Welcome:
Temple Bioinformatics Teachers Workshop

- Funded as part of an NSF grant awarded to Arun Sethuraman and Jody Hey
- A 1 week introduction to the field of bioinformatics, together with guidance in developing a lesson plan.
    - participants working in pairs
    - Each pair will develop a lesson plan to bring back to your students to introduce them to the field of bioinformatics.
- Introductions
- Computing basics

J. Hey introduction

- Current Research Program:
  - Evolutionary genomics
  - Use genomic data to figure out evolutionary history
  - Develop statistical methods
  - Lots of computer programming
- Previous research milestones:
  - Ran a DNA sequencing lab at Rutgers from 1989-2004
  - Switched completely to computational/statistical genomics in 2004
  - Published over 100 research papers & books
- Moved from Rutgers to Temple in 2013
  - Established the Center for Computational Genetics and Genomics, CCGG
  - Established a new Professional Science Masters program in Bioinformatics
    - http://bioinformatics.cst.temple.edu/

CCG Center for Computational Genetics and Genomics

J. Hey computing history

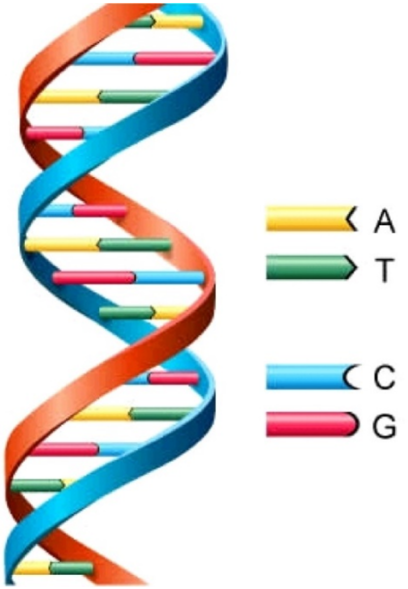
- Computer language history:
    - Basic 1976
    - pop2 1977
    - fortran 1981
    - pascal 1983
    - C 1993
    - C++ 2005
    - python 2008
- Operating systems:
    - 1975 IBM 5100
    - Max operating system 1981
    - DOS 1983
    - OS/2 1992
    - Windows & Linux since 1995
- Computer Programs authored and distributed:
    - SITES
    - HKY
    - IM, IMa,IMa2

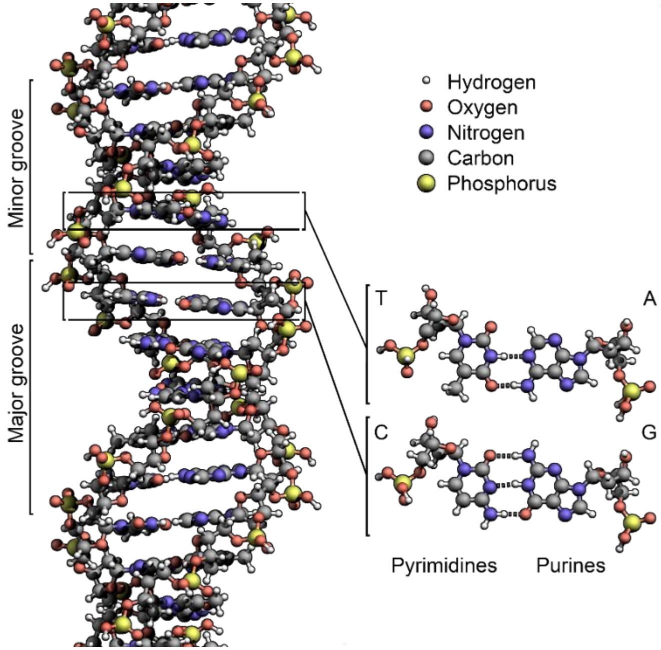
What is a gene?

