FOCUS COMPSCI 004G
The Genome Revolution and its Impact on Society

Owen Astrachan
http://www.cs.duke.edu/courses/cps004g/fall05
http://www.cs.duke.edu/~ola

Where are we going?
• What is computer science?
• What is biology?
• What is computational biology?
• What is bioinformatics?
• What tools do scientists need?
• What is a scientist?

What is Bioinformatics?
• Synonym?: computational biology www.pasteur.fr
• The application of computational techniques to the management and analysis of biological information. www.informatics.jax.org
• Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine. www.pasteur.fr

What is Bioinformatics?
• Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface. www.bisti.nih.gov/CompuBioDef.pdf
• Because of the great success of genome-sequencing projects, the quantity of DNA sequence data that are now available greatly exceeds the tools that are available to process those data. Consequently the analysis of those data presents one of today’s great scientific challenges.
What is Computer Science?

- Computer science is no more about computers than astronomy is about telescopes.
  
  Edsger Dijkstra

- Computer science is not as old as physics; it lags by a couple of hundred years. However, this does not mean that there is significantly less on the computer scientist's plate than on the physicist’s: younger it may be, but it has had a far more intense upbringing!
  
  Richard Feyneman

Scientists and Engineers

- Scientists build to learn, engineers learn to build.
  
  Fred Brooks

Computer Science and Programming

- Computer Science is more than programming
  - The discipline is called informatics in many countries
  - Elements of both science and engineering
  - Elements of mathematics, physics, cognitive science, music, art, and many other fields

- Computer Science is a young discipline
  - Fiftieth anniversary in 1997, but closer to forty years of research and development
  - First graduate program at CMU (then Carnegie Tech) in 1965

- To some programming is an art, to others a science, to others an engineering discipline

What is Computer Science?

What is it that distinguishes it from the separate subjects with which it is related? What is the linking thread which gathers these disparate branches into a single discipline? My answer to these questions is simple --- it is the art of programming a computer. It is the art of designing efficient and elegant methods of getting a computer to solve problems, theoretical or practical, small or large, simple or complex.

C.A.R. (Tony) Hoare
What is Computer Science?

- What does a computer scientist do?
- What does a programmer do?
- What does a systems administrator do?
- What do you want to do?

Algorithms as Cornerstone of CS

- Step-by-step process that solves a problem:
  - more precise than a recipe
  - eventually stops with an answer
  - general process rather than specific to a computer or a programming language
- Searching: for phone number of G. Samsa, whose number is 929-9338, or for the person whose number is 489-6569
  - Are these searches different?
- If the phone book has 8 million numbers in it (why are there only 7.9 million phone numbers with area code 212?):
  - How many queries to find phone number of G. Samsa?
  - How many queries to find person with number 929-9338?
  - What about IP addresses?

Search, Efficiency, Complexity

- Think of a number between 1 and 1,000
  - respond high, low, correct, how many guesses needed?
- Look up a word in a dictionary
  - Finding the page/word, how many words do you look at?
- Looking up a phone number in the Manhattan, NY directory
  - How many names are examined?
- How many times can 1,024 be cut in half?
  - $2^{10} = 1,024$, $2^{20} = 1,048,576$

Themes and Concepts of CS

- Theory
  - properties of algorithms, how fast, how much memory
  - average case, worst case: sorting cards, words, exams
  - provable properties, in a mathematical sense
- Language
  - programming languages: C++, Java, C, Perl, Fortran, Lisp, Scheme, Visual BASIC, ...
  - Assembly language, machine language,
  - Natural language such as English
- Architecture
  - Main memory, cache memory, disk, USB, SCSI, ...
  - pipeline, multi-processor
Abstraction, Complexity, Models

- What is an integer?
  - In mathematics we can define integers easily, infinite set of numbers and operations on the numbers (e.g., +, -, *, /) 
    ...[-3, -2, -1, 0, 1, 2, 3, ...]
  - In programming, finite memory of computer imposes a limit on the magnitude of integers.
    - Possible to program with effectively infinite integers (as large as computation and memory permit) at the expense of efficiency
    - At some point addition is implemented with hardware, but that’s not a concern to those writing software (or is it?)
  - C++ doesn’t require specific size for integers, Java does

- Floating-point numbers have an IEEE standard, it’s more expensive to do arithmetic with 3.14159 than with 2

Complexity: What’s hard, what’s easy?

- What is a prime number?
  - 2, 3, 5, 7, 11, 13, ...
  - Largest prime?

- 4811295837082048697
- 671998030559713968361666935769

- How do we determine if these numbers are prime?
  - Test 3, 5, 7, ...
  - If we can test one million numbers a second, how long to check a 100 digit #?

- Why do we care?
- 671998030559713968361666935767 is not prime, I can prove it but I can’t give you the factors.

- Finding factors is “hard”, determining primality is “easy”
  - What does this mean?
  - Why do we care?

- Encryption depends on this relationship, without encryption and secure web transactions where would we be?

What is digital?

- What’s the difference between:
  - Vinyl LP and CD/DVD?
  - Rolex and Timex?

- Sampling analog music for CD’s
  - 44,100 samples/channel/second * 2 channels * 2 bytes/sample * 74 minutes * 60 seconds/minute = 783 million bytes

- How does MP3 help?

Alan Turing (1912--1954)

- Instrumental in breaking codes during WW II
- Developed mathematical model of a computer called a Turing Machine (before computers)
  - solves same problems as a Pentium III (more slowly)
- Church-Turing thesis
  - All “computers” can solve the same problems
- Showed there are problems that cannot be solved by a computer
- Both a hero and a scientist/mathematician, but lived in an era hard for gay people
Lincoln Stein, bioinformatician

Lincoln D. Stein is awarded the 2004 Benjamin Franklin Award for his creation of a great number of open-source bioinformatics programs and for championing open-source principals in many venues...

"[bioinformatics is] Biologists using computers, or the other way around."

Citation: WormBase: a multi-species resource for nematode biology and genomics (2004)

Thomas Kepler (Computational Biologist)

Division Chief of Computational Biology of the Department of Bioinformatics and Biostatistics at Duke University, is a Professor of Bioinformatics and Biostatistics

Citation: "Invertebrate immune systems--not homogeneous, not simple, not well understood"

Chips, Central Processing Unit (CPU)

- CPU chips
  - Pentium (top)
  - G4 (bottom)
  - Sound, video, ...
- Moore’s Law
  - chip “size” (# transistors) doubles every 12–18 months (formulated in 1965)

Why is programming fun?

What delights may its practitioner expect as a reward?

First is the sheer joy of making things

Second is the pleasure of making things that are useful

Third is the fascination of fashioning complex puzzle-like objects of interlocking moving parts

Fourth is the joy of always learning

Finally, there is the delight of working in such a tractable medium. The programmer, like the poet, works only slightly removed from pure thought-stuff.

Fred Brooks