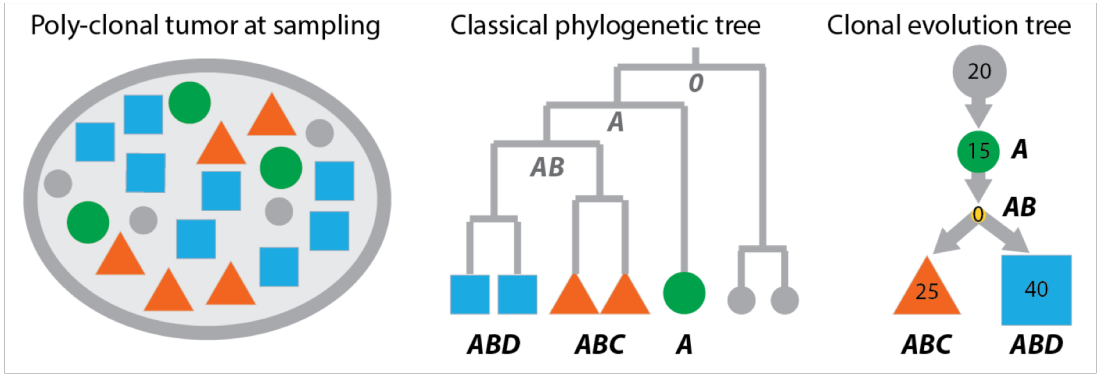
Course Topic #3: Phylogenetics



https://en.wikipedia.org/wiki/Phylogenetic_tree

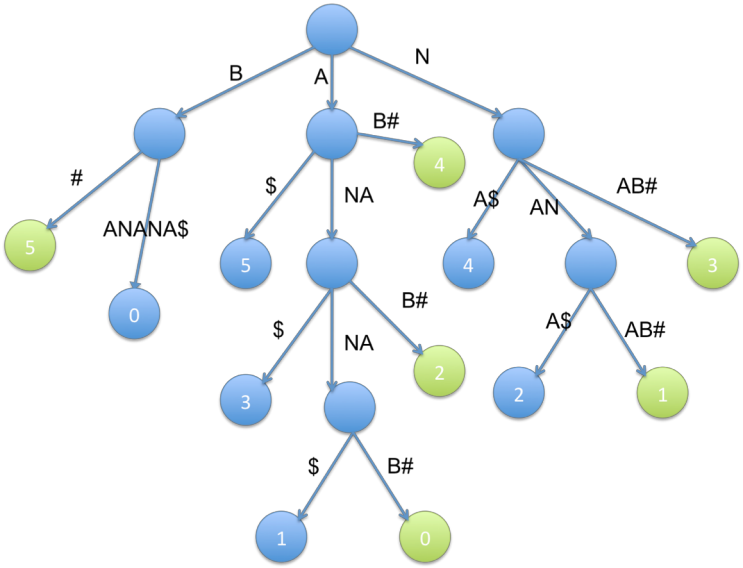
Question: Can we reconstruct the evolutionary history of different species?

Question: Can we recover how a tumor has evolved overtime?



https://scientificbsides.wordpress.com/2014/06/09/inferring-tumour-evolution-2-comparison-to-classical-phylogenetics/

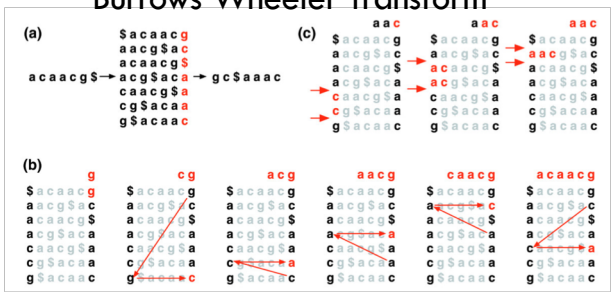
Course Topic #4: Pattern Matching



Suffix Trees

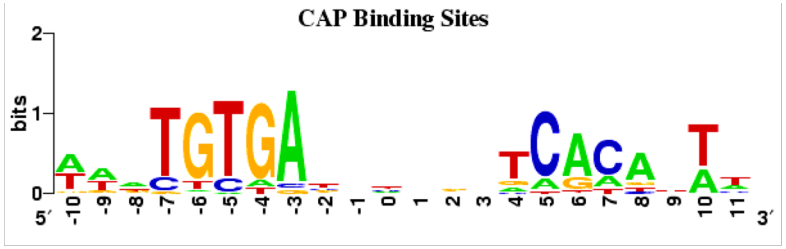
Question: How do we start to make sense of all these sequences?

