

CSC 121 Computers and Scientific Thinking

Fall 2005

Computers in Biology and Bioinformatics

1

Biology



biology is roughly defined as "the study of life"

- it is concerned with the characteristics and behaviors of organisms, how species and individuals come into existence, and the interactions they have with each other and with the environment (en.wikipedia.org/wiki/Biology)

biology encompasses a broad spectrum of academic fields that are often viewed as independent disciplines

- *ecology* and *evolutionary biology* study life at the habitat or population level
- *developmental biology* and *genetics* study life at organism level
- *physiology*, *anatomy*, and *histology* study life at the multicellular level
- *cell biology* studies life at the cellular level
- *molecular biology*, *biochemistry*, and *molecular genetics* study life at the atomic and molecular level

2

Impact of Computers



the history of biology dates as far back as the rise of various civilization

- while computers are relatively new, they have had a monumental impact on biological research

3 examples of impact:

1. computer technology is rapidly advancing the tools of scientific research
2. computer models are being used to study complex systems
3. computers are being used to store, process, and analyze large collections of biological data

note: this list is in no way exhaustive

- many aspects of biology and computer science are converging
- biology researchers must be savvy computer users and even programmers
- computer scientists must be able to solve interdisciplinary problems

3

Technology Tools/Resources



many of the traditional tools of biological research are integrating computer technology

- e.g., the confocal microscope
- invented by Marvin Minsky (computer science pioneer)
 - works by focusing a laser on a dyed sample and measuring the fluorescent light emitted
 - can be used to build up a 3-D model of a sample, stored on a computer



e.g., [DNA Microarrays](#) to measure the expression levels of genes

the Internet and the Web allow researchers to share data and publications

- speeds the dissemination of information and the advancement of science

e.g., [PubMed](#), from the National Library of Medicine

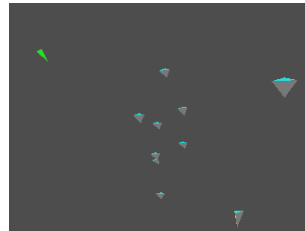


4

System Modeling

as computer memory and processing power has increased, it has become possible to model complex biological systems in software

- can attempt to discern natural laws or behaviors by observing the model under varying conditions
 - e.g., models of plant or seashell growth
 - e.g., the evolution of cooperative behavior in species, such as bird flocking
- can predict the effects of actions over long periods
 - e.g., the effects of automobile emissions on global warming
 - e.g., the effects of increased fishing on worldwide fishery stocks
- can avoid infeasible, unethical, or costly experimentation
 - e.g., predict the toxicity of a new drug based on a chemical/biological model as opposed to animal testing
 - e.g., study brain trauma using a neural network model

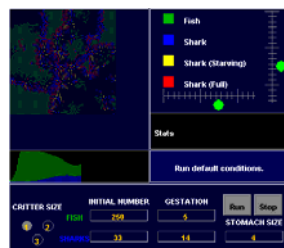
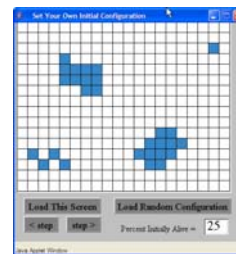


5

Ecosystem Modeling

in the late 1960s, John Conway showed that a simple model of an environment could produce complex and interesting behavior

- the environment is modeled as a 2-D grid of cells
- a cell can be alive (contain an organism) or dead
- simple rules model evolution
 1. a dead cell becomes alive in the next generation if it has exactly 3 neighbors
 2. a living cell survives in the next generation if it has 2 or 3 neighbors



Conway's ideas have been extended to a variety of ecosystems

- here, different colored cells denote different organisms (sharks & fish)
- other systems have modeled:
 - ✓ the growth of viruses
 - ✓ the spread of infectious diseases in a population
 - ✓ the behavior of an ant colony

6

Bi o i n f o r m a t i c s



perhaps the biggest impact of computers in biology is in storing, accessing, and processing large amounts of biological data

the new field of bioinformatics bridges biology and computer science (or informatics, as it is known in Europe)

