Introduction to Bioinformatics

by Eka Antonius Kurniawan

Outline

- What is Bioinformatics?
- Impacts on Daily Life
- Fields in Bioinformatics
- Why Bioinformatics?
- Challenges
- Jobs in Bioinformatics
- Bioinformatics Algorithms

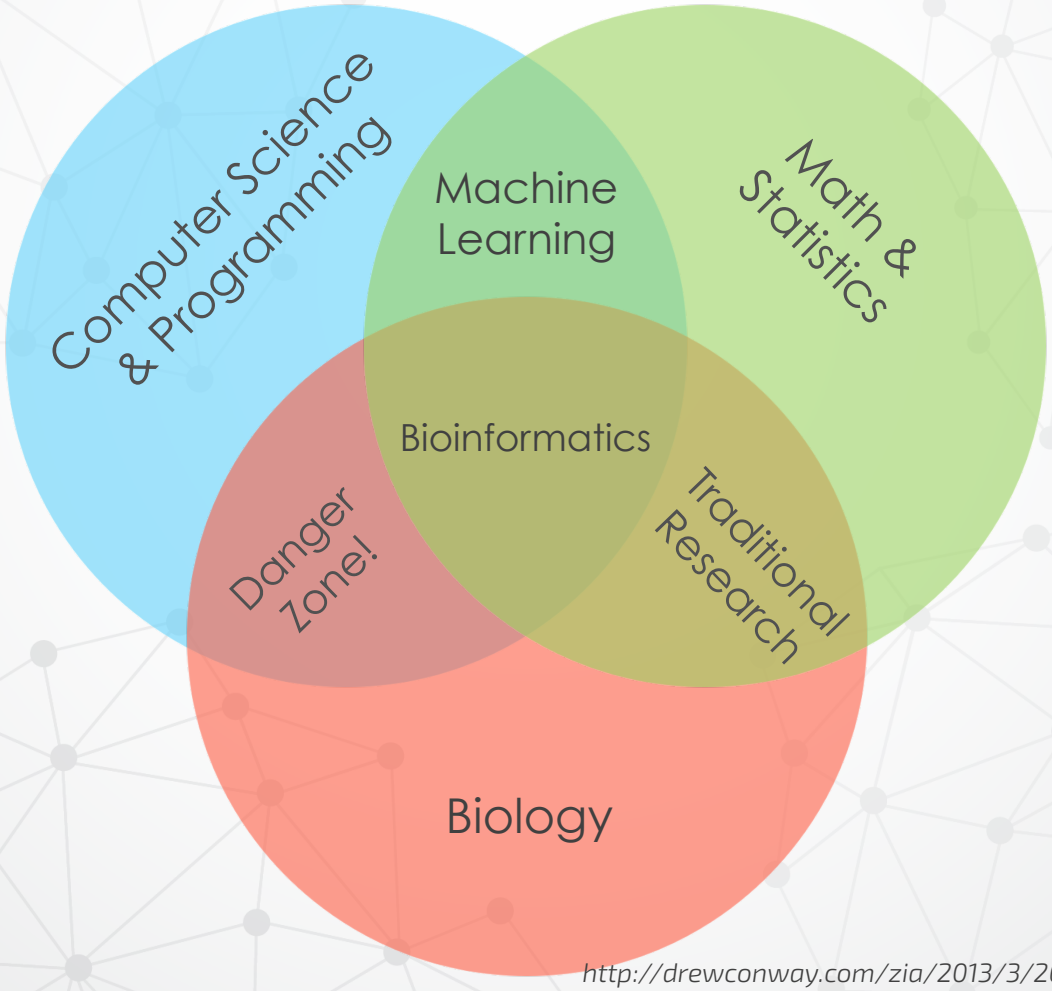
What is Bioinformatics?

What is Bioinformatics?

Bioinformatics is an **interdisciplinary** field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to study and **process** biological data.

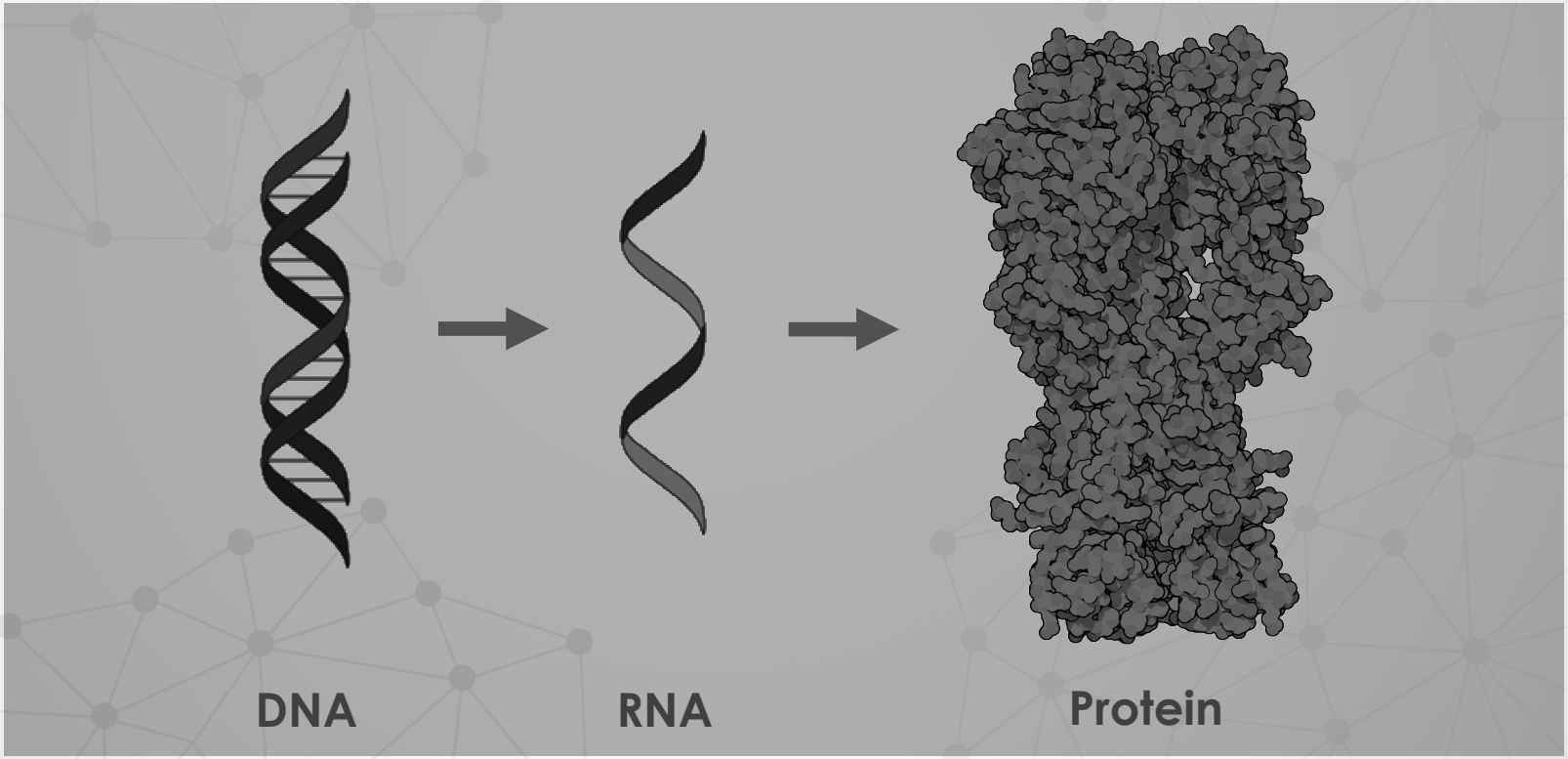
http://en.wikipedia.org/wiki/Bioinformatics

Adjusted Data Science Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Central Dogma of Molecular Biology



