TCSS 588
Bioinformatics

Ka Yee Yeung
kayee@uw.edu
Institute of Technology, UW-Tacoma
3/26/2018

1

Objectives of this course

- Statistical computing in R
  - Expect a steep learning curve initially…
  - Why R?
  - R tools for visual studio: https://blogs.msdn.microsoft.com/visualstudio/2016/03/22/introducing-r-tools-for-visual-studio-3/
- Jupyter notebooks
  - An open source web application that allows you to create and share documents that contain live code
- Please install Jupyter notebook with the R kernel on your laptop before the next class
- Machine learning methods
- Applications in biology
- Working with biological data
- Experience contributing to crowd sourcing projects in bioinformatics

2

Bioinformatics (aka Computational Biology)

- What is bioinformatics?
  - An *interdisciplinary* field that develops methods and software tools for understanding biological data (source: wikipedia)

3

Applications of Bioinformatics

- Study biological processes in organisms
- Determine how these processes go wrong in diseases
- Discover and develop drugs to treat, cure and prevent diseases

4

An Explosion in Bioinformatics Career (Science Career Advice 2014)

- "…high demand for talented, experienced professionals at the crossroads of biology, statistics, and computer science."
- "Scientists who can analyze large amounts of information and present it in a clear manner to decision makers are finding the sky is the limit in terms of jobs and career pathways".
- "Bring your expertise to health care," says Telthorst, "and you'll know you're going to make a difference, at the patient level and at the societal level."

http://sciencecareers.sciencemag.org/career_magazine/previous_issues/articles/2014_06_13/science.opms.r1400143

5

Spotlight on Bioinformatics (NatureJobs 2016)

- https://www.nature.com/naturejobs/science/articles/10.1038/nj0478
- 2 paths to careers in bioinformatics:
  - Computer scientists must become fluent in the life science terminology of genetics, genomics and cellular biology.
  - Biologists must pick up skills in data analysis, including statistics, logic and programming.
- The skill set needed by a bioinformatician continues to evolve.

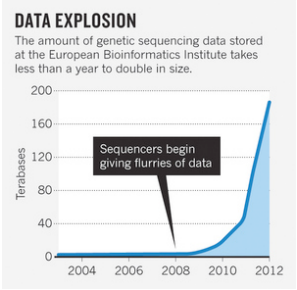
SKILLS SPECTRUM

There are **three** essential skill sets bioinformaticians need. Here's where to start.

1. COMMAND	2. PROGRAM	3. DATA
Understand how **Unix** commands work.	Learn **Python**, a basic language. Then consider **R**, a useful language for handling statistics.	Understanding what type of data is in different kinds of **databases**, and how to mine it, is essential. Learning relational database techniques is another **plus**.

6

Big Data in Biology
(reference: Marx 2013)

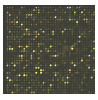


http://www.nature.com/nature/journal/v498/n7453/full/498255a.html 7

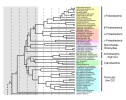
NIH Big Data to Knowledge (BD2K)

- An initiative launched by NIH in 2012.
- Addresses challenges in using biomedical big data:
  - Locating data and software tools.
  - Getting access to the data and software tools.
  - Standardizing data and metadata.
  - Extending policies and practices for data and software sharing.
  - Organizing, managing, and processing biomedical Big Data.
  - ***Developing new methods for analyzing & integrating biomedical data.***
  - ***Training*** researchers who can use biomedical Big Data effectively.

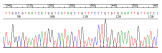
http://bd2k.nih.gov/about_bd2k.html#sthash.xs2j0lpi.dpbs 8



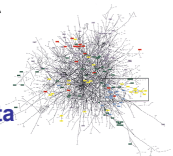
"High-Throughput BioTech"



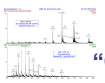
Sensors
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



Controls
- Cloning
- Gene knock out/knock in
- RNAi



Floods **of data**



"Grand Challenge" problems

Courtesy: Larry Ruzzo 9

How much biology do I need to know to work on these data?

