

Genomic Data Analysis

CSC 315-01

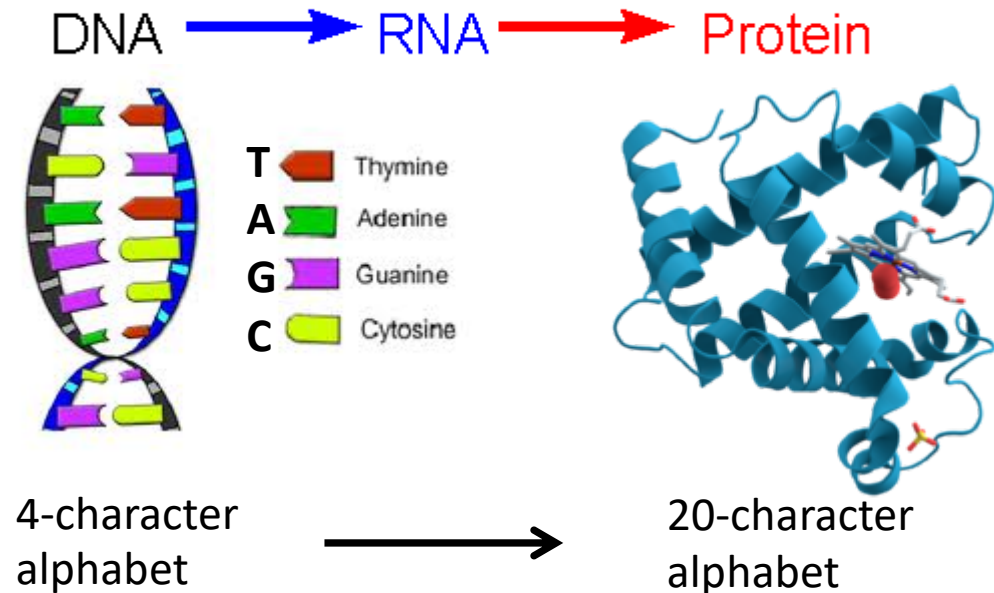
Dr. Garrett Dancik
Fall 2021

Course Notes: <https://gdancik.github.io>

What is bioinformatics

- Bioinformatics:

- Biology + information
- the study and utilization of methods for storing, retrieving and analyzing biological data
- **This class: gene expression data**