Burrows Wheeler Transform

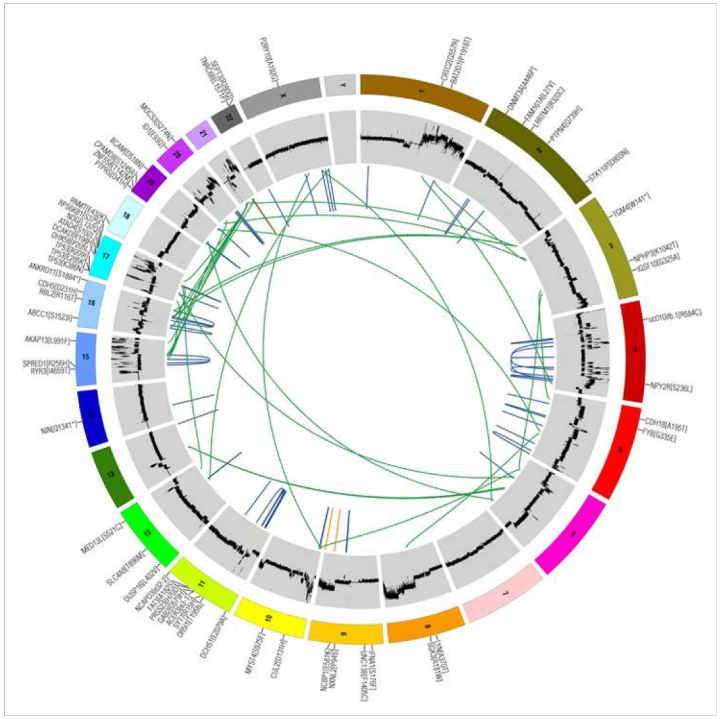


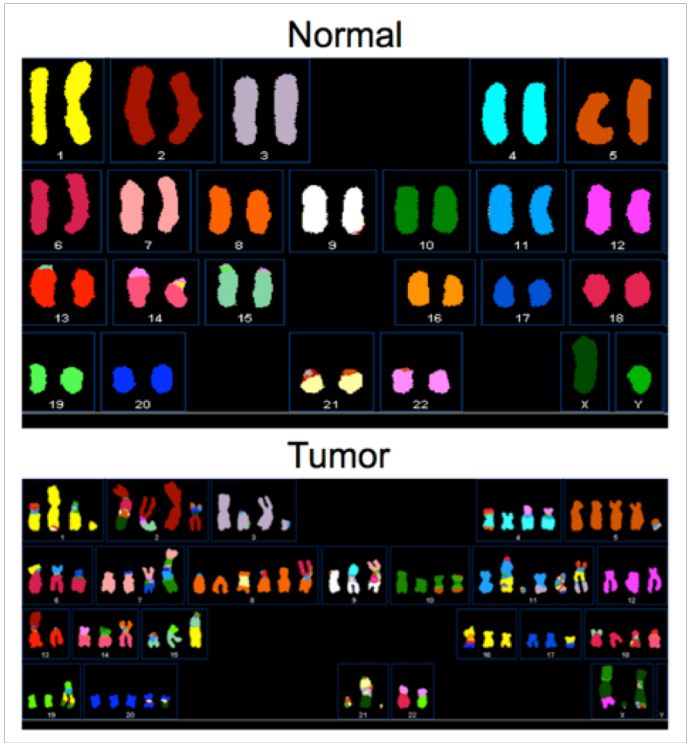
http://www.genomebiology.com/2009/10/3/R25/figure/F1?highres=y

Motif Finding



Course Topic #5: Cancer Genomics





Question: How can we analyze available data to determine what drives tumor growth and how to treat or prevent it?

Course Topics

1. Sequence alignment
   'How do we compare two genes/genomes?'

2. Genome assembly
   'How do we put all the pieces back together?'

3. Phylogenetics
   'What is the evolutionary history of different sequences?'

4. Pattern matching
   'How do we start to make sense out of all these sequences?'

5. Cancer genomics
   'How do we identify what drives tumor growth and how to treat/prevent it?'

Course Topics

1. Sequence alignment
   Dynamic programming: edit distance

2. Genome assembly
   Graphs: de Bruijn graph, Eulerian and Hamiltonian paths

3. Phylogenetics
   Trees and distances: distance matrices, neighbor joining, hierarchical clustering.
   Phylogenies: Sankoff/Fitch algorithms, perfect phylogeny and compatibility

4. Pattern matching
   Suffix trees/arrays. Burrows-Wheeler transform, Hidden Markov Models (HMMs)

5. Cancer genomics
   Cancer phylogenies: Integer linear optimization and graph algorithms

Bioinformatics & Computational Biology Group




Top: Mohammed El-Kebir, Jian Peng, Saurabh Sinha, Tandy Warnow

Bottom: ChangXiang Zhai, Jiawei Han, and Olgica Milenkovic

And others!



https://cs.illinois.edu/research/bioinformatics-and-computational-biology

Saurabh Sinha: gene regulation, big data to knowledge

Two broad areas:

How is information about us encoded in our DNA ?

How do we bring the latest and greatest in machine learning and graph mining to the biologist's desktop computer?

Research questions:

- **Gene regulation:** How are genes turned on and off in precisely orchestrated ways?
- **Regulatory evolution**: Can we model evolution of regulatory sequences?
- **Genomics of behavior:** How does DNA encode animal behavior ?
- **Cancer pharmacogenomics:** Can a person's DNA predict the best drug treatment?
- **Big Data To Knowledge (BD2K):** Build a **"**Knowledge Engine for Genomics".

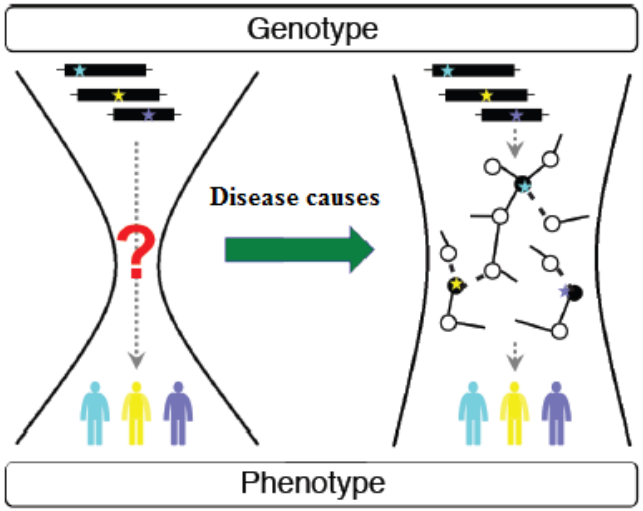
http://www.sinhalab.net/

Algorithmic Network Medicine

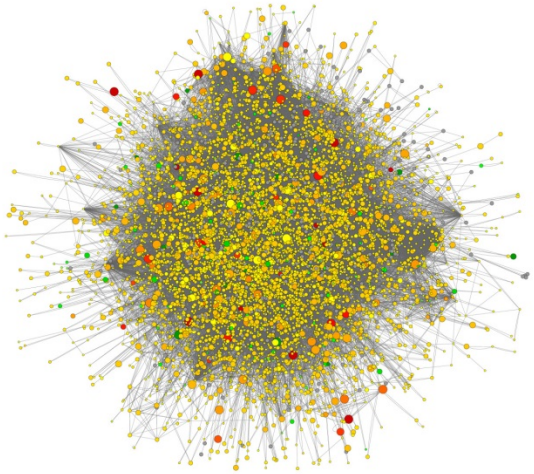
- Understanding human diseases from gene network and DNA
- Patient stratification for personalized medicine
- Acceleration of drug design

DNA data

Predictive Modeling



Gene Network

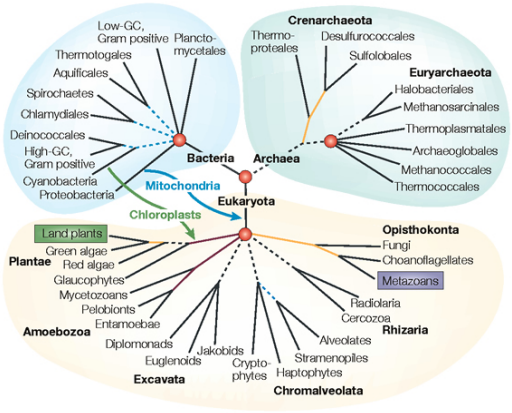


Precision Treatment

http://jianpeng.web.engr.illinois.edu

CS @ ILLINOIS

Tandy Warnow
The Tree of Life: *Multiple Challenges*

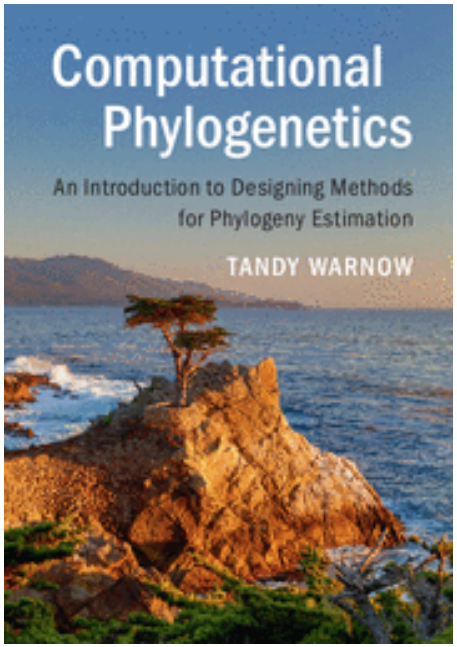


Nature Reviews | Genetics

Large datasets:

