

# Introduction

# Computational Biology? Bioinformatics? Genomics?

- explosion of new molecular biological information
  - sequence information - genomes
  - microarray information - expression
- far too much for humans to comprehend
- rapidly increasing progress in computer technology
- how to combine the two so that they work together?
- very exciting time to be at the cutting edge in two fields

# Computational Biology

- **computational biology** - seeks to use computational methods to analyze biological processes
- includes extraction of information directly from sequences, sometimes called **Bioinformatics**
- includes comparing sequences and structures to each other and to **databases** to look for relationships
- includes discerning **evolutionary relationships** among organisms through sequence analysis
- includes comparing whole genomes to each other, sometimes called **Functional Genomics**
- includes predicting structures and interactions based on those structures, sometimes called **Structural Biology**
- includes finding **functional relationships** among sets of genes
- includes understanding **networks** of regulatory interactions

# Looking for Similar Sequences

- what sequence is similar to my sequence?
- is part of it similar or is the whole thing similar (local vs global similarity)?
- what if sequences are of vastly differing length?
- what about rearrangements? [AxxxxBxxx vs. xxxBxxAxx]
- how do you define similarity (need to define a scoring system)?
- Runtime - need to consider how fast the comparison runs - some compromises have to be made
- similar structure even if sequence is not very similar?

# Looking for Similar Structures

- many of the same issues as for sequence comparisons
- local vs global similarity
- granularity - what size should the basic structural unit be [e.g an amino acid? an alpha helix?]
- need scoring system(s)

# Defining Function

- function of whole molecule? or of pieces?
- consider binding to DNA
  - can be the main function of the molecule (e.g. histones)
  - can be part of a molecule's function - transcription factor (interacts with DNA and with other TFs)
  - can be part of a DNA polymerase or RNA polymerase

# Looking for Patterns

- repeated sequences  
AAAA; ACAC; ACCACCACC; ACC<sub>n</sub>ACC<sub>nn</sub>ACC
- inverted repeated sequences - sometimes with spaces  
ACGCGT; ACC<sub>nnn</sub>GGT
- patterns based on physical characteristics - hydrophobic or hydrophilic AAs
- repeated arrangements of AAs that lead to higher order structures like alpha-helices or beta-sheets, leucine zippers
- repeated appearances of higher order structures such as helix-turn-helix found in DNA binding proteins or beta-barrels found in some enzymes
- patterns of regulatory sequences upstream of genes

# Gene Expression Patterns

- using microarray data or SAGE
- cell cycle
- cancer vs. normal tissue
- during development
- drug treatments
- how do you measure similarity of expression?

# Data Mining

- The process of exploring large amounts of data to look for information and relationships is called **data mining**.
- It is more similar to other branches of research than to traditional molecular biology research.
- Finding relationships among sequences or other patterns might lend some insight into the function of the molecules being studied.
- Available to anyone on Internet.

# What is a Database?

- stores information and allows for rapid retrieval and searching
- provides for comparisons and complex searches
- information can be in many formats
- database information is stored in records
- a record can contain a number of individual pieces of information
- a relational database allows for connections between different records or pieces of data

# Information Retrieval

- all records of people living in zip code 03755
- all DNA sequences from humans
- all human DNA sequences entered by a specific author
- literature references to papers by that author
- all genes for ATPases found in both humans and yeast but not Arabidopsis

# Data Warehousing

- Different database are established at different locations (and times) by different investigators, but much of the information is interrelated.
- How can all of these different databases be “related” to each other in such a way that information can be obtained easily from all of them?
- Data warehousing projects are designed to address the issues involved in tying the different databases together

# Reconciling Data Formats

- GenBank stores sequence information associated with a specific GeneID.
- KEGG databases store enzymes by specific IDs that are not the same as GenBank IDs
- Structure databases use an older GenBank name or an entirely different name
- Author names can be stored differently
  - Gross, Robert H.
  - Gross, R.H.
  - Gross, RH
  - R.H. Gross

# Some Day...

- Find sequence similarities for all genes whose expression is inhibited by a specific drug
- Find similarities in gene expression patterns for all genes that are phosphorylated in both humans and yeast after inhibition of cell division after G1
- Find common promoter motifs for all genes mentioned in a set of published papers on a particular topic

# Computer Architecture and Operating Systems

Bob Gross, Bio 39 / 139

# The CPU

- The Central Processing Unit
- Intel, PowerPC
- Has a number of “registers” that can store values
- Calculations occur by adding or subtracting values in registers or to a value in memory
- Works in binary (0 or 1 are the only allowed values)
- Has a speed that is usually measured in billions of operations per second (gigahertz, GHz)
- Many computers have multiple CPUs which can work in parallel

# Bits and Bytes

- a single unit of memory is called a **bit**. It stores a value of 1 or 0.
- 2-bits can have 4 different configurations - 00, 01, 10, 11  
the rightmost digit is  $2^0$ , the leftmost is  $2^1$   
00 => 0, 01 => 1, 10 => 2, 11 => 3
- typically, bits are combined into larger units to enable us to think about them more conveniently. 8-bits make up a **byte**. There can be  $2^8$  (256) different values for a byte.
- ASCII character set consists of 256 different letters, numbers and punctuation marks
- Unicode system can have as many as 100,000 characters as well as information about text direction and other properties - is becoming the standard.

# Some Terms

- **RAM** = Random Access Memory: the memory inside your computer that can temporarily store data and computer instructions.
- **Virtual memory** is used to run programs or hold data that require more memory than that which is physically available. Virtual memory is space on the hard disk that the computer treats as if it were actually RAM.
- **ROM** = Read Only Memory: a memory chip that actually contains instructions (or sometimes data). This is usually some specific instructions for helping the operating system work. This “memory” stays intact and is not lost when you turn off the computer.
- **Hard disk space** = sometimes erroneously called memory. It is the space available on your hard disk for storing information (including virtual memory). Organized by sectors and a disk directory.

# Network Connections

- **modems**: modulation-demodulation. Converts digital signals from your computer into analog signals that can be sent over phone lines. Frequency or amplitude changes can be used to signify a 0 or a 1.
- **FTP** (File Transfer Protocol) - designed to facilitate transfer of files from one specific location to another.
- **HTTP** (HyperText Transfer Protocol) - for displaying web pages
- **TCP/IP** (Transmission Control Protocol/Internet Protocol) can send and receive internet data packets.

# Computer Input

- disk
- network connection
- mouse
- keyboard
- audio
- video
- dataglove
- sensors/instruments
- drawing tablet
- ambient conditions (light, usage)
- other programs

# Computer Output

- screen
- printer
- audio
- video
- instruments
- other programs
- robotic devices
- vector vs. bitmapped (rasterized) graphics