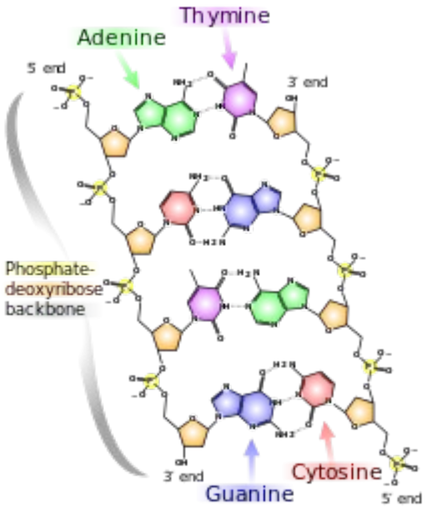
- A common traditional definition:
  - a part of the DNA in a cell that codes for a specific protein
- But DNA also carries other kinds of information that is needed for development, and that is not part of a protein sequence
  - Regulatory sequences
  - Many kinds of genes that code for specific RNA molecules
- Definition of "gene" can be hard to pin down
- Two takeaways:
  - Genes exist in the DNA
  - Genes are where the DNA encodes information about how the cell functions

What is DNA

- An organic molecule
- A chain of smaller molecules (bases)
  - There are four kinds of bases: A,C, G and T
- Exists as two complementary chains
  - A pairs with T and G pairs C







What do we mean by a “DNA sequence”

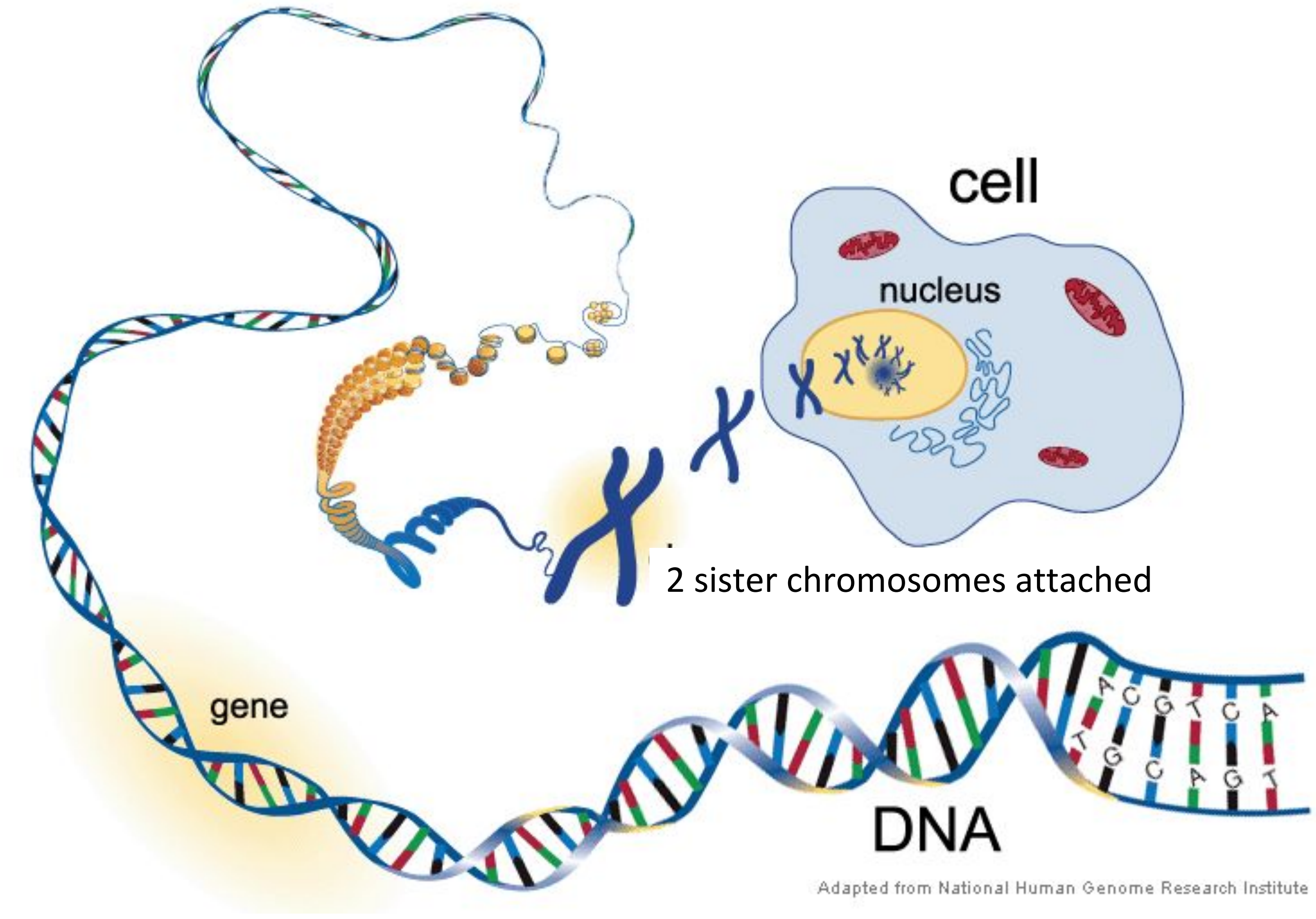
- DNA usually exists as two strands stuck together along their length.
- The bases on one strand pair with those on the other strand following the rules: A with T, and C with G
- So if you know the sequence of bases on one strand then you can also write down the sequence on the other strand
- So we only need to write down the sequence for one strand
- Each strand of DNA has a molecular orientation with one end called 5’ (5 prime) and the other called 3’ (3 prime)
- The orientation of one strand of DNA is in the reverse direction of the strand that it is paired with.
- For example we can write a short sequence as:
    - 5’ TGAAGCTGA3’
    - 3’ ACTTCGACT5’
- By convention we only write down one strand, and by convention we put the 5’ end on the left, e.g. TGAAGCTGA