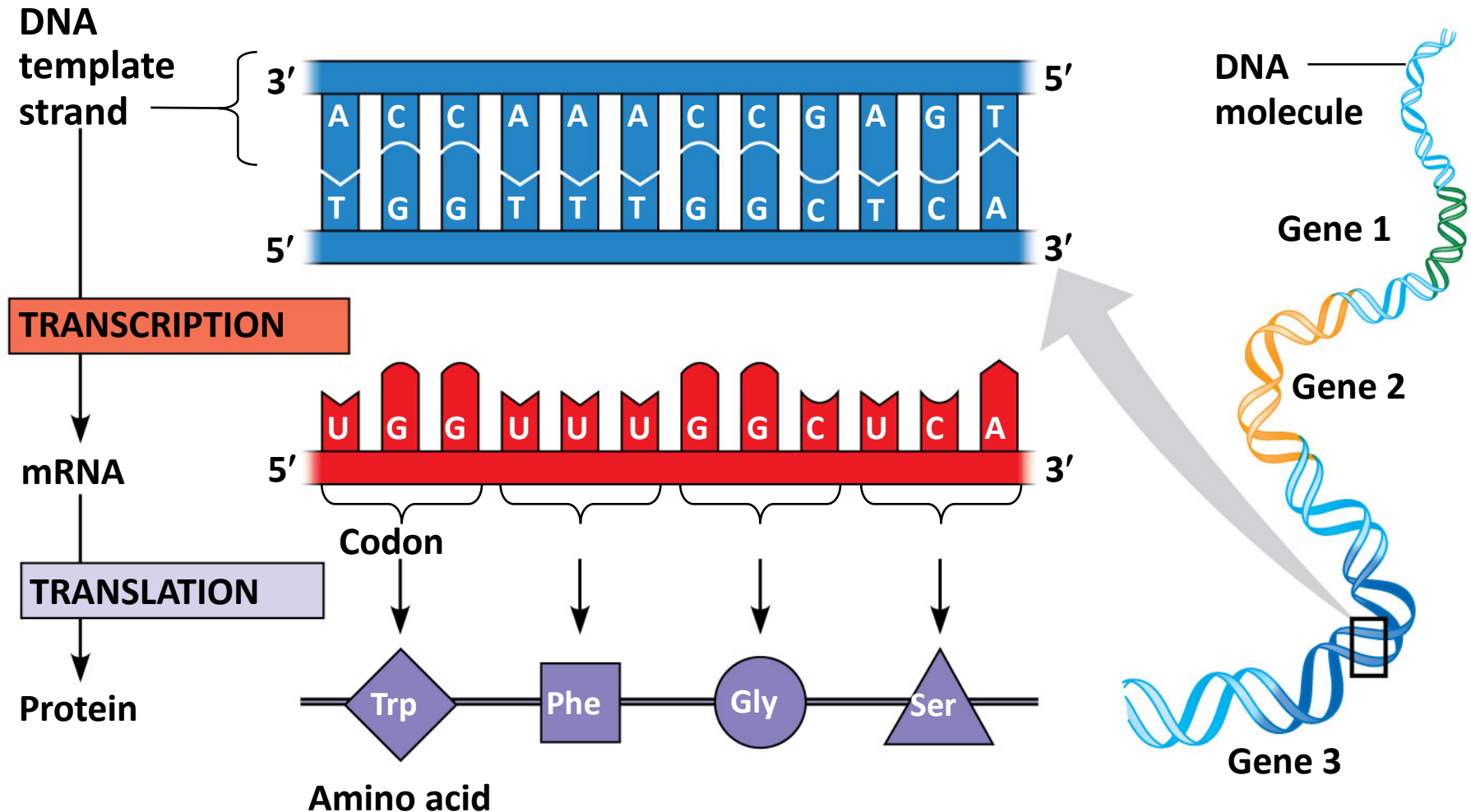
- How much information:

- Human genome: 3 billion nucleotides
- ~20,000 genes
 - many more when considering “junk DNA” and alternative splicing
- >10 million sites of DNA variation
- Countless possible interactions between DNA, RNA, and proteins

Intro to Genetics and Gene Expression

- What are genes?
 - http://www.youtube.com/watch?v=ubq4eu_TDFc

Gene expression: the production of a functional gene product from DNA



Why is bioinformatics important?

- A lot of research in genetics and molecular biology is driven by genomic sequencing and gene expression analysis
 - What does this gene or protein do?
 - What genes are associated with a condition?
- Personalized medicine, based on genomic or gene expression data
 - Is a patient at risk for a particular disease
 - Does a patient with a disease have a good/poor outcome
 - Is a patient with a disease likely to respond to a particular treatment