- Short answer: the more the better
- You are all on the right track for being in this class.
- My take:
  - We can focus on the computational aspects of the problems. Just learn enough jargon to get started.
  - Adopt the lazy algorithm to learn more biology as we go.

10

A VERY quick intro to molecular biology

http://www.yourgenome.org/video/from-dna-to-protein

Check out related content as well.

11

The Genome

- The hereditary info present in every cell DNA molecule -- a long sequence of *nucleotides* (A, C, T, G)
- Human genome -- about 3 x 10^9 nucleotides
- The genome project -- extract & interpret genomic information, apply to genetics of disease, better understand evolution, ...

Courtesy: Larry Ruzzo

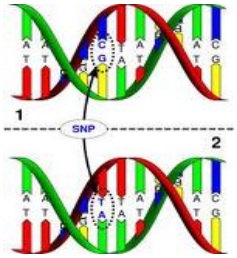
12

DNA - Sequence

```
.....acctc ctgtgcaaga acatgaaaca cctgtggttc ttccttctcc
tggtggcagc tcccagatgg gtcctgtccc aggtgcacct gcaggagtcg
ggcccaggac tggggaagcc tccagagctc aaaaccccac ttggtgacac
aactcacaca tgcccacggt gcccagagcc caaatcttgt gacacacctc
ccccgtgccc acggtgccca gagcccaaat cttgtgacac acctccccca
tgcccacggt gcccagagcc caaatcttgt gacacacctc ccccgtgccc
ccggtgccca gcacctgaac tcttgggagg accgtcagtc ttcctcttcc
ccccaaaacc caaggatacc cttatgattt cccggacccc tgaggtcacg
tgcgtggtgg tggacgtgag ccacgaagac cccgaggtcc agttcaagtg
gtacgtggac ggcgtggagg tgcataatgc caagacaaag ctgcgggagg
agcagtacaa cagcacgttc cgtgtggtca gcgtcctcac cgtcctgcac
caggactggc tgaacggcaa ggagtacaag tgcaaggtct ccaacaaagc
aaccaagtca gcctgacctg cctggtcaaa ggcttctacc ccagcgacat
cgccgtggag tgggagagca atgggcagcc ggagaacaac tacaacacca
cgcctcccat gctggactcc gacggctcct tcttcctcta cagcaagctc
accgtggaca agagcaggtg gcagcagggg aacatcttct catgctccgt
gatgcatgag gctctgcaca accgctacac gcagaagagc ctctc.....
```


SNP

A **single-nucleotide polymorphism** (**SNP**, pronounced *snip*) is a DNA sequence variation occurring when a single nucleotide — A, T,C, or G — in the genome (or other shared sequence) differs between members of a species or paired chromosomes in an individual.

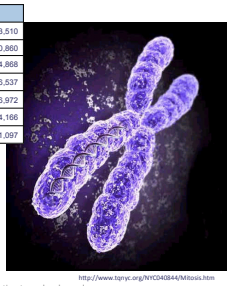


Chromosome

A *chromosome* is a very long, continuous piece of DNA, which contains many genes, regulatory elements and other intervening nucleotide sequences.

Chrom.	Genes	Bases
1	2968	245,203,898
2	2288	243,315,028
3	2032	199,411,731
4	1297	191,610,523
5	1643	180,967,295
6	1963	170,740,541
7	1443	158,431,299
8	1127	145,908,738
9	1299	134,505,819
10	1440	135,480,874
11	2093	134,978,784
12	1652	133,464,434
13	748	114,151,656
14	1098	105,311,216
15	1122	100,114,055
16	1098	89,995,999
17	1576	81,691,216