A DNA sequence is fundamentally digital

- A DNA sequence, regardless of its length, is just a sequence of 4 different symbols, e.g. TGAAGCTGA
- It is like a number written in base 4.
- E.g. let A=1, T = 3, C= 0 and G=2 then TGAAGCTGA = 341140341
- Digital computers operate on numbers that take two values, 0 and 1. (this is base 2).
- It is not difficult to convert a number in base 2 to its equivalent value in base 4 or base 10.
- DNA sequences and digital computers go really well together

What do we mean by a ‘genome’

- Every person starts as a single cell formed by the union of a sperm and an egg, each which carries a single set of 23 distinct chromosomes.
- As the zygote develops into a person, every cell carries copies of both original sets chromosomes :one that came from the biological mother and the other from the biological father
- Each chromosome includes a single long DNA molecule
- A genome sequence is the DNA sequence of one complete set of chromosomes.
- A single human genome exists as 23 chromosomes (23 DNA molecules) with a total length of about 3.2 billion bases.



2 sister chromosomes attached

Adapted from National Human Genome Research Institute

What is the physical length of the human genome?

- Distance between two base pairs (bp) in double stranded DNA: 3.38 angstroms (= 3.38×10^{-8} meters)
- Length of 3.2 billion bases: $3.38\times10^{-8} \times 3\times10^{9}$ =1.082 meters
- A cell has two copies of the genome, so 2.163 meters.
- If you stretched all the DNA in a human body end to end, how far would it reach?
  - 2.163 meters of DNA per cell
  - Approximately 37 trillion cells in a human body (Bianconi et al 2013)
  - $2.163 \times 37\times10^{12} = 160 \times10^{12}$meters = 160×10^{9} kilometers
  - Or about 100 billion miles (this is over 1000 times the distance between earth and the sun)
  - Light would take about 6 days, 5 hours to travel this far

What is the information content of the human genome?

- Consider a basic component of computer memory – the bit. One bit can represent two numbers, either a zero or a one.
    - Two bits can represent four possible numbers: 00, 01, 10, 11
    - N bits can store 2^N possible numbers
- DNA has 4 states, so one base position can represent four numbers
    - N bases can represent 4^N different numbers
- The English alphabet has 26 states (letters) so N letters can represent 26^N different numbers
- A human genome, with 3.2×10^9 bases can represent $4^{3200000000}$ different numbers (approximately $10^{1926591972}$ numbers) in base 10
- In other words, you can think of the DNA sequence of one of your genomes (e.g. the one you got from mom), as a number with $10^{1926591972}$ digits
- This is about the same as you could represent using a 26 letter alphabet, with 3.4×10^8 letters (340,000,000 letters)
- This is about as many letters as you would find in 500 books (each with 1400 characters per page and 500 pages)
- So the human genome can hold about as much information as a library of 500 books

What is bioinformatics?