- 100,000+ sequences
- 10,000+ genes

"BigData" complexity



Large-scale statistical phylogeny estimation

- Ultra-large multiple-sequence alignment
- Estimating species trees from incongruent gene trees
- Supertree estimation
- Genome rearrangement phylogeny
- Reticulate evolution

Visualization of large trees and alignments

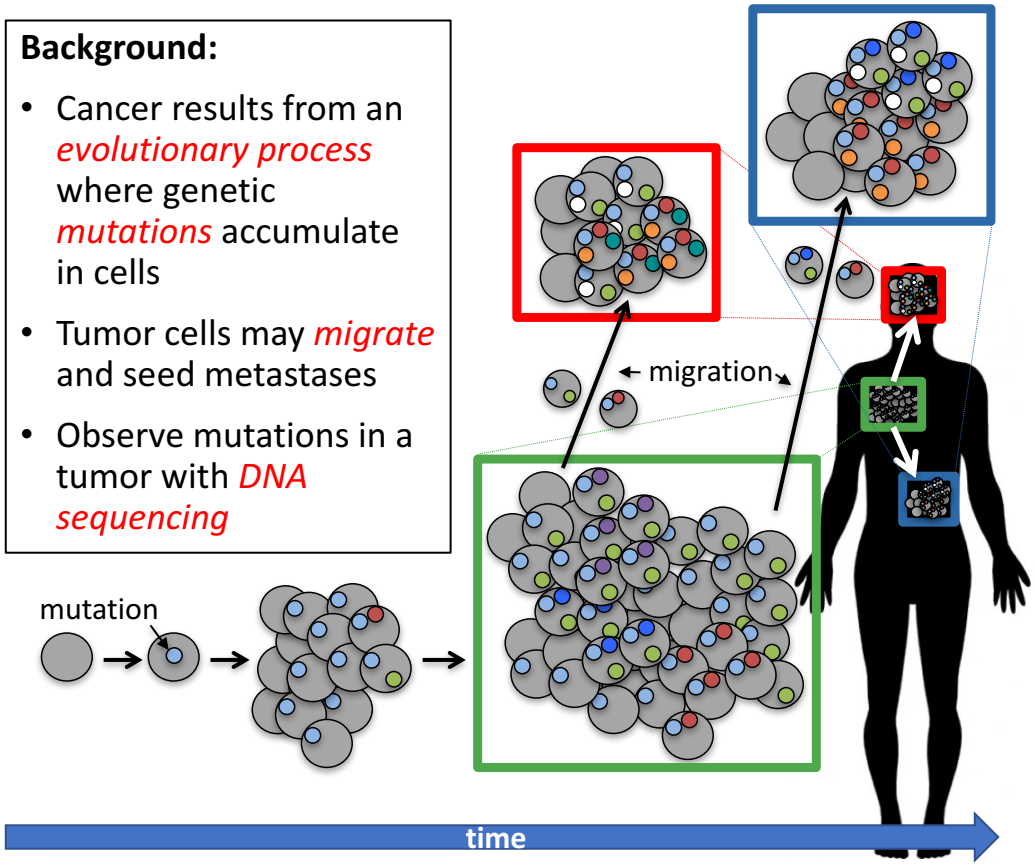
Data mining techniques to explore multiple optima

http://tandy.cs.illinois.edu

Computational Cancer Genomics -- Mohammed El-Kebir

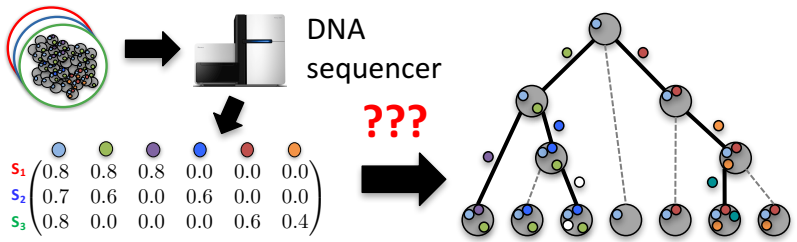
Background:

- Cancer results from an *evolutionary process* where genetic *mutations* accumulate in cells
- Tumor cells may *migrate* and seed metastases
- Observe mutations in a tumor with *DNA sequencing*

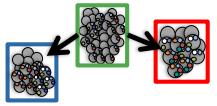


Research topics:

- How does a tumor *evolve*?
  - Tumor phylogeny inference from mutation frequencies



- How does *metastasis* take place?
  - Migration analysis of tumor cells



- How does a tumor become *resistant* to treatment?
  - Comparative analysis of pre and post treatment data

- Which mutations *drive* cancer progression?
  - Classify mutations into drivers and passengers

http://www.el-kebir.net