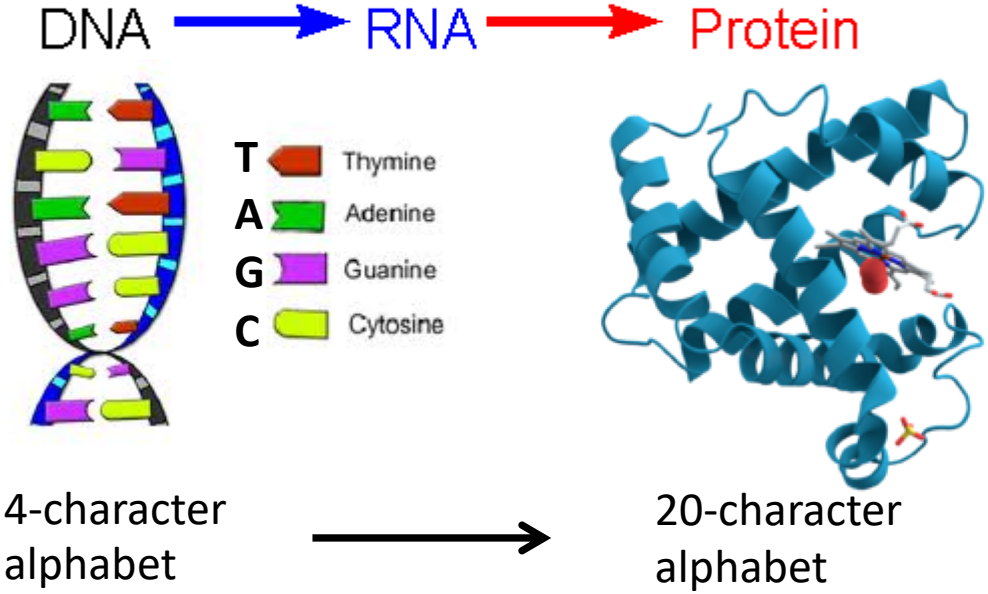
Bioinformatics Programming and Analysis
CSC 315-01

Dr. Garrett Dancik

What is bioinformatics

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

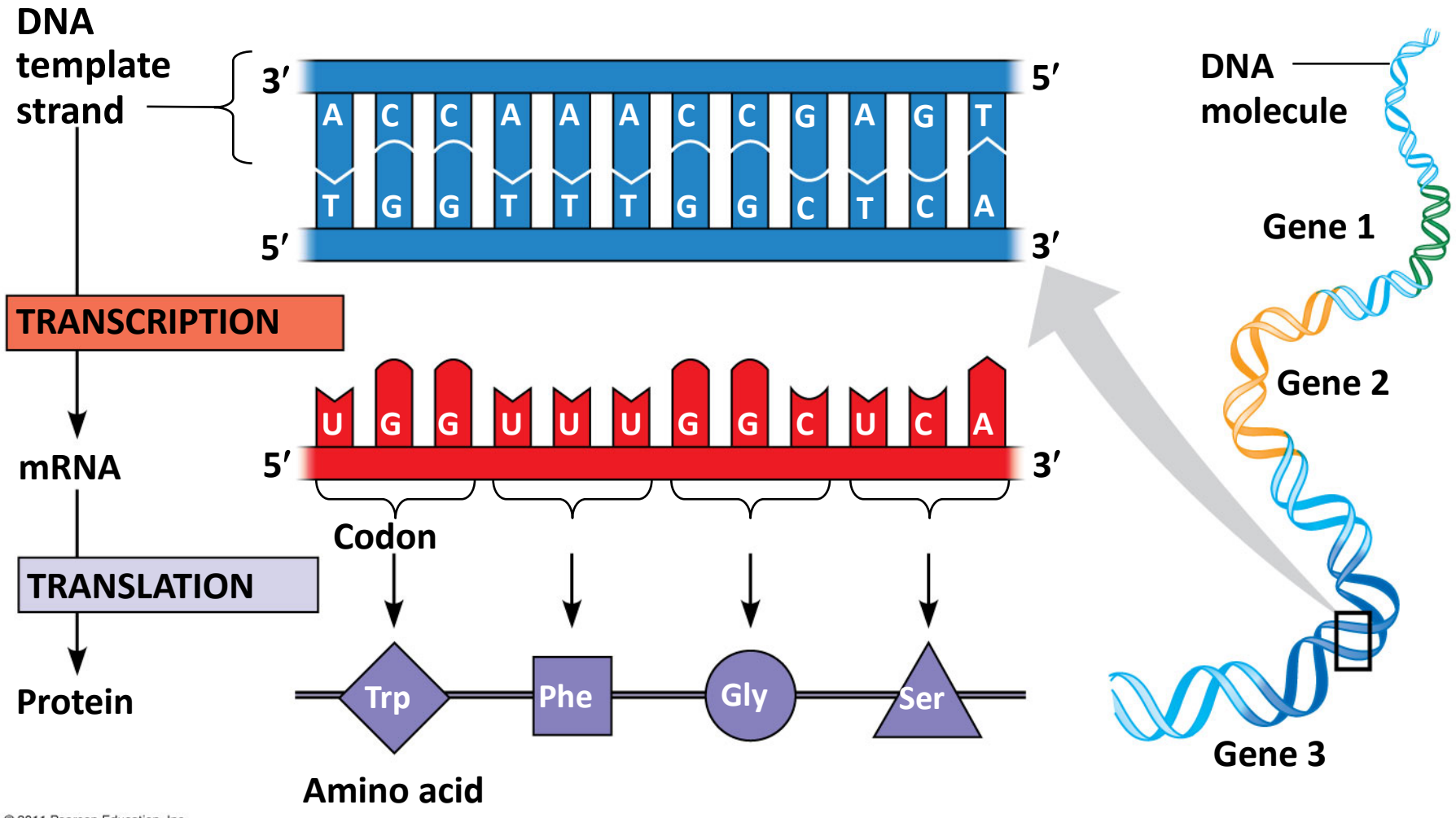


- How much information:
  - Human genome: 3 billion nucleotides
  - ~20,000 genes
    - many more when considering “junk DNA” and alternative splicing
  - >10 million sites of DNA variation
  - Countless possible interactions between DNA, RNA, and proteins

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?

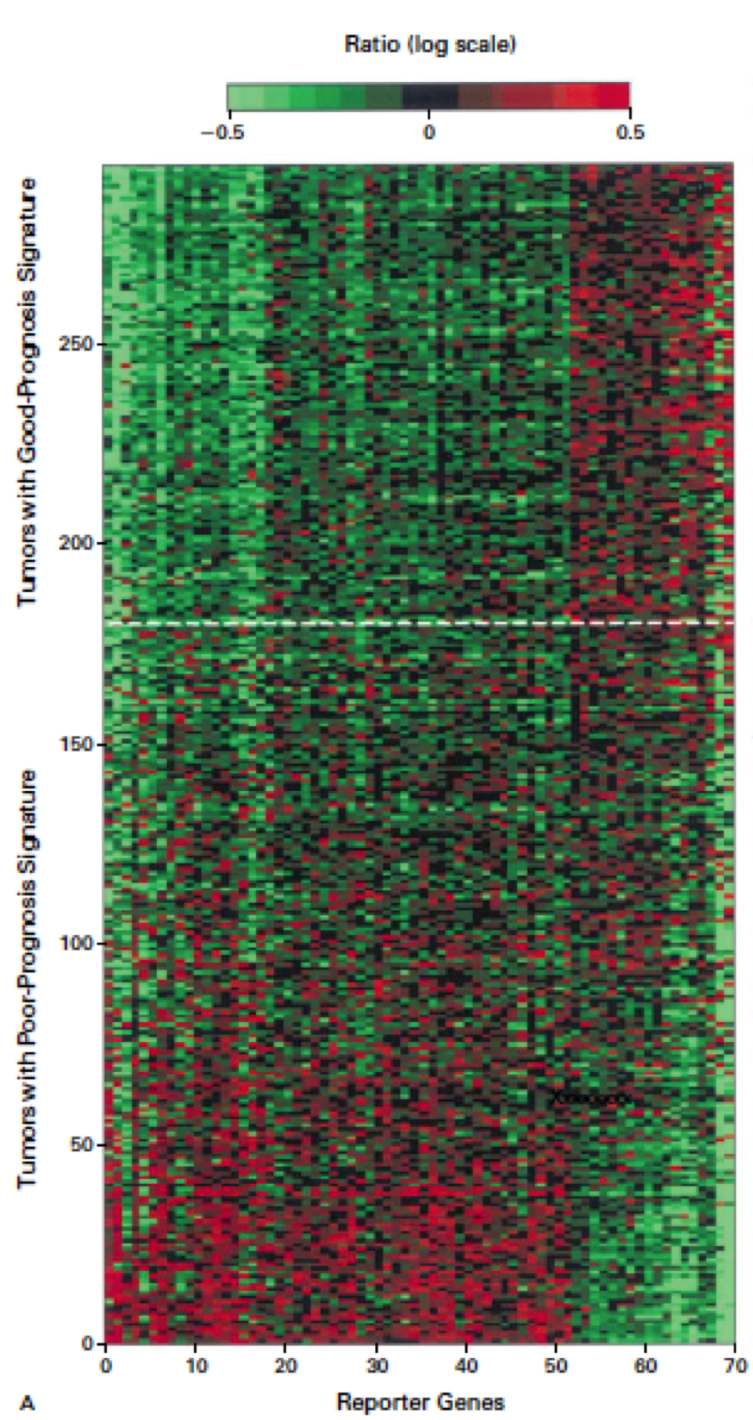
- Basic research in genetics and molecular biology depends on genomic sequencing and gene expression analysis
  - What does this gene or protein do?
- Personalized medicine, based on genomic or gene expression data
  - Is a patient at risk for a particular disease
  - Does a patient with a disease have a good/poor outcome
  - Is a patient with a disease likely to respond to a particular treatment