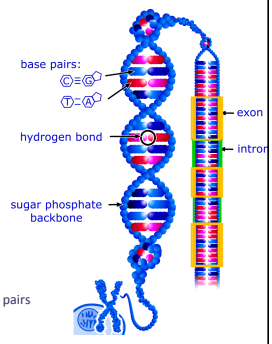
Chrom.	Genes	Bases
18	766	77,753,510
19	1454	63,790,860
20	927	63,644,868
21	303	46,976,537
22	288	49,476,972
X	1184	152,634,166
Y	231	50,961,097



http://www.tqnyc.org/NYC040844/Mitosis.htm

DNA - Deoxyribonucleic acid

- Role as carrier of genetic information
- Deoxyribonucleic acid (DNA) forms a double stranded helix.
- The two strands of DNA run in opposite directions.
- Bases face towards each other and form hydrogen bonds
- carries the generic instructions (genes)



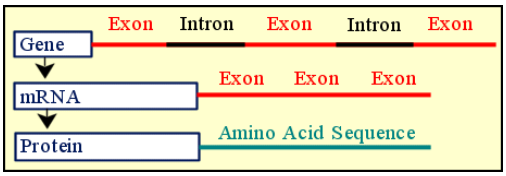
free Bases
Cytosine - C
Guanine - G
Adenine - A
Thymine - T
complementary base pairs

Ulf Schmitz, Introduction to molecular and cell biology

16

Introns vs. Exons

An **exon** is any nucleotide sequence encoded by a gene.



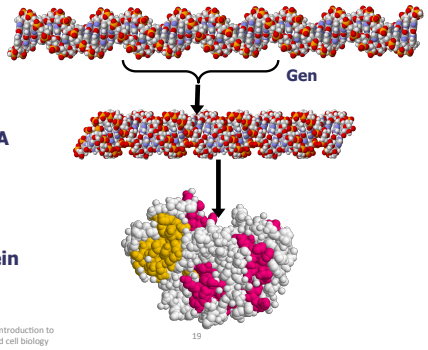
17

The "Central Dogma"

- Genes encode proteins.
- DNA transcribed into messenger RNA (mRNA) translated into proteins

18

From Genes to Proteins

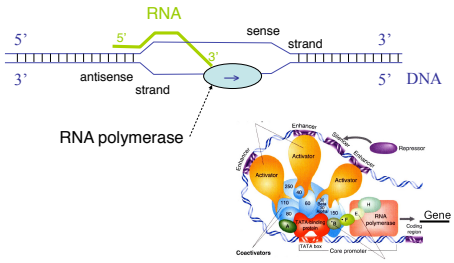


RNA – Ribonucleic acid

- **Messenger RNA** (**mRNA**) is transcribed directly from a gene's DNA and is used to encode proteins.
- RNA is almost exclusively found in the single-stranded form.

Bases:
Cytosine - C
Guanine - G
Adenine - A
Uracil - U

Transcription: DNA → RNA

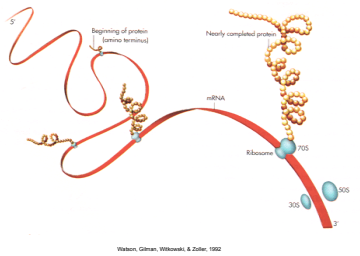


Gene Expression - Transcription

Messenger RNA (mRNA)
Messenger RNA is RNA that carries information from DNA.

Non-coding RNA or "RNA genes"
RNA genes (sometimes referred to as non-coding RNA or small RNA) are genes that encode RNA that is not translated into a protein. The most prominent examples of RNA genes are transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are involved in the process of translation.

Translation: mRNA → Protein



Watson, Gilman, Witkowski, & Zoller, 1992

Proteins

Proteins have a variety of roles that they must fulfil:

1. they are the enzymes that rearrange chemical bonds.
2. they carry signals to and from the outside of the cell, and within the cell.
3. they transport small molecules.
4. they form many of the cellular structures.
5. they regulate cell processes, turning them on and off and controlling their rates.