DNA Sequencing Process

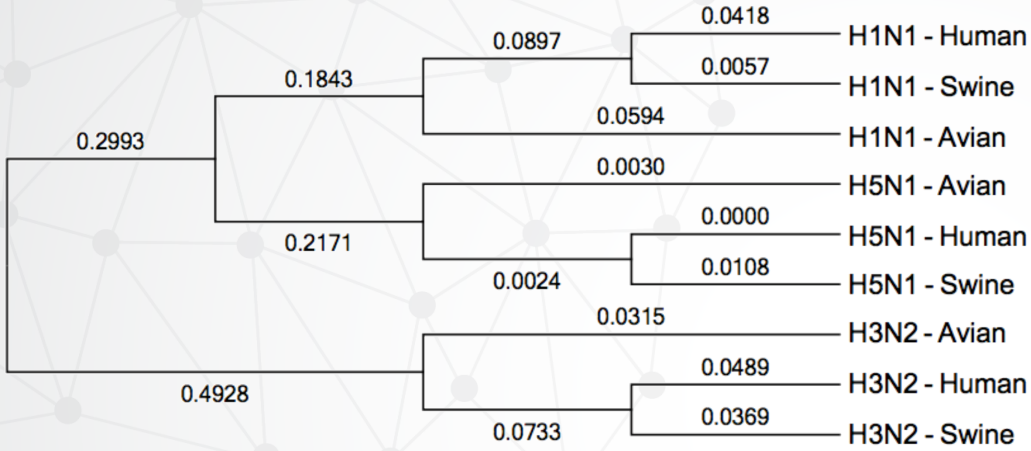
Cell Preparation

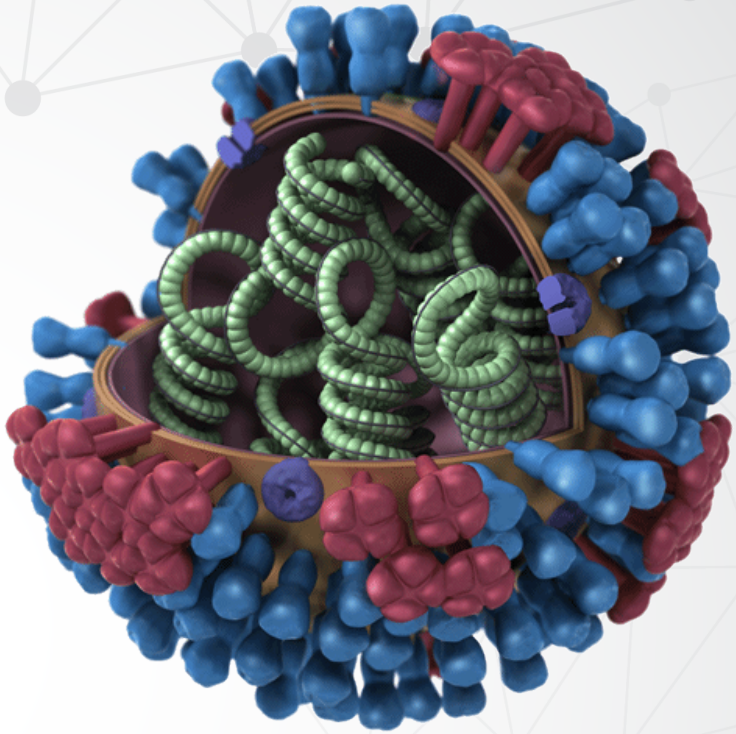
Multiplication and Sequencing

Assembling

http://en.wikipedia.org/wiki/File:Mapping_Reads.png

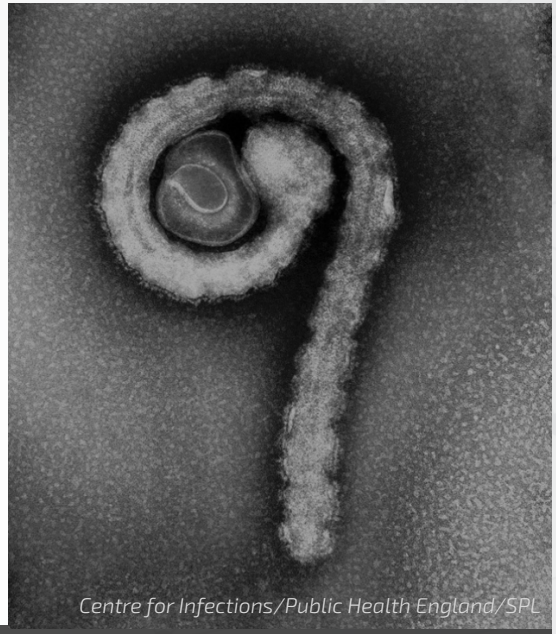
Impacts on Daily Life





Triple-reassortant Influenza A Virus

http://www.cdc.gov/flu/images.html

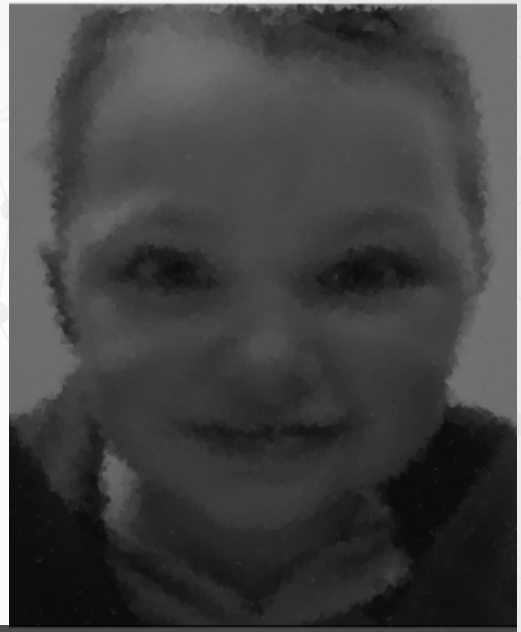


Centre for Infections/Public Health England/SPL

Sequencing the Ebola virus

> The mutations do not seem to be affecting the efficacy of experimental drugs and vaccines, some of which have been given to patients in this outbreak. **Some changes have occurred in regions of the genome** that are targeted by diagnostic tests. This does not mean the tests are ineffective, but confirming this and continuing to monitor such mutations will be crucial, Chiu says.

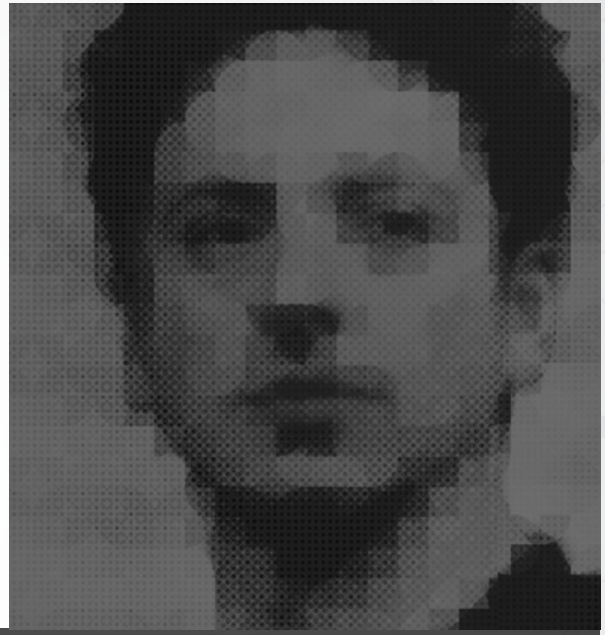
Ebola virus mutating rapidly as it spreads by Erika Check Hayden
http://www.nature.com/news/ebola-virus-mutating-rapidly-as-it-spreads-1.15777



Nicholas Volker: The First Child Saved By DNA Sequencing

This is the first clear example of what many doctors have been predicting for several years: that for patients with rare, undiagnosable diseases, **DNA sequencing** will become the court of last resort.

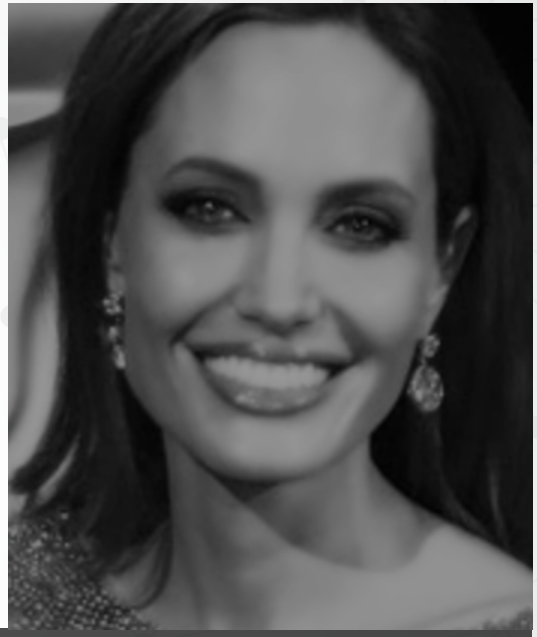
The First Child Saved By DNA Sequencing by Matthew Herper
http://www.forbes.com/sites/matthewherper/2011/01/05/the-first-child-saved-by-dna-sequencing/



Sergey Brin: Search for a Parkinson's Cure

Not everyone with Parkinson's has an LRRK2 mutation; nor will everyone with the mutation get the disease. But it does increase the chance that Parkinson's will emerge sometime in the carrier's life to **between 30 and 75 percent**. (By comparison, the risk for an average American is about 1 percent.) Brin himself splits the difference and figures his DNA gives him about 50-50 odds.

Sergey Brin's Search for a Parkinson's Cure by Thomas Goetz
http://www.wired.com/2010/06/ff_sergeys_search/



Angelina Jolie: My Medical Choice

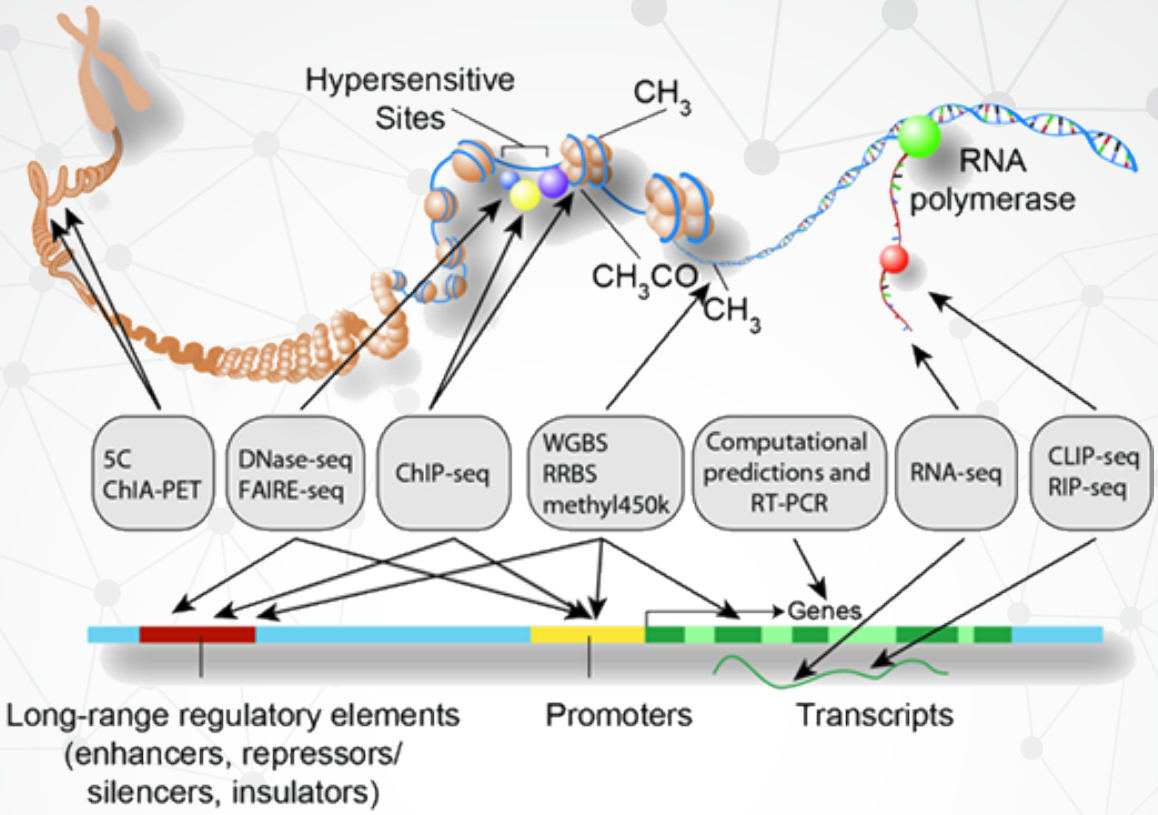
We often speak of "Mommy's mommy," and I find myself trying to explain the illness that took her away from us. They have asked if the same could happen to me. I have always told them not to worry, but the truth is I carry **a "faulty" gene, BRCA1**, which sharply increases my risk of developing breast cancer and ovarian cancer.

My Medical Choice By Angelina Jolie
http://www.nytimes.com/2013/05/14/opinion/my-medical-choice.html

Fields in Bioinformatics

Sequence Analysis

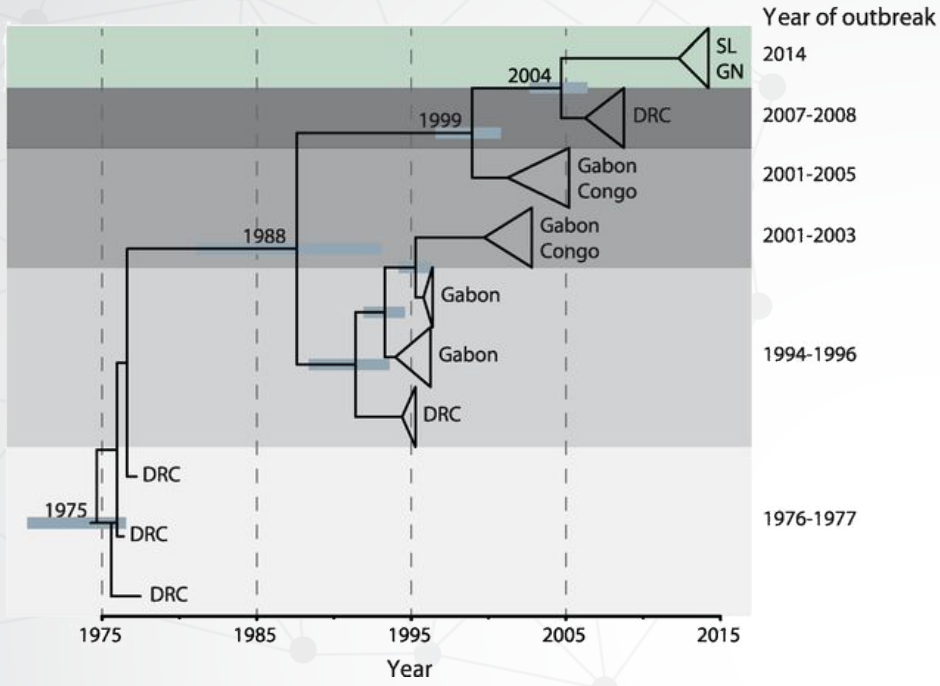
http://en.wikipedia.org/wiki/Bioinformatics by Alexbateman

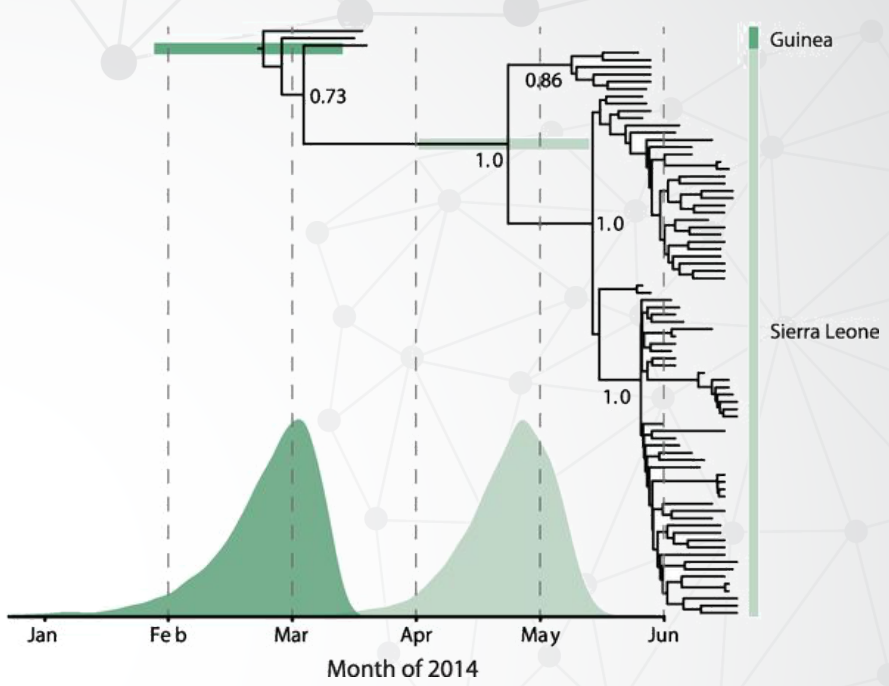


ENCODE: Encyclopedia of DNA Elements

The goal of ENCODE is to build a comprehensive parts **list of functional elements** in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

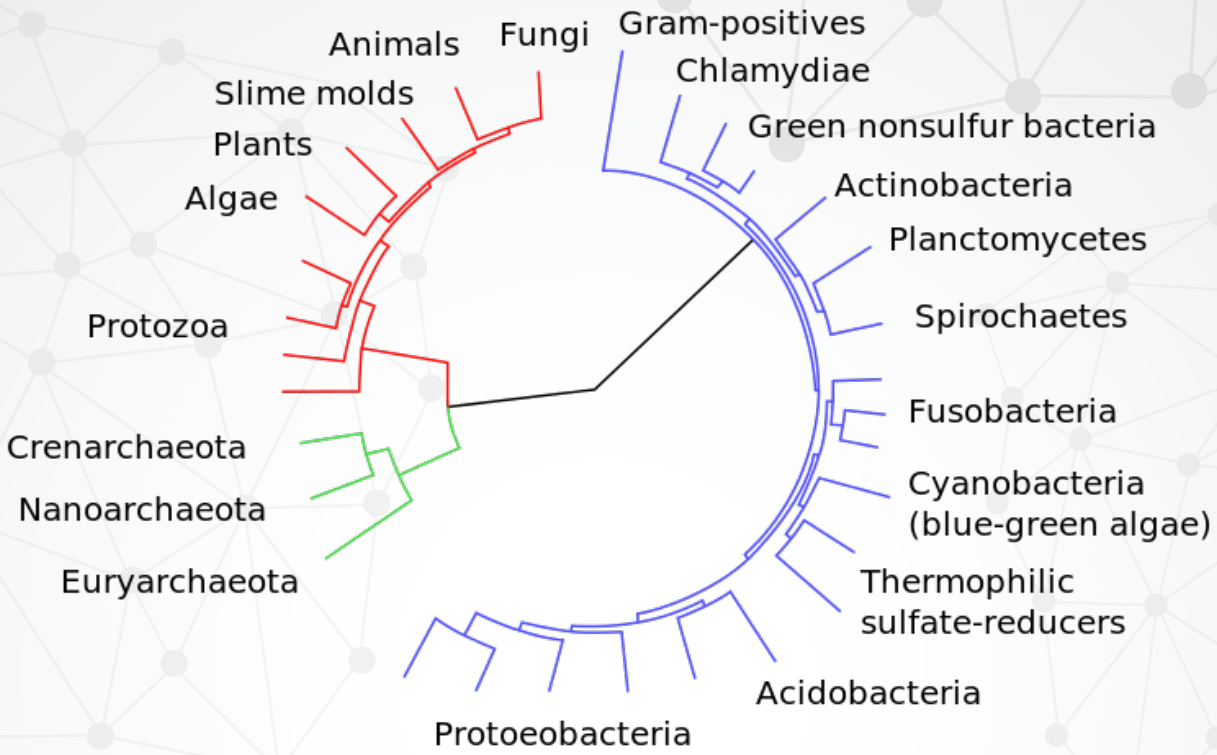
https://www.encodeproject.org





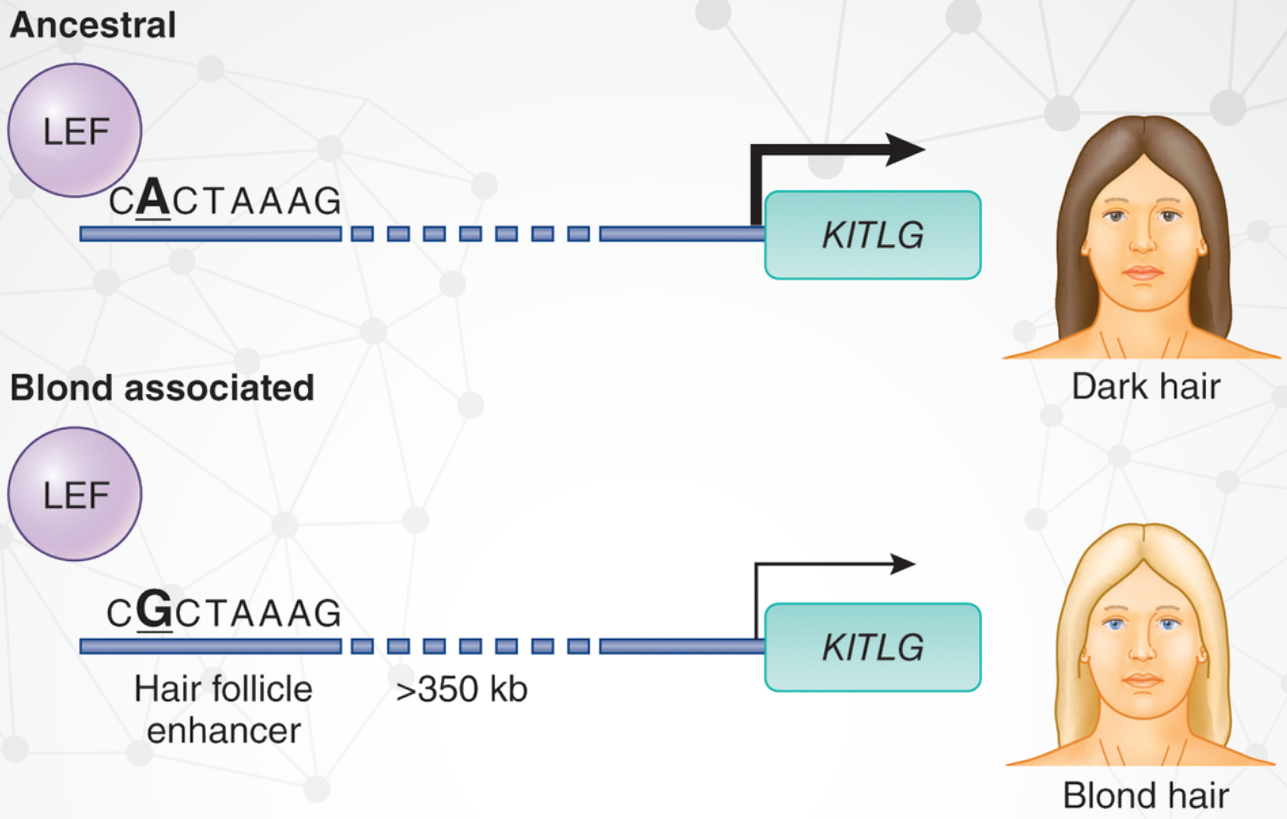
Computational Evolutionary Biology

*S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P.-P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti, "**Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak**," Science, vol. 345, no. 6202, pp. 1369–1372, Sep. 2014.*



Comparative Genomics

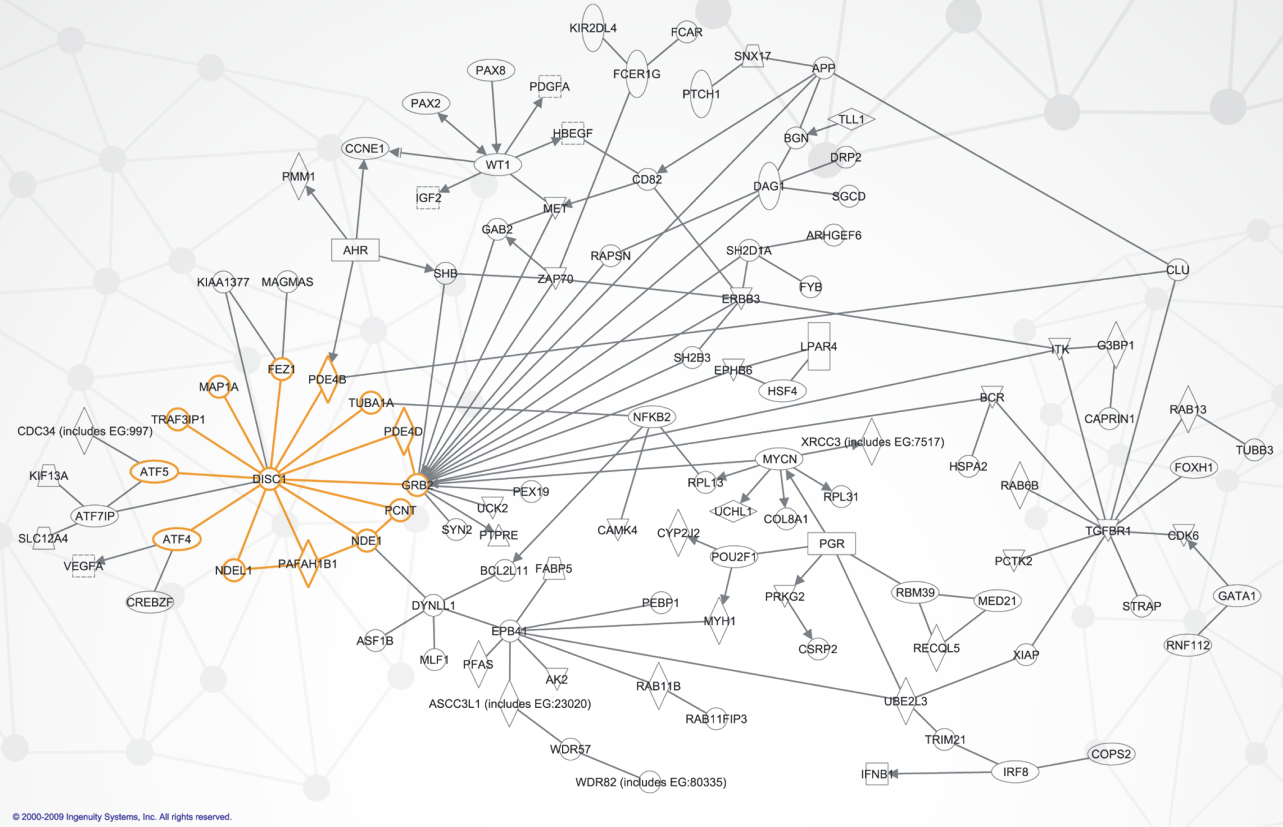
http://en.wikipedia.org/wiki/Bioinformatics Uploaded by Hgrosser



Genetics of Disease

The distal upstream hair follicle enhancer of KITLG contains a predicted LEF-binding site overlapping the blond-associated SNP (underlined nucleotide). LEF strongly binds the ancestral allele, whereas weak binding of the derived allele **reduces gene expression by ~20%**, resulting in lighter hair color.

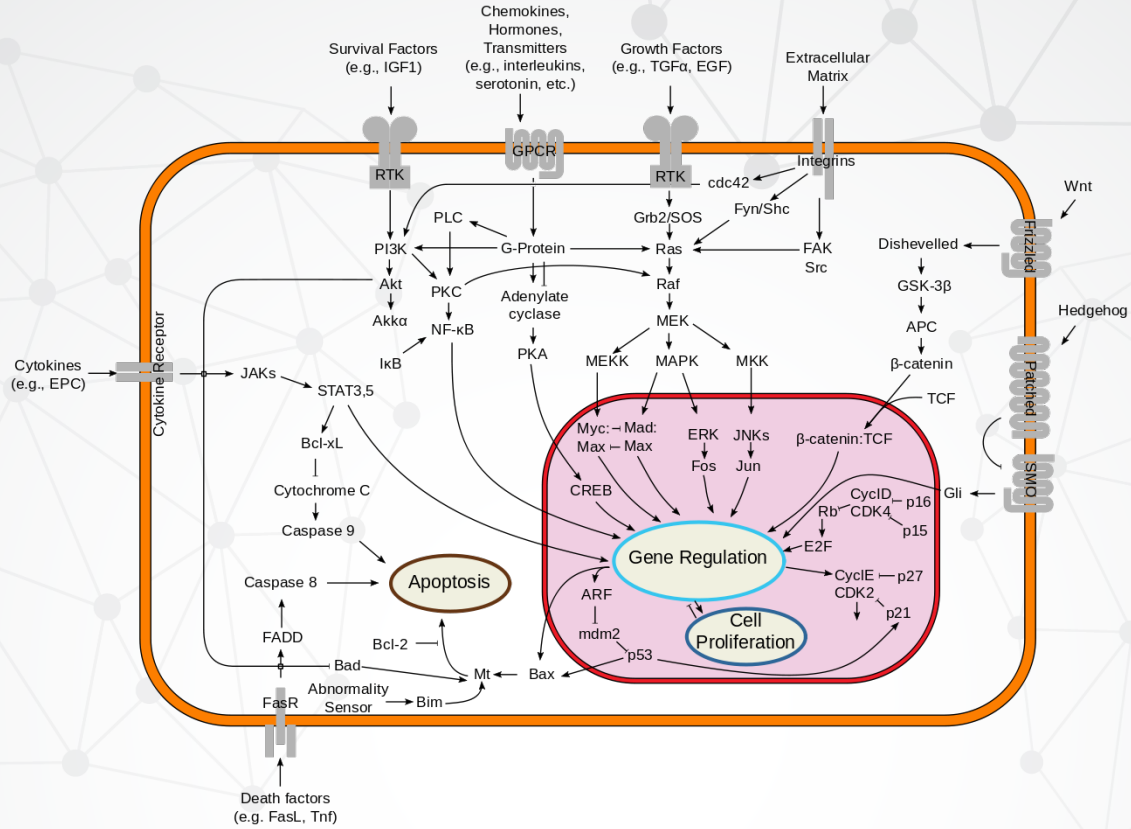
*H. Hoekstra, "**The secret of a natural blond**," Nat Genet, vol. 46, no. 7, pp. 660–661, Jul. 2014.*



Network Analysis

Network of how 100 of the 528 genes identified with significant differential expression relate to **DISC1** and its core interactors

*W. Hennah and D. Porteous, "**The DISC1 Pathway Modulates Expression of Neurodevelopmental, Synaptogenic and Sensory Perception Genes**," PLoS ONE, vol. 4, no. 3, p. e4906, Mar. 2009.*



Systems Biology

http://en.wikipedia.org/wiki/File:Signal_transduction_v1.png

Why Bioinformatics?

Why Bioinformatics?

4P in Medicine

- Predictive
- Personalized
- Preventive
- Participatory

*L. Hood and S. H. Friend, "**Predictive, personalized, preventive, participatory (P4) cancer medicine**," Nat Rev Clin Oncol, vol. 8, no. 3, pp. 184–187, Mar. 2011.*

Challenges

7V in Big Data on Human

- Volume (6 billion bases in human genome = 6GB file size)
- Variety (technologies, devices, file formats)
- Velocity (356 000 babies were born per day in 1997*)
- Veracity (>99%**)
- Validity (high)
- Volatility (DNA is very low but RNA is high)
- Value (very high up to someone life)

* http://www.who.int/whr/1998/media_centre/50facts/en/

** M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "***A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers***," *BMC Genomics*, vol. 13, pp. 341–341, Jul. 2012.

Overall Cost

- Device (USD 100K up to USD 1M*)
- Reagents
- Maintenance (electricity, vibration, humidity)
- Skilled People

* M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "***A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers***," *BMC Genomics*, vol. 13, pp. 341–341, Jul. 2012.

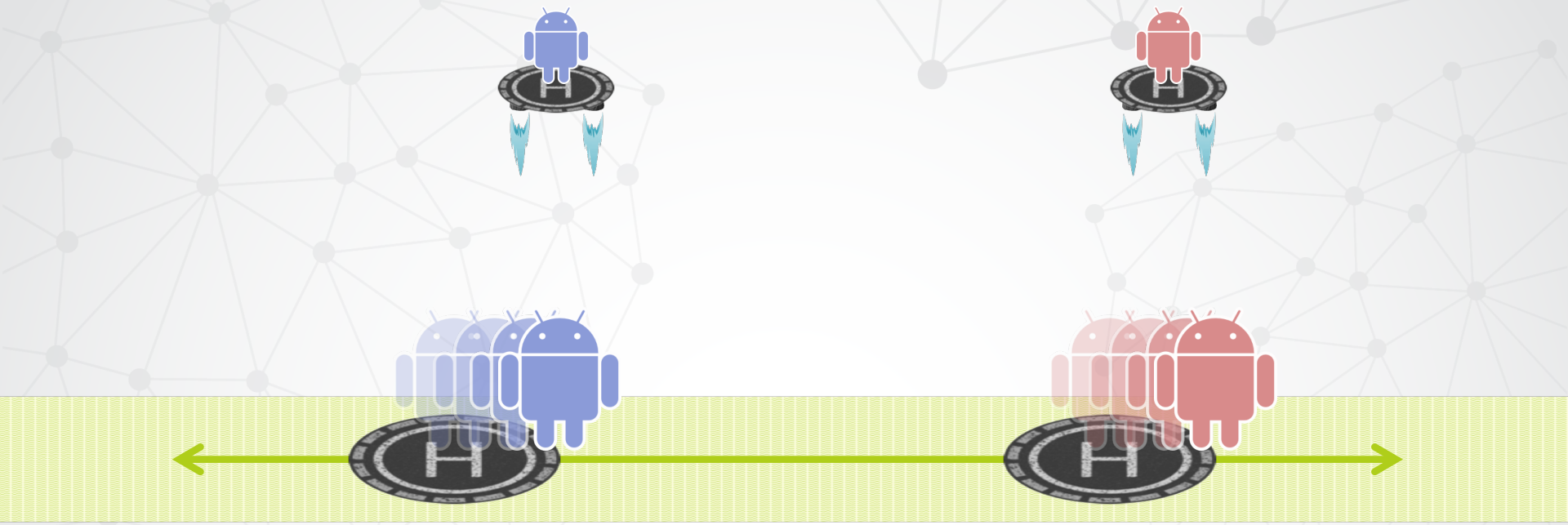
Encoding DNA

Travelling Robots Problem

- Two robots visit one dimension boundless planet separately
- They need to see each other when they arrive
- They have a same set of command and sensor
- Assume that they have a same landing time

Travelling Robots Problem

- Commands: Move Left, Move Right, Double Speed
- Sensor: Detect Landing Pad

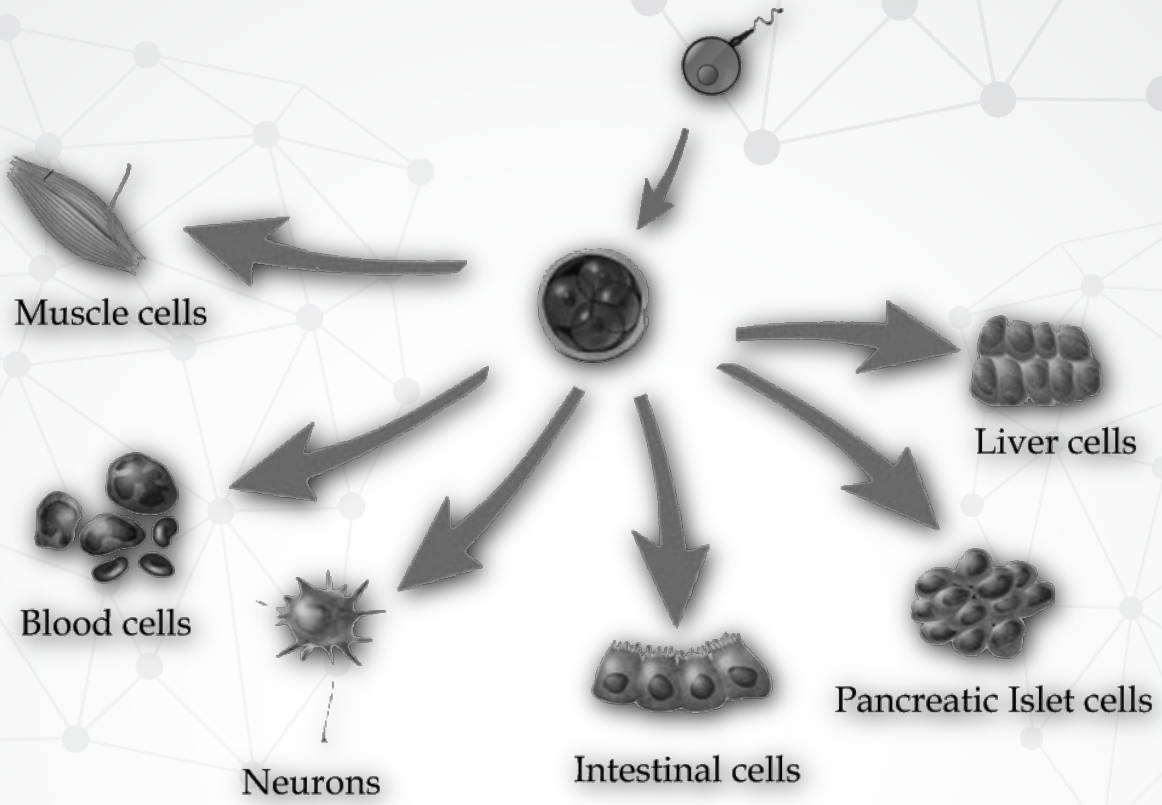


Travelling Robots Problem

- Commands: Move Left, Move Right, Double Speed
- Sensor: Detect Landing Pad
- Algorithm:

  After landing **Move Left**

  If **Detect Landing Pad** then **Double Speed**



Cell Differentiation

Jobs in Bioinformatics

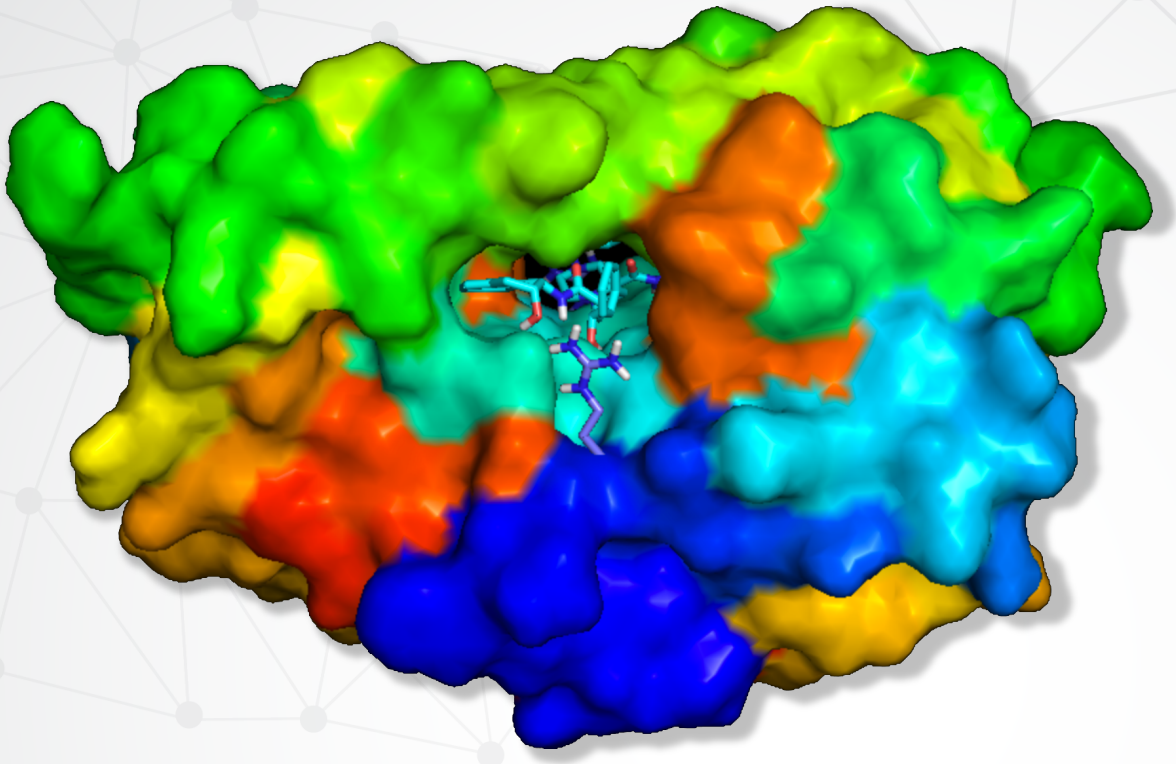
Jobs in Bioinformatics

- Bioinformatician
- Data Scientist
- Optimization, Simulation, Prediction
- Researcher, Lecturer

Bioinformatics Algorithms

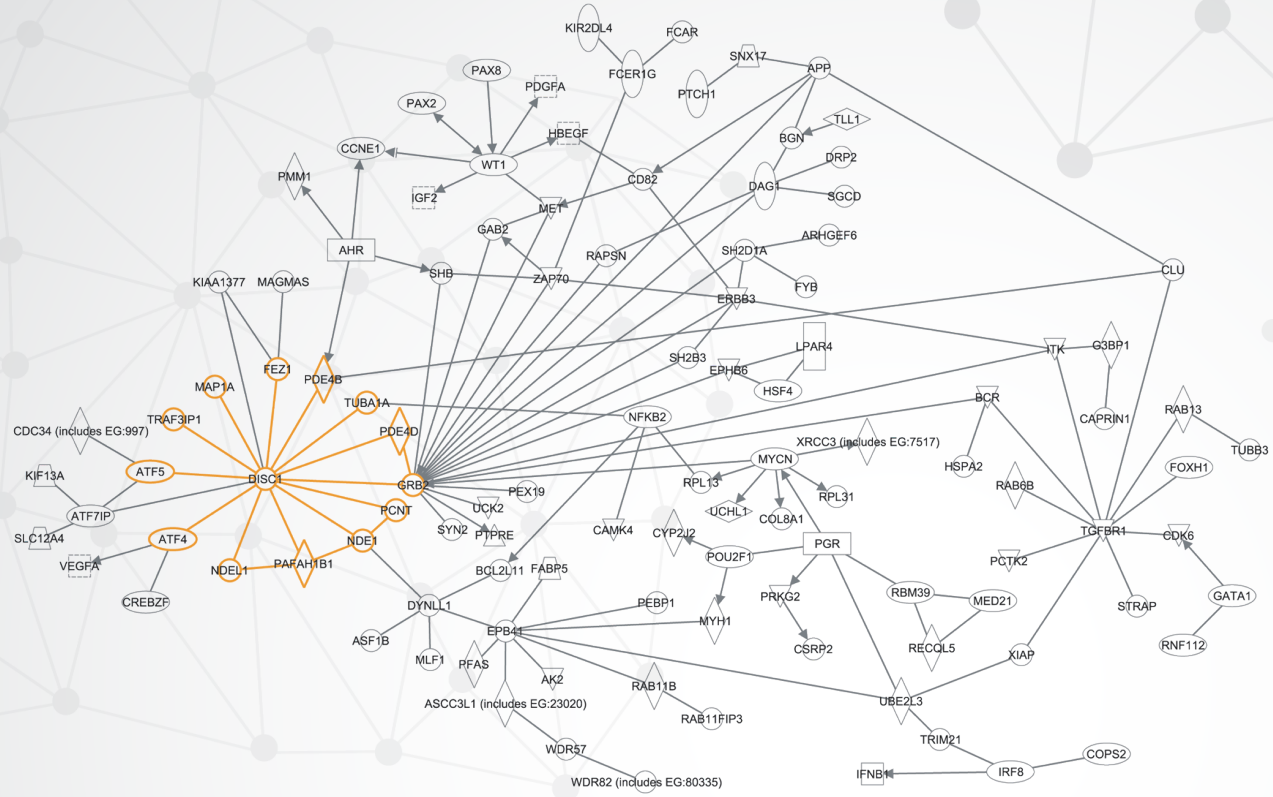
Brute force Algorithms

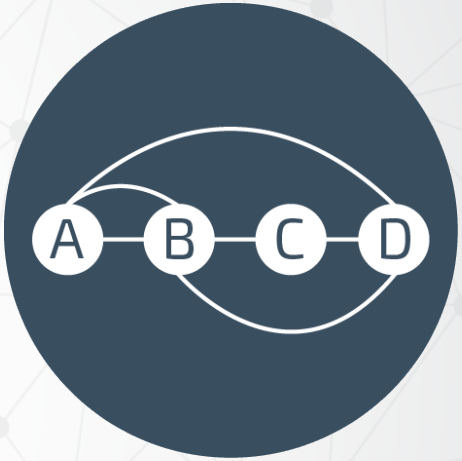
To sequence antibiotics



Greedy and Randomized Algorithms

To find DNA patterns that act as cellular clocks

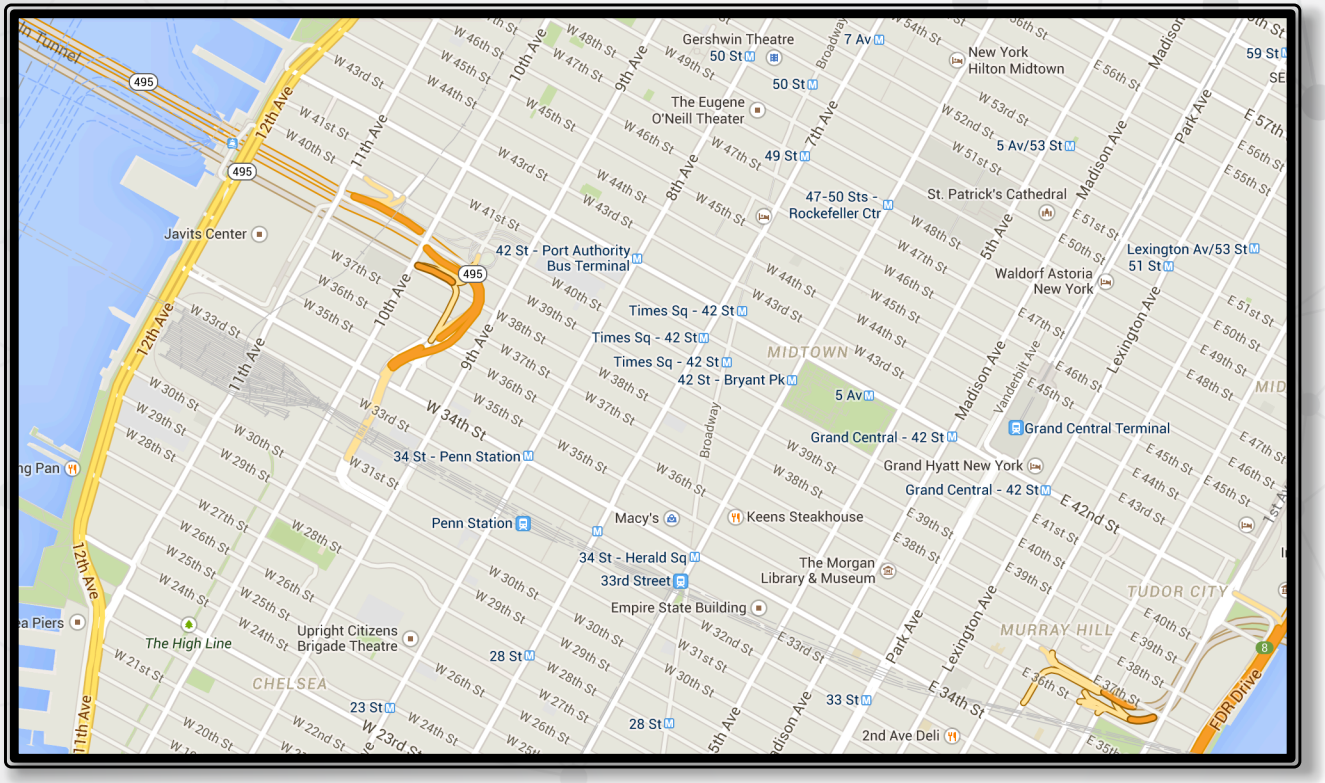




Graph Algorithms

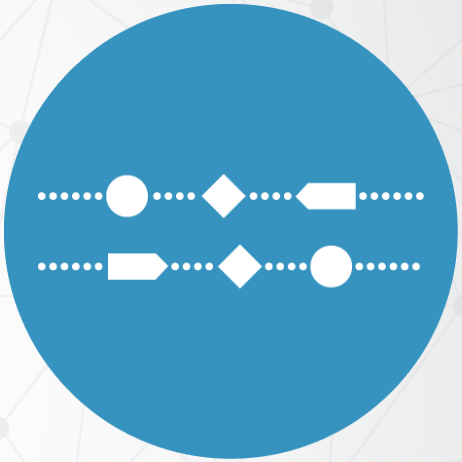
To assemble genomes, also useful to define path for mail carriers, garbage collectors and airplane pilots

*W. Hennah and D. Porteous, "**The DISC1 Pathway Modulates Expression of Neurodevelopmental, Synaptogenic and Sensory Perception Genes**," PLoS ONE, vol. 4, no. 3, p. e4906, Mar. 2009.*



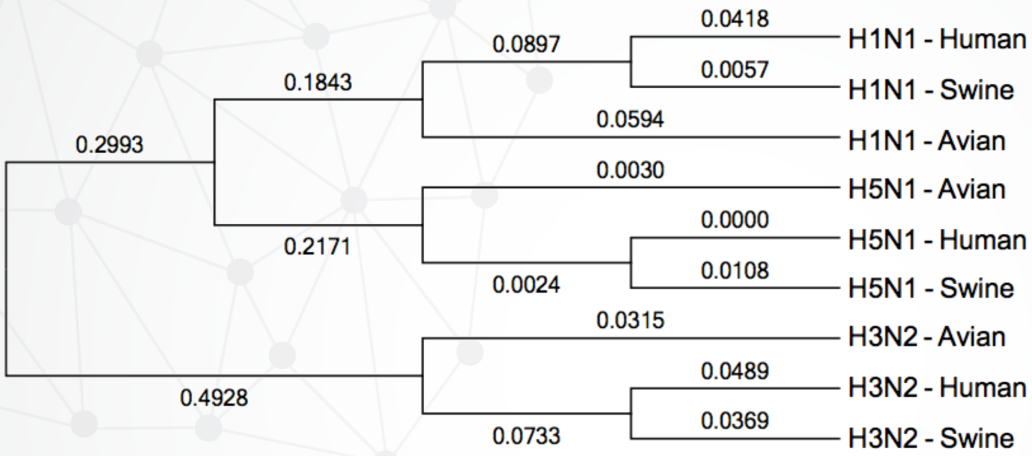
Dynamic Programming Algorithms

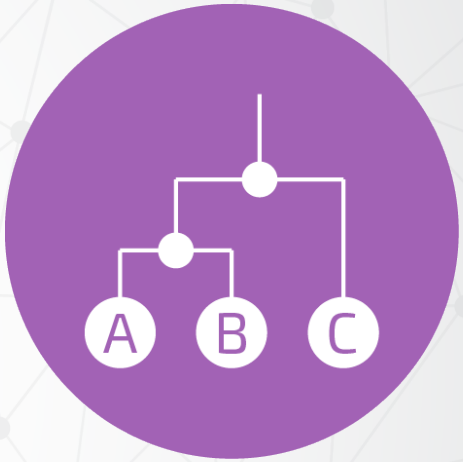
To compare DNAs, also for optimization problems like supply chain, managing storage and for tourist to enjoy maximum attraction



Combinatorial Algorithms

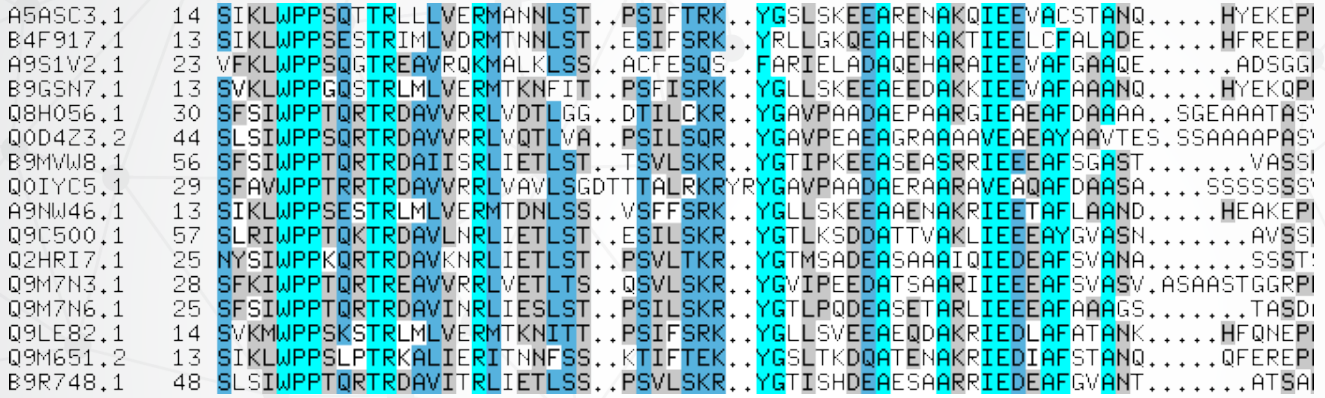
To find fragile regions in human genome, also many scheduling problem





Evolutionary Trees

To find the animal gives us SARS



Combinatorial Pattern Matching

To locate disease-causing mutation

http://en.wikipedia.org/wiki/Bioinformatics by Alexbateman

Clustering Algorithms

To find the reason Yeast is a good wine brewer

Q&A