The New England Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 347

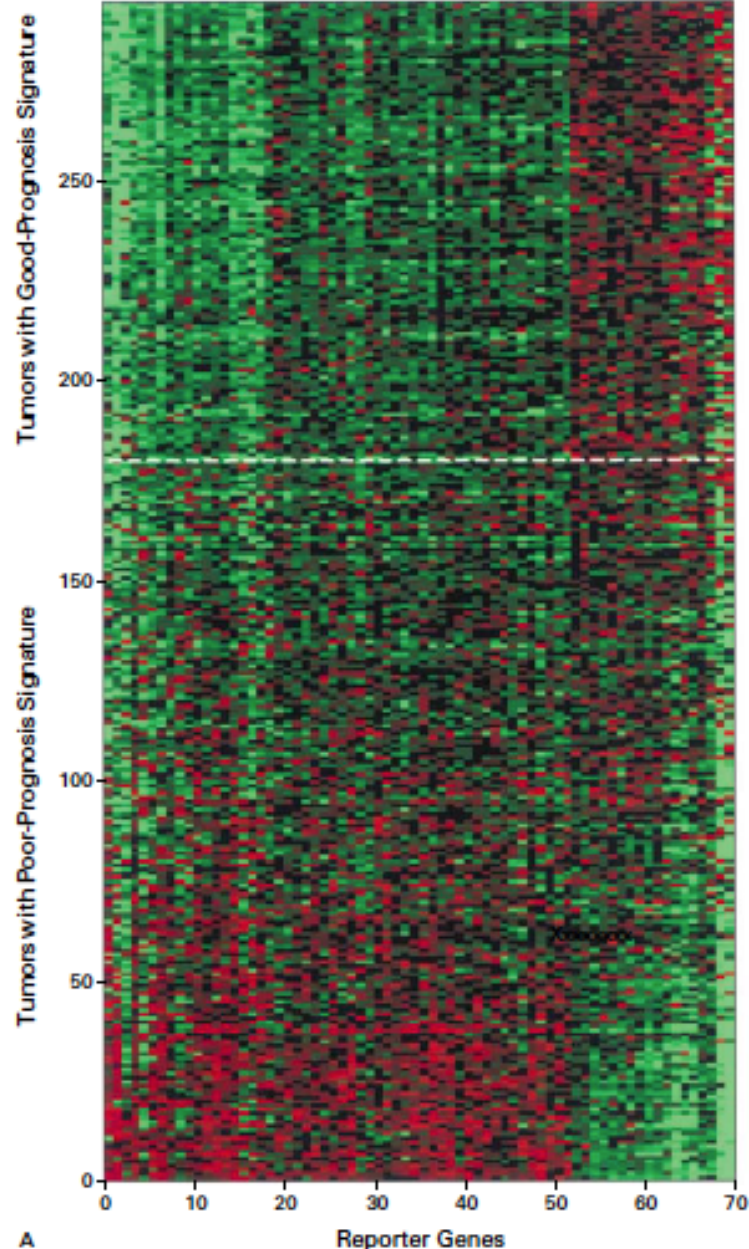
DECEMBER 19, 2002

NUMBER 25

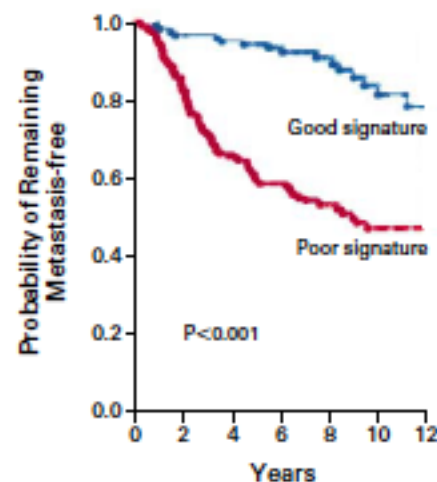


A GENE-EXPRESSION SIGNATURE AS A PREDICTOR OF SURVIVAL IN BREAST CANCER

MARC J. VAN DE VIJVER, M.D., PH.D., YUDONG D. HE, PH.D., LAURA J. VAN 'T VEER, PH.D., HONGYUE DAI, PH.D.,
AUGUSTINUS A.M. HART, M.SC., DORIEN W. VOSKUIL, PH.D., GEORGE J. SCHREIBER, M.SC., JOHANNES L. PETERSE, M.D.,
CHRIS ROBERTS, PH.D., MATTHEW J. MARTON, PH.D., MARK PARRISH, DOUWE AT SMA, ANKE WITTEVEEN,
ANNUSKA GLAS, PH.D., LEONIE DELAHAYE, TONY VAN DER VELDE, HARRY BARTELINK, M.D., PH.D.,
SJOERD RODENHUIS, M.D., PH.D., EMIEL T. RUTGERS, M.D., PH.D., STEPHEN H. FRIEND, M.D., PH.D.,
AND RENÉ BERNARDS, PH.D.



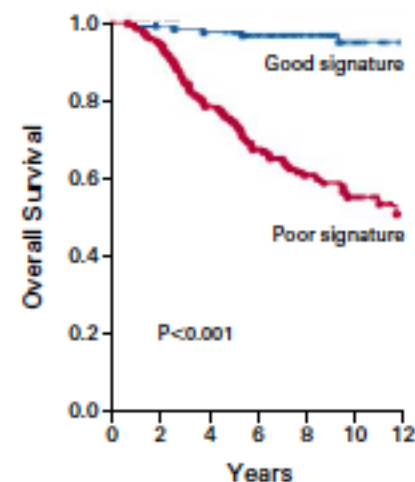
A All Patients



NO. AT RISK

Good signature	115	111	107	87	59	36	19
Poor signature	180	146	111	84	52	33	17

B All Patients



NO. AT RISK

Low risk	115	114	112	91	65	43	23
High risk	180	167	134	100	62	40	19

nature

International weekly journal of science

Published online 7 February 2007 | Nature | doi:10.1038/news070205-

News

Genetic test gets approval

United States gives thumbs up to breast cancer prognostic.

Bioinformatics is Interdisciplinary

*course objectives

Calculus

2. Data analysis

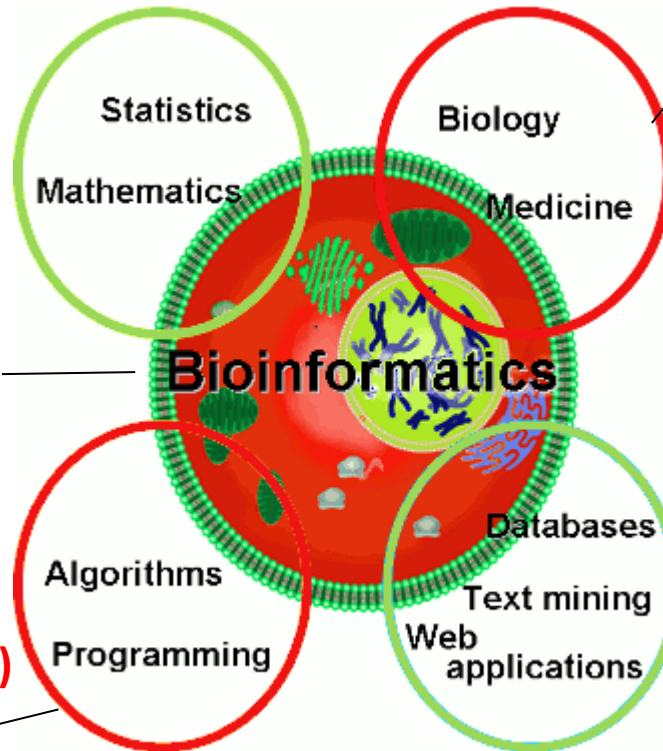
3. Biology, Genetics

Chemistry
Physics

4. Genomic data analysis

1. Computer programming (R)

Computer algorithms



Database design

Data analysis more generally

- Very large datasets are being generated by scientific studies, technology, and commerce
- From technology (2017)
 - Overview: <https://web-assets.domo.com/blog/wp-content/uploads/2020/08/20-data-never-sleeps-8-final-01-Resize.jpg>
 - Facebook (<https://www.brandwatch.com/blog/47-facebook-statistics/>)
 - Generates 4 petabytes of data each day
 - Averages 350 million photo uploads per day
 - Users generate 4 million likes every minute
 - <https://www.internetlivestats.com/>
- Bioinformatics: human genome sequencing experiments requires ~150 gigabytes per individual

Data Analysis Examples

- What does FB know about you?
 - <http://www.nbcnews.com/science/gay-conservative-high-iq-your-facebook-likes-can-reveal-traits-1C8805606>
- What does Target know about you?
 - <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>
- What information does your genome reveal about you?
 - Sex
 - Paternity and ancestry
 - Eye color
 - Whether you are lactose intolerant
 - Relative Risk of breast cancer
 - Relative Risk of alcoholism

(Note that many genetic relationships are complex, not well understood, can be difficult to interpret, and can be influenced by environmental factors)

R programming language

- R (<http://www.r-project.org>) is a free environment for statistical computing and graphics
- R is an interpreted programming language
- Many *packages* are available for specialized analyses (<http://cran.r-project.org/web/packages/>)
- Bioconductor (<http://www.bioconductor.org>) provides packages for the analysis of genomic (and other biological) data
- We will use Rstudio, an IDE for R: <http://www.rstudio.com>
- We will go through R examples in class, but to learn more on your own, check out <http://swirlstats.com>