Functional Modeling of Genes and Cellular Processes

Matt DeJongh
Summer 2004
Hope College
Bioinformatics

“Bioinformatics is the term coined for the new field that merges biology, computer science, and information technology to manage and analyze [genomic] data, with the ultimate goal of understanding and modeling living systems.”

Genomics and Its Impact on Medicine and Society – A 2001 Primer
U.S. Department of Energy Human Genome Program
DNA
the molecule of life

Trillions of cells

Each cell:
- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- **33,000** genes code for proteins that perform all life functions
The Human Genome Project

- Launched in 1990 by the Department of Energy and the National Institutes of Health as a 15 year effort.

- Primary Goal: To determine the “consensus” sequence of the 3.2 billion DNA bases (A, C, T and G) that comprise the 24 human chromosomes (1-22, X, Y)
The Human Genome Project

- 1982: 606 sequences, 680K base pairs
- 1990: 39K sequences, 49M base pairs
- 2002: 22M sequences, 28.5B base pairs
- 2004: 30M sequences, 36B base pairs
From Genes to Their Functions

• Full Human Genome sequence published in 2003.

• Next goal of the Human Genome Project: determine what the function of each gene is.

How do we represent the functions of genes so that databases of genetic information can answer the “what, where and how” questions?
Representing Gene Function

- Textual Annotations on gene sequence records; journal articles
  
  human readable

- Functional Models: data structures that represent the roles genes play in cellular processes, and algorithms that use them
  
  computer readable
Questions

• Do you know about any other genes with a function similar to X?
  indexing, searching, pattern matching

• What happens when lactose-intolerant people eat ice cream?
  simulation

• Which genes should I target with drugs in order to prevent heart disease?
  prediction
Gene Ontology

- controlled vocabulary
- hierarchical representation
- