Introduction to Bioinformatics

By Nick Giangreco, MA, PhD Candidate
Website LinkedIn Github Meetup
11/13/2018
NYC Medical Research and Bioinformatics Meetup

My bioinformatics trajectory

Biochemistry B.S.
- Computational structural biology
- Molecular modeling
- Proteomic software development

Cancer bioinformatics trainee
- NGS analysis
- Ovarian cancer

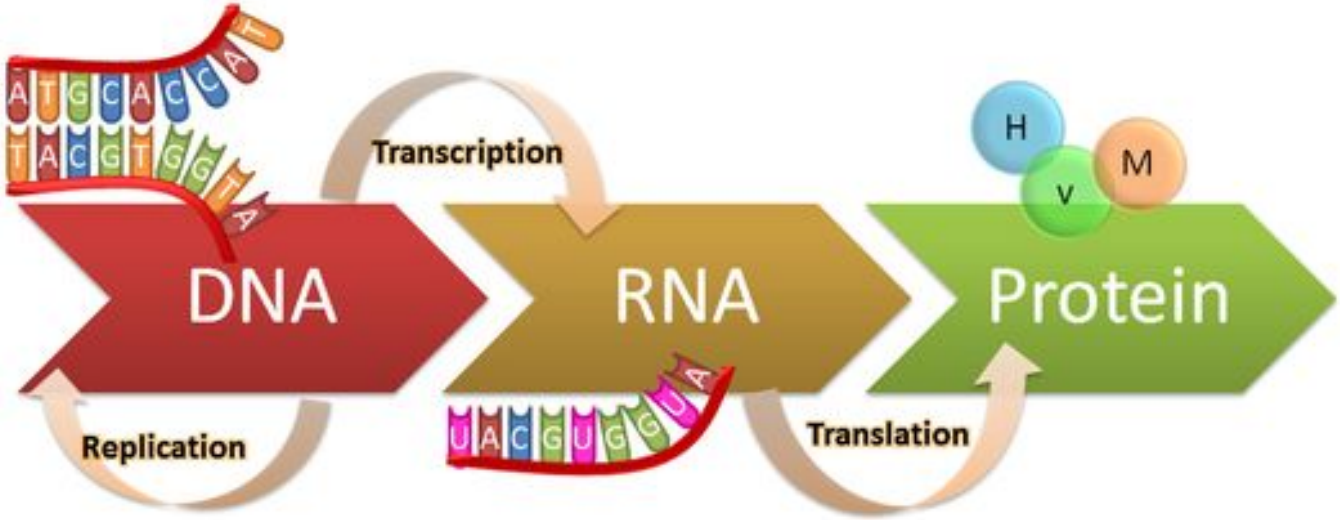
- Integrated Program in Cellular, Molecular, and Biomedical Studies
- Systems Biology PhD Candidate
- Tatonetti lab

- Current PhD project areas:
  - Translational bioinformatics
  - Clinical proteomics
  - Drug Safety in Children

Outline

- Brief history
- Topics overview
- Tools
- Data
- Application

Essential foundation for Bioinformatics: Central Dogma of Molecular Biology



Links:
[Central dogma video lectures]
[NCBI central dogma course module]
[Essay on Crick's discovery in PLoS]

Source:
https://genius.com/Biology-genius-the-central-dogma-annotated

Brief History of Bioinformatics

- Biological information processing

References

Brief history of bioinformatics

Bioinformatics-wikipedia page

The Roots of Bioinformatics in Theoretical Biology

Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment

Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment
- DNA sequencing and comparative genomics

Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment
- DNA sequencing and comparative genomics
- The computing revolution meets advances in biology

Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment
- DNA sequencing and comparative genomics
- The computing revolution meets advances in biology
- The Human Genome Project

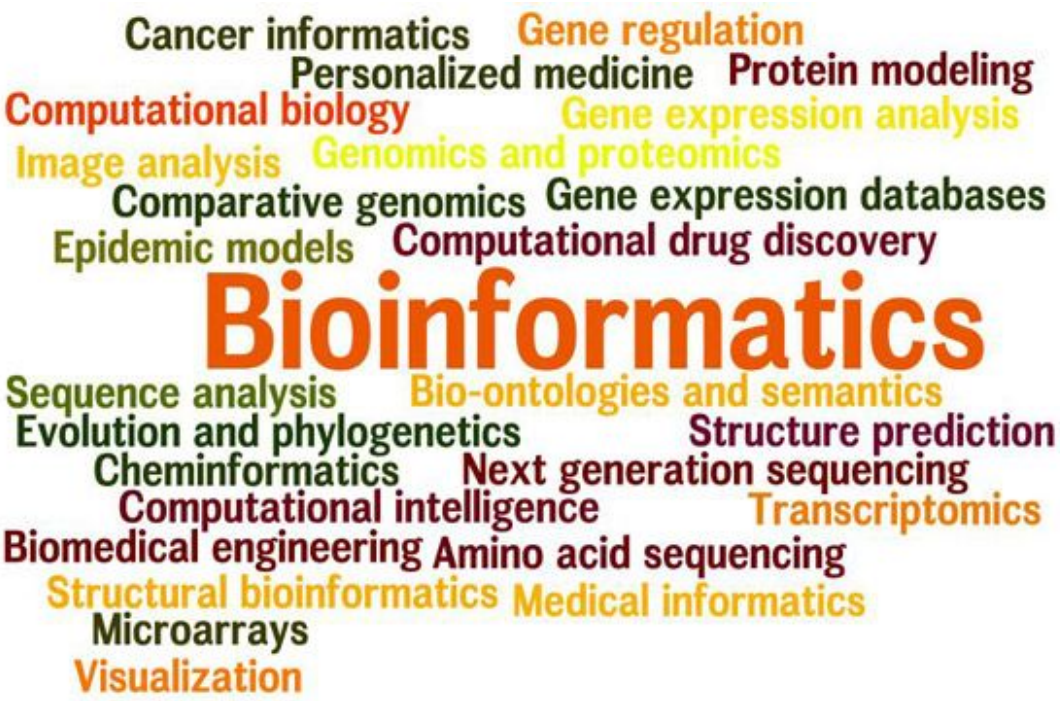
Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment
- DNA sequencing and comparative genomics
- The computing revolution meets advances in biology
- The Human Genome Project
- Acceleration of bioinformatics via the Internet

Brief History of Bioinformatics

- Biological information processing
- Margaret Dayhoff and protein sequence alignment
- DNA sequencing and comparative genomics
- The computing revolution meets advances in biology
- The Human Genome Project
- Acceleration of bioinformatics via the Internet
- High throughput sequencing yields biological big data

Bioinformatics topics overview



Open Bioinformatics Foundation

Source: http://www.erewise.com/current-affairs/what-is-bioinformatics_art5310866888028.html

Bioinformatics tools

Databases

- NCBI databases (Pubmed, OMIM, and friends; topic-centric)
- EMBL-EBI databases (Genomes, Proteins, etc.; entity-centric)

Bioinformatics tools

Databases

Command-line tools

- Variant Discovery - GATK
- Genome alignment - HISAT2
- Metagenomics sequence classification - Kraken
- Many, many more...

Bioinformatics tools

Databases

Command-line tools

Web tools

- Comprehensive list of bioinformatics tools - OMICTOOLS
- Interaction with functional genomics - HumanBase
- Many resources - Ma'ayan Lab, Mt. Sinai

Bioinformatics tools

Databases

Command-line tools

Web tools

Programming Languages

- Java
  - BioJava
- Python
  - BioPython
- R
  - Bioconductor

Bioinformatics tools

Databases

Command-line tools

Web tools

Programming Languages

APIs

- NCBI database utilities - API & Python wrapper
- Food & Drug Administration data - openFDA
- Awesome Bioinformatics - github repository of many things...

Open datasets

- GigaDB - Open datasets from the open-access journal GigaScience
- NCBI data + software
- Biomedical + clinical datasets at Kaggle
- Brain datatypes from human and mouse @ Allen Brain Institute
- Google dataset search
- Many, many more...

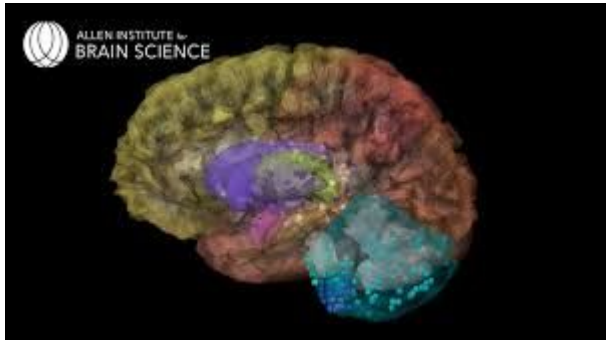
An example bioinformatics project:

Characterization of tissue specific gene expression differences in the human adult brain

Github Papers

Data

- Gene expression from 414 brain structures



- Gene expression from Alzheimers disease brain tissue

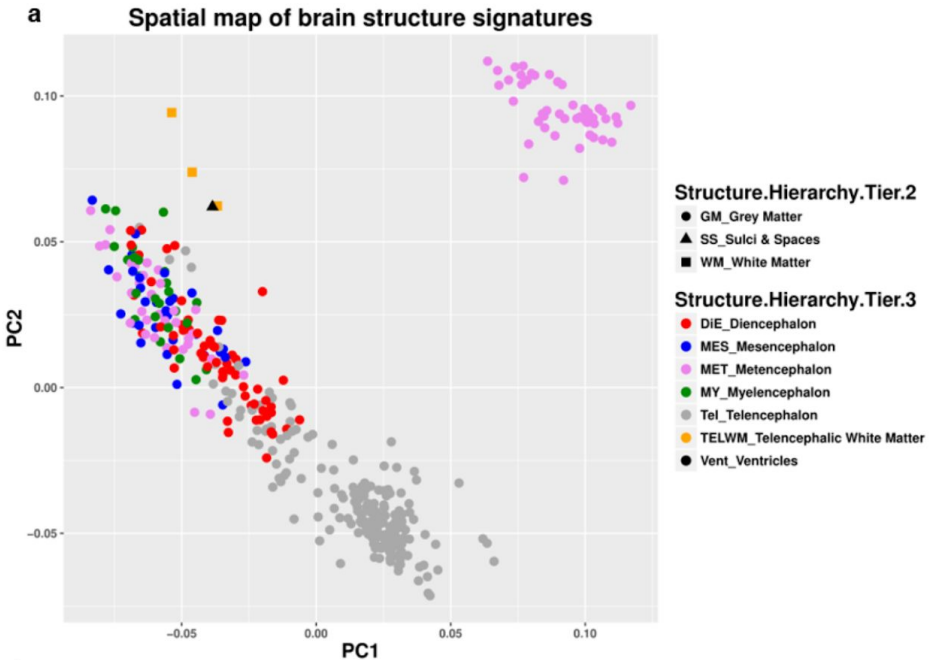
Gene Expression Omnibus

Research project findings

- Association of gene expression with embryonic origin of tissue
- Pathways with highly variable gene expression, compared to low variability, seem to indicate tissue specific gene expression.
- The cerebral cortex harbors highly variable pathways seen in Alzheimers disease
- Distinct expression profile of metancephelon-originated structures, including the pons and cerebellum

