Biology 212 General Genetics **Spring 2007**

Lecture 24: Human Genome Project

Reading: Chap. 10 pp. 368-379

Presentation: Powerpoint version of lecture posted with course documents on Blackboard.

Lecture Outline:
1. **Genomes and proteomes**
2. **Human genome project**
3. **Genome projects of model organisms**
4. **Bioinformatics**
5. **Functional genomics**

Lecture:

1. Genomes and proteomes

Genome: all the DNA in an organism.

Proteome: the complete set of proteins in an organism.

Simple organisms	**Complex organisms**
Smaller genomes	Larger genomes
Encode fewer genes for fewer proteins	Encode many genes for proteins
Little non-coding DNA	Much of the genome is noncoding: repetitive DNA, introns, etc.

Gene families
- Most eukaryotes contain families of related proteins
- Encode proteins with similar DNA sequences = homologous genes
- Multiple genes arise by gene duplication

2. Human genome project

human genome 3.2 billion base pairs
22 different autosomes, X and Y

Goals of human genome project
- Develop a high resolution map of the human genome
- Determine the base sequence of the DNAs from each of the human chromosomes
- Develop new computing technologies for storing and analyzing DNA sequences
- Consider ethical issues that arise

How was the human genome project accomplished?

DNA mapping

- Identified and mapped many "DNA markers", short DNA sequences that map to specific chromosome regions
- Improved methods for identifying the locations of genes on chromosomes

DNA sequencing

- Automated DNA purification and DNA sequencing
- Used cycle sequencing, a PCR-based modification of dideoxysequencing
- Used fluorescent tags
- Scanned data from sequencing gels into high speed computers
- Competition between U.S. govt. program (led by Francis Collins) and private company (Celera, led by Craig Venter)

Computing

- Develop databases for storing/retrieving and indexing DNA sequence data
- Develop new computer programs for analyzing data
- Apply state of the art computing technology to the problem

Ethical issues

- Who should have access to data? Private vs. public databases
- Genome diversity project
- Current project represents small number of individuals; shouldn't it represent the entire diversity of the species?
- How will genome project data be used in the future?
  - DNA based ID
  - Issues of discrimination in insurance and health care

Findings of human genome project

- 3.2 billion base pairs
- 31,000-39,000 genes
- more proteins encoded than genes due to alternative splicing: where different mRNAs are produced from a gene--> different proteins
- coding sequences represent less than 5% of the genome
- repeat sequence represent >50% of the genome

3. Genome projects of model organisms

Compare genes and proteins to determine conserved functions.

a. bacteria

E. coli

Organisms that cause diseases such as cholera, meningitis, tuberculosis, anthrax, syphilis

b. model eukaryotes

yeast
protozoan that causes malaria
C. elegans (nematode)
fruit fly
pufferfish (*Fugu* species)
human
Arabidopsis: small plant
rice

In progress: mouse, zebrafish

Conclusions from comparative genomics

- # of genes increases slightly with complexity of organism
- proteins encoded by metazoans larger and more complex
- functions we learn for proteins in simple organisms often provide clues for proteins in more complex organisms

4. Bioinformatics (to be covered in lab)

- use of computers to study biological processes or analyze biological data
- mainly focused on storage and analysis of DNA and protein sequences

databases: where DNA or protein sequences or other biological data are stored.

NCBI: National Center for Biotechnology Information= U.S. government sponsored site where DNA sequences are stored.

Databases are organized based on

- Species of origin
- Genomic sequence data vs. cDNA sequence data
- Complementary DNAs often represented as partially completed sequences=expressed sequence tags (ESTs)
- DNA markers (genes or unique DNA sequences)= STSs: sequence tagged sites

NR=non-redundant database; database where all sequences are represented only once

Programs/algorithms

BLAST programs

- Basic local alignment sequence tool
- Compares sequence to those in database
- Can search DNA sequences vs. DNA sequences or can predict protein sequences and search against derived or known protein sequences

Utility programs (useful programs for analyzing sequences)