- *broad definition of bioinformatics*: the use of computer science techniques to solve biological problems
- *narrower but common definition*: the application of computer science techniques to the representation and processing of biological data

as research tools advance, biologists are generating enormous amount of data

- a single experiment with genetic material can produce thousands or millions of data points
- computational and statistical tools are needed to analyze and understand such volumes of data

7

DNA Overvi ew



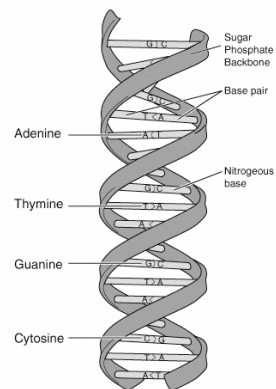
DNA is the genetic blue-print of life

- made of nucleotides with four bases (A, T, G, C), organized in a double-helix
- the two strands match A+T and C+G base pairs
- can think of DNA as encoding information in base 4

a gene is a region of DNA that encodes the chemical structure of a protein

it is currently believed that there are 20,000-30,000 different genes in human DNA

- roughly 3 billion base pairs



"If our strands of DNA were stretched out in a line, the 46 chromosomes making up the human genome would extend more than six feet. If the ... length of the 100 trillion cells could be stretched out, it would be ... over 113 billion miles. That is enough material to reach the sun and back 610 times." [Source: Centre for Integrated Genomics]

8

DNA → RNA → Proteins

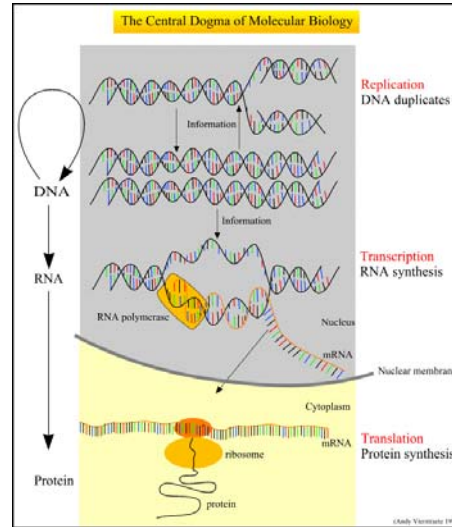
in cell division,

- the two strands of DNA are split
- each strand is paired with free nucleotides in the nucleus to complete copies of the original DNA
- each cell gets a complete set of the DNA

in mapping DNA to proteins,

- the DNA strands are split and copied into mRNA (using the same bases except U replaces T)
- this mRNA is then "read" by a ribosome to build the specified protein
- proteins are commonly represented using a 20 letter alphabet (for the different types of amino acids)

see www.dnai.org/a/index.html for a series of online animations



DNA Replication

DNA replication and transcription are basically information processing on a biological level

- if errors occur in the reading or replication of DNA information, then mutations and diseases are the result
- fortunately, DNA replication and transcription are INCREDIBLY reliable

Table 6-1 Error Rates

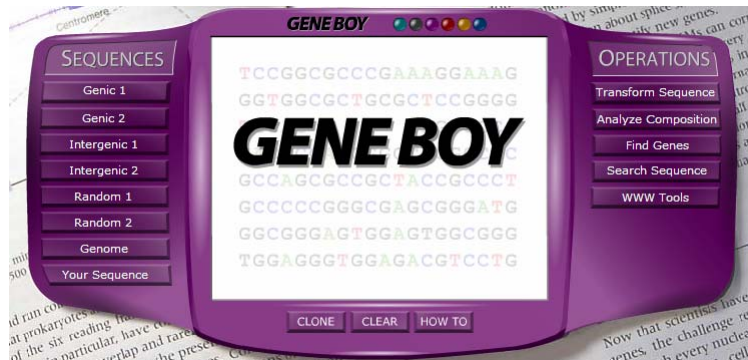
U.S. Postal Service on-time delivery of local first-class mail	13 late deliveries per 100 parcels
Airline luggage system	1 lost bag per 200
A professional typist typing at 120 words per minute	1 mistake per 250 characters
Driving a car in the United States	1 death per 10 ⁴ people per year
DNA replication (without mismatch repair)	1 mistake per 10 ⁷ nucleotides copied
DNA replication (including mismatch repair)	1 mistake per 10 ⁹ nucleotides copied