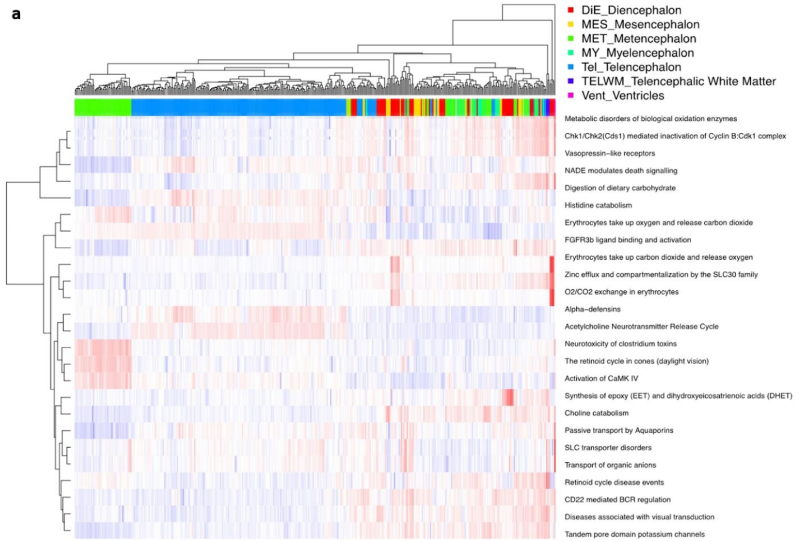
Research methods

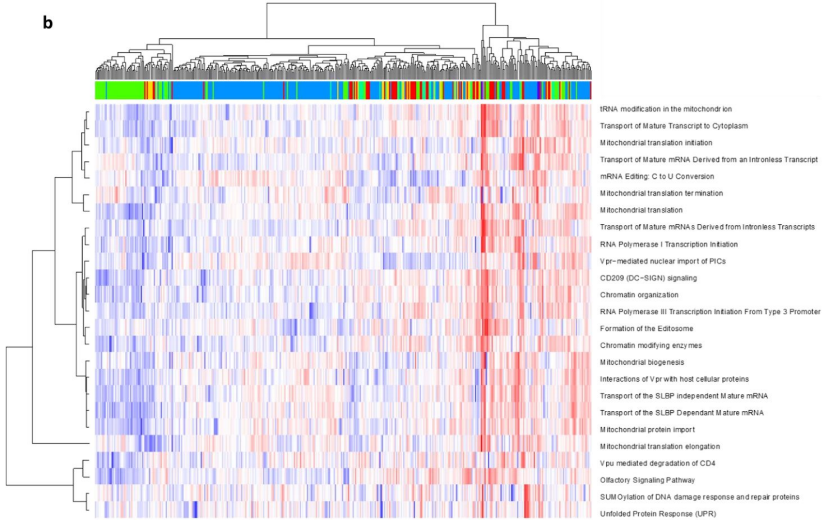
- Association of gene expression with embryonic origin of tissue
  - Methods:
    - Principal components analysis
      - Software 1 2
      - Tutorial 1 2
    - Independent component analysis
      - Software 1 2
      - Tutorial 1 2
    - Dimensionality Reduction Resources



Research methods

- Association of gene expression with embryonic origin of tissue
- Pathways with highly variable gene expression, compared to low variability, seem to indicate tissue specific gene expression.
  - Methods:
    - Differential expression analysis 1 2
    - Visual Check: Hierarchical Clustering





Research methods

- Association of gene expression with embryonic origin of tissue
- Pathways with highly variable gene expression, compared to low variability, seem to indicate tissue specific gene expression.
- The cerebral cortex harbors highly variable pathways seen in Alzheimers disease
  - Methods:
    - Enrichment (association) tests
    - Reactome pathway analysis

Top 8 Enriched Reactome Pathways in the Alzheimer's dataset	Corrected p-values
Immune System	4.37e-07
Gene Expression	4.83e-06
Signaling by Rho GTPases	8.04e-05
Signaling by NGF	1.61e-04
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	1.68e-04
Cellular responses to stress	1.68e-04
Innate Immune System	5.78e-04

Table 1. Top 8 enriched Reactome pathways in Alzheimer's GEO dataset, ranked by order of significance.

Brain structure category	Structures Harboring Enriched High-Variance Pathways	P-value
Cerebral Cortex	39.7%	2.0E-03
Cerebellum	13.2%	3.2E-2
Cerebral Nuclei	9.9%	1.4E-2
Pons	8.8%	0.95
Hypothalamus	7.2%	6.6E-4

Table 2. Proportion of structures in each brain category harboring high-variance enriched pathways.

Parting thoughts

- The bioinformatics community is very open
  - Open access Journals 1 2
  - Open datasets (see slides for links)
  - APIs
  - Databases
  - Tutorials
  - Many more resources
- Innovative methods can promote new biological knowledge
- Tailor method to biologically motivated question/hypothesis
- We need stronger communication between biologists and informaticists
- Course materials for Stanford Intro to Bioinformatics course

Thank you!

Keep in touch! Email