Proteins – Amino Acids

- there are 20 different types of amino acids (see below).
- different sequences of amino acids *fold* into different 3-D shapes.
- Proteins can range from fewer than 20 to more than 5000 amino acids in length.
- Each protein that an organism can produce is encoded in a piece of the DNA called a "gene".
- Humans are believed to have about 20,000 different genes (the exact number as yet unresolved),

Gene Expression - Translation

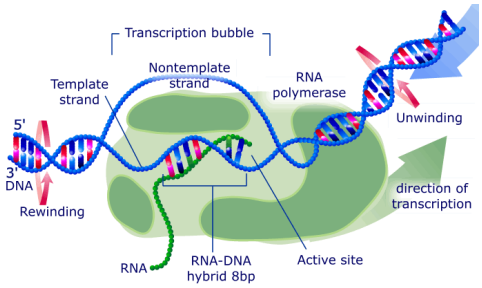
- The genetic code is made up of three letter 'words' (termed a codon) formed from a sequence of three nucleotides (e.g.. ACT, CAG, TTT).
- These *codons* can then be translated with messenger RNA and then transfer RNA, with a codon corresponding to a particular amino acid.
- Since there are 64 possible codons, most amino acids have more than one possible codon.
- There are also three 'stop' or 'nonsense' codons signifying the end of the coding region.

Name	1-Letter Nickname	Triplet
Glycine	G	GGT,GGC,GGA,GGG
Alanine	A	GCT,GCC,GCA,GCG
Valine	V	GTT,GTC,GTA,GTG
Leucine	L	TTG,TTA,CTT,CTC,CTA,CTG
Isoleucine	I	ATT,ATC,ATA
Histidine	H	CAT,CAC
Serine	S	TCT,TCC,TCA,TCG,AGT,AGC
Threonine	T	ACT,ACC,ACA,ACG
Cysteine	C	TGT,TGC
Methionine	M	ATG
Glutamic Acid	E	GAA,GAG
Aspartic Acid	D	GAT,GAC,AAT,AAC
Lysine	K	AAA,AAG
Arginine	R	CGT,CGC,CGA,CGG,AGA,AGG
Asparagine	N	AAT,AAC
Glutamine	Q	CAA,CAG
Phenylalanine	F	TTT,TTC
Tyrosine	Y	TAT,TAC
Tryptophan	W	TGG
Proline	P	CCT,CCC,CCA,CCG
Terminator	*	TAA,TAG,TGA

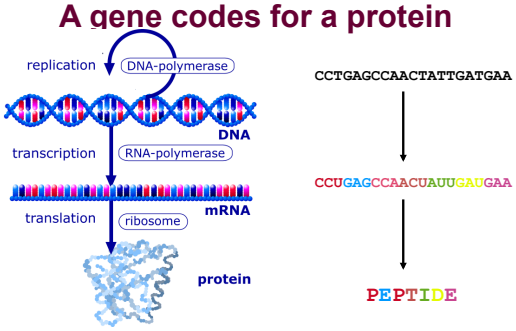
Proteins - Summary

- DNA sequence determines protein sequence
- Protein sequence determines protein structure
- Protein structure determines protein folding and function

Transcription

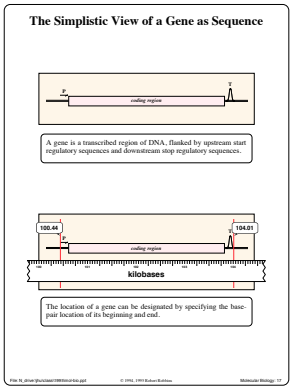


A gene codes for a protein

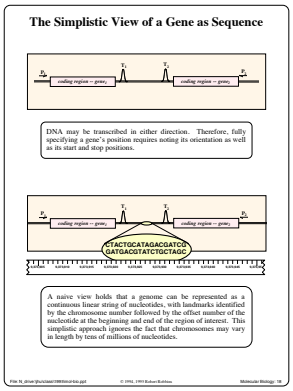


What is a gene?

- **Classical Definition:** fundamental unit of heredity, mutation, and recombination.
- **Physiological Definition:** fundamental unit of function
- **Cistronic Definition:** fundamental unit of expression.
- **Sequence Definition:** the smallest segment of the gene- string consistently associated with the occurrence of a specific genetic effect.



31



32

Genome Sizes

	Base Pairs	Genes
Mycoplasma genitalium	580,073	483
MimiVirus	1,200,000	1,260
E. coli	4,639,221	4,290
Saccharomyces cerevisiae	12,495,682	5,726
Caenorhabditis elegans	95,500,000	19,820
Arabidopsis thaliana	115,409,949	25,498
Drosophila melanogaster	122,653,977	13,472
Humans	3.3×10^9	~25,000

33

DREAM Challenges

- A crowd sourcing effort to encourage research communities to contribute to challenges.
- Challenges
  - Previously hosted at Columbia http://wiki.c2b2.columbia.edu/dream/index.php/Challenges
  - Sage Bionetworks (closed) http://dreamchallenges.org/project-list/closed/
  - Sage Bionetworks (open) http://dreamchallenges.org/challenges/

34

"Crowdsourcing Biomedical Research: The DREAM Challenges" by Dr. Gustavo Stolovitzky (2015)

- https://www.youtube.com/watch?v=u-sqDfSWRrE
- 2:10-7:00

35

DREAM 5 (2010) Gene network inference

- The goal is to infer gene networks, given a compendium of simulated data:
  - Genetic perturbations: deletions, over-expression
  - Time series: multi-factorial, drug perturbations
  - Repeats: noise, measurement errors
- Publications:
  - From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis. Pinna et al. PLOS One 2010.
  - Wisdom of crowds for robust gene network inference. Marbach et al. Nature Methods 2012.

36

DREAM 9 (2014)
Acute Myeloid Leukemia Outcome Prediction

- Given clinical data and protein measurements, the goal is to predict outcome and survival time.
- Sub-challenge 1: Predict
  - Primary resistant: resistant to the given therapy
  - Complete remission (CR): no signs or symptoms of the disease
- UW-Tacoma team: Claire Gu, Hong Hung, Ka Yee, Kiyana Zolfaghar, Maciej Fronczuk, Vivian Oehler

37

DREAM 9 (2014)
Acute Myeloid Leukemia Outcome Prediction

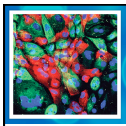
- Data description
  - Binary response: resistant (Y=0), CR (Y=1)
  - Training data:
    - 40 clinical variables, 231 protein measurements
    - 191 patients = 136 CR + 55 resistant
  - Test data: 100 patients, same attributes
- Can bring in any additional data sources as you wish! ☺
- Major challenge: major variations in different subsets of the training data
- A Crowdsourcing Approach to Developing and Assessing Prediction Algorithms for AML Prognosis. Noren et al. PLoS Computational Biology 2016, 12(6): e1004890. We served as a member of the DREAM 9 AML-OPC Consortium (as a collaborator)

38

Group Project (Spring 2015)

https://www.synapse.org/#!Synapse:syn2813558

39



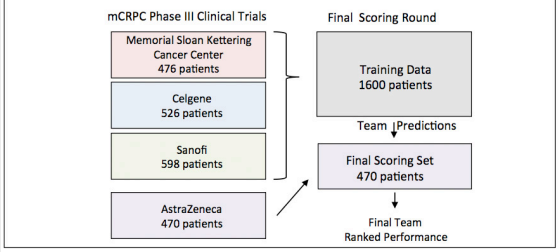
Prostate Cancer DREAM Challenge

- Motivation
  - Prostate cancer is the most common kind of cancer among men in developed countries
  - androgen deprivation therapy (ADT)
  - Resistant to ADT → metastatic castrate-resistant prostate cancer (mCRPC)
  - Goal: improve the treatment of patients with mCRPC
- Data
  - four phase III clinical trials
  - 150+ clinical variables and 2,000+ mCRPC patients
- Challenges

1) Predict overall survival of mCRPC patients
2) Predict discontinuation of docetaxel treatment due to adverse events at early time points.

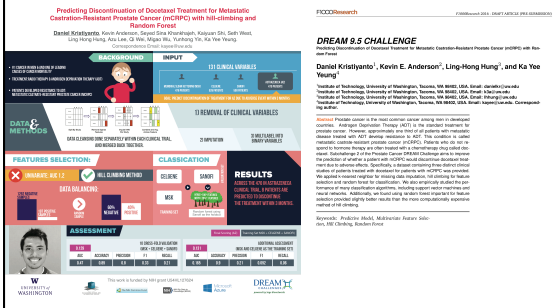
40

Subchallenge 2: Predict Treatment Discontinuation



- Solvers are required to predict discontinued treatment for patients due to adverse events with 3 months of starting treatment.
- Predictions are submitted in the form of a continuous variable with bigger values corresponding to higher probability of discontinuation due to adverse events

41



Kristiyanto et al. [F1000Research 2016, 5:2673]

42

Respiratory Viral DREAM Challenge

Group Project (Spring 2016)

Discovering dynamic molecular signatures in response to virus exposure

https://www.synapse.org/#!Synapse:syn5647810/wiki/399103

43

Motivation

Respiratory viruses are highly infectious and cause acute illness in millions of people every year.

- However, there is wide variation in the physiologic response to exposure at the individual level. It is not well understood what characteristics may protect individuals from respiratory viral infection.
- This Challenge aims to develop early predictors of susceptibility and contagiousness based on expression profiles that were collected prior to and at early time points prior to, and following viral exposure.

44

Respiratory viral DREAM challenge

- https://www.synapse.org/#!Synapse:syn5647810/wiki/393443
- Given: time series gene expression data (time 0, 24 hours) across 4 different viruses (H1N1, H3N2, RSV, Rhinovirus) in 7 studies.
- Goal: build predictors to distinguish people who become contagious after exposure to flu and other respiratory viruses.
- Participated in sub-challenge 1 and sub-challenge 2.

45

TCSS 588 (Spring 2016) Respiratory Viral DREAM Submission

- Team Espoir: Xiao Liang, Reem Almugbel, Abeer Almutairy, Lan Lyu, Mohammad Rahman
- Leaderboard results
  - Sub-challenge 1: Ranked 4 (time 0) and 9 (time 24) out of 125 submissions.
  - Sub-challenge 2: Ranked 15 (time 0) and 26 (time 24) out of 115 submissions.
- Independent test phase (sub-challenge 2, RSV only): Ranked 3 (time 0) and 11 (time 24) out of 13.
- Co-authors on a journal paper under review.

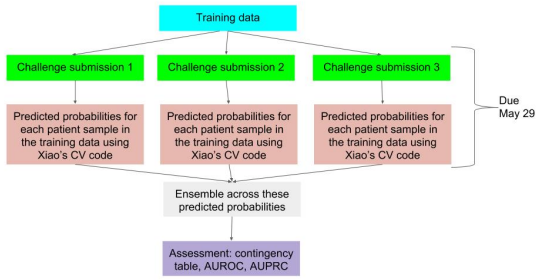
TCSS 588 (Spring 2017)
Respiratory Viral DREAM Submissions Reproducibility + Ensemble predictions

- Downloaded the code and documentation from the submissions to the DREAM challenge, and try to reproduce the results using Jupyter notebooks and provided cross validation code
- Total 28 teams, 2 time points. ~50 total submissions. Each team signed up for minimum 3 submissions. Bonus credits if sign up for extra.

TCSS 588 (Spring 2017)
Respiratory Viral DREAM Submissions Reproducibility + Ensemble predictions

Reproducibility issues:

- Some code are inaccessible
- Missing input data files
- Undefined variables
- Programming language: R, python, matlab
- Assume external annotation files, libraries, or executables
- Jupyter notebook: memory issues



49

Before next class

- Install Jupyter with R kernel on your computer.
  - Use the canvas Discussion board if you have any issues/concerns.
- Learn basic R commands. Resources are available from canvas.
- Bring your computer to next class. We will start using R to perform data analysis.

50

Credits

- Textbook
- Inspired by slides from Larry Ruzzo @ UW-Seattle, Garima Bajetha Joshi, Ulf Schmitz

51