©1996 GARLAND PUBLISHING

Bioinformatics Tools

many tools are available for searching and manipulating genetic sequences

- e.g., the GeneBoy program (www.dnai.org/geneboy)
 - demonstrates DNA → RNA transformations
 - analyzes the composition of a sequence
 - searches for specific patterns in the sequence



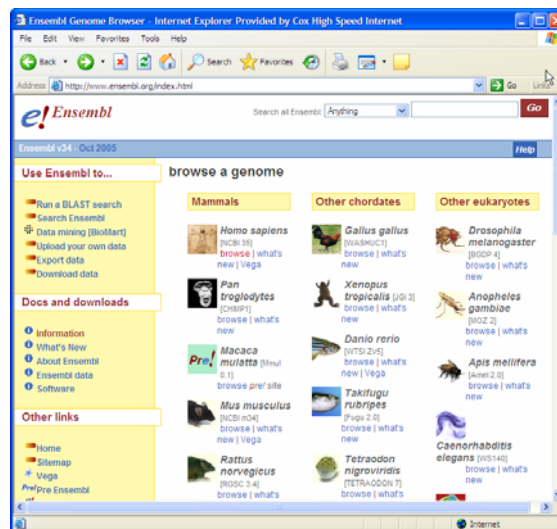
11

DNA Databases and Tools

often, the source or purpose of a DNA sequence can be determined by comparing it with documented genetic material

- several large databases are available online
- tools for visualizing and/or searching the databases are also available

e.g., the Ensemble site (www.ensembl.org) contains visualizations of the human genome and other DNA sequences



12

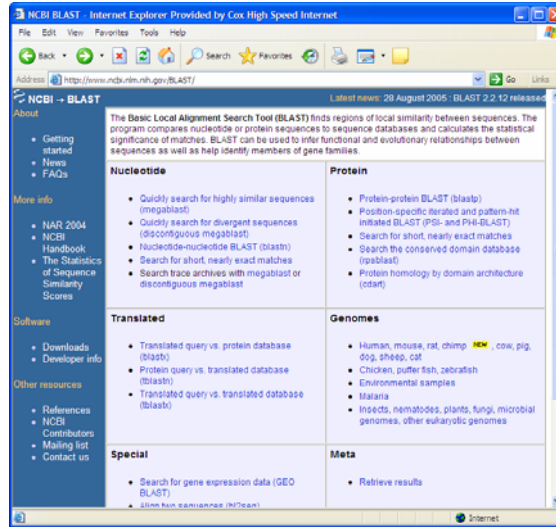
GenBank



the GenBank public repository of DNA and RNA sequence data contains

- partial or complete genomes for more than 165,000 organisms
- more than 1 trillion bases of sequence data
- roughly 3 million new DNA sequences are added per month

the database can be accessed and searched using various tools at www.ncbi.nlm.nih.gov



13

BLAST Search



Jurassic Park example

14

Bi oinformatics i n the News



Researchers at the University of Bath have won a £261,000 grant to use the latest software to produce a blueprint of a designer drug that could stop influenza and some other diseases from replicating in humans.

UCSD biochemists have developed a computer program that helps explain a long-standing mystery: how the same proteins can play different roles in a wide range of cellular processes, including those leading to immune responses and cancer.

Blue Gene is an IBM Research project dedicated to exploring the frontiers in supercomputing: in computer architecture, in the software required to program and control massively parallel systems, and in the use of computation to advance our understanding of important biological processes such as protein folding.

- will utilize 65,536 processors working in parallel
- will be able to perform 360 trillion operations/sec (greater than the total computing power of the world's current 500 most powerful supercomputers)