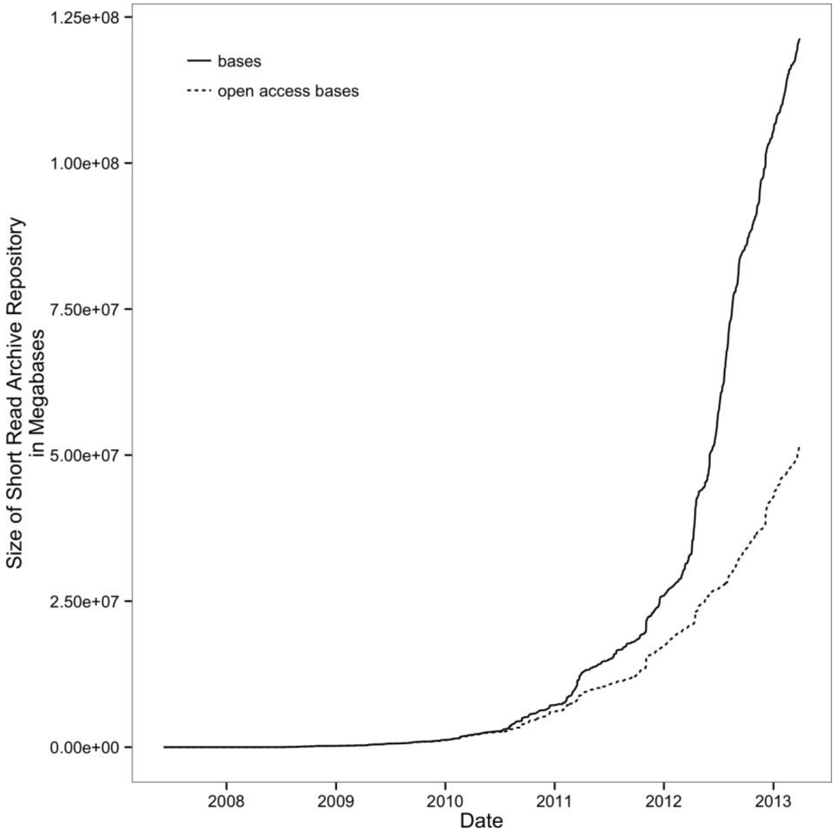
- The meaning of “bioinformatics” can be hard to pin down.
- There are many vague descriptions of the term
- For example:
- “Bioinformatics is a fast evolving field that integrates elements of biology, chemistry, computer science, and statistics, and that has become an essential part of the biotechnology and pharmaceutical industries.”

What is bioinformatics?

- The basic idea behind “Bioinformatics” (*‘Biology’ + ‘Information’)* is simply the idea of using computers to deal with lots of biological data.
- In the 1990’s scientists began collecting DNA sequence data (and other kinds of biological data) using machines that provide very large amounts of data very rapidly
- Consider:
  - Computers usually encode a letter of text with 1 byte (a ‘byte’ is a computer term for a basic unit of computer memory: 1 byte = 8 bits).
  - So one human genome sequence of 3 billion bases (A’s,C’s,G’s and T’s) takes up 3 billion bytes when stored in a computer.
  - A 3 Gigabyte file is pretty big.
  - An iphone 6 or 7 might have 32 or 64 Gigabytes of memory.
  - Back in the old days (1990’s) 3 Gigabytes was huge.
- Now human genome sequences are being obtained very commonly.
  - https://www.scientificamerican.com/article/full-genome-sequencing-for-newborns-raises-questions/
- We need lots of computing power to deal with genome sequences!
- And we need people, teachers and scientists, who understand biological data *and* understand how to work with computers!

Bioinformatics is growing rapidly

- In a few years we will probably all have our genomes sequenced.
- Increasing rate of appearance of:
  - New sequencing technologies
  - New methods
  - New programs and platforms for analysis
- Every new discovery opens the door to new kinds of bioinformatic work



What does it mean to be a Bioinformatician?

- To understand biological data and to have the knowledge and ability to make a computer do what you want to biological data
- Depends upon:
    - A good understanding of the data
        - Bioinformaticians usually understand genetics very well
    - A logical mindset - the capacity to organize a problem and design a path to a solution
    - Knowledge of the operating system and programs that are available to you
    - Knowledge of how to connect different programs and to write new programs for processing data

You don't have to be an expert bioinformatician to introduce high school students to bioinformatics

- Ten Simple Rules for Teaching Bioinformatics at the High School Level
  - http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002243

1. Rule 1: Keep It Simple
2. Rule 2: Familiarity: Use Activities to Explore Examples That Are Familiar to Students
3. Rule 3: Link Activities to Preexisting Science Curricula
4. Rule 4: Develop Activities That Build on Each Other
5. Rule 5: Use Activities to Build Skills and to Provide Information through Inquiry-Based Research
6. Rule 6: Provide Opportunities for Individualization
7. Rule 7: Address Multiple Learning Styles
8. Rule 8: Empower Students
9. Rule 9: Model Processes Using Pen and Paper before Using the Computer
10. Rule 10: Produce a Product