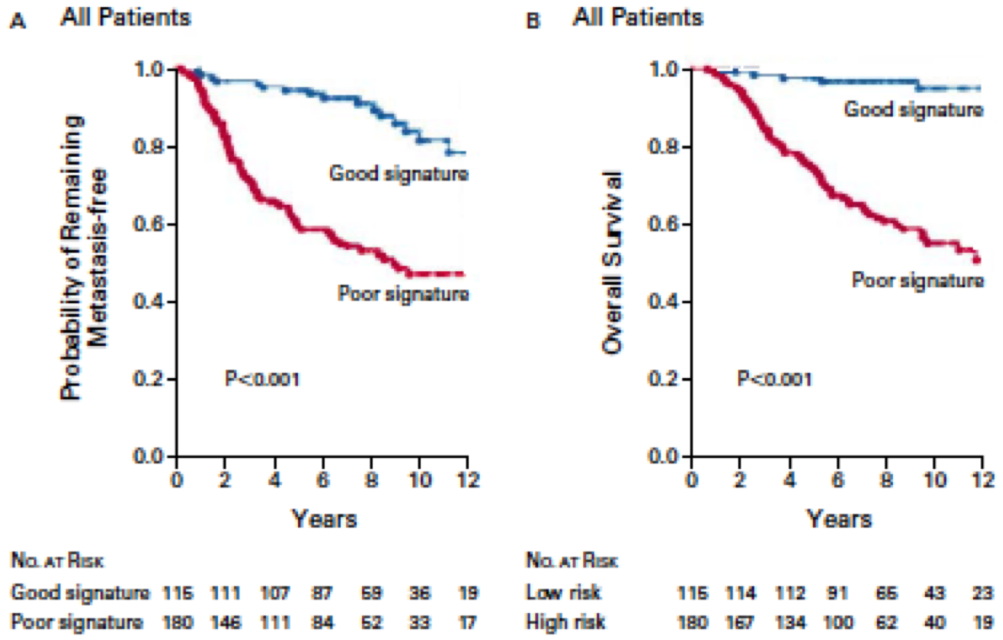
The New England
Journal of Medicine

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 ATSMA, 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.





nature International weekly journal of science

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

News

Genetic test gets approval

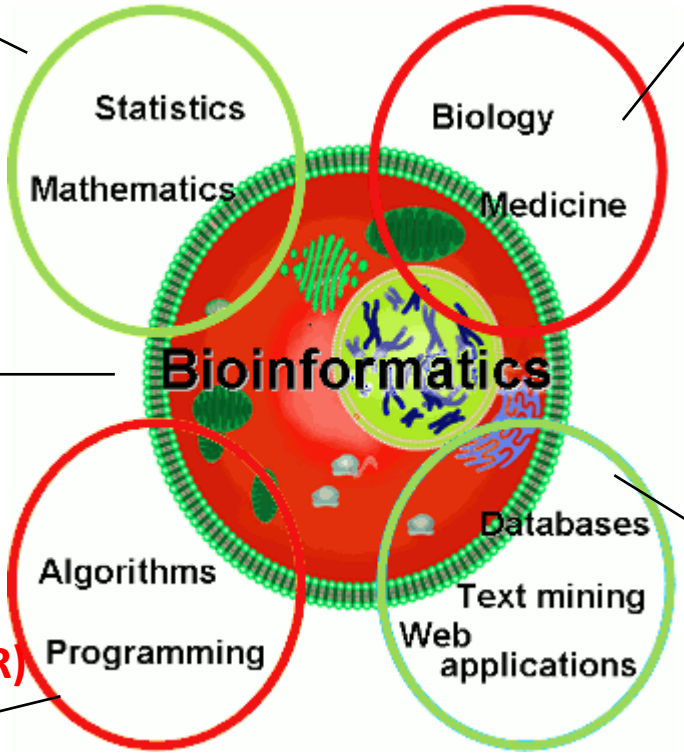
United States gives thumbs up to breast cancer prognostic.

Bioinformatics is Interdisciplinary

*course objectives

Calculus

2. Data analysis



3. Biology, Genetics

Chemistry

Physics

Database design

1. Computer programming (R)

Computer algorithms

4. Bioinformatics Programming (Gene expression analysis)

5. Gene sequencing analysis (time permitting)

Source: http://www.amberbio.com/

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/2017/07/17_domo_data-never-sleeps-5-01.png
  - 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
- Bioinformatics: genome sequencing experiment requires ~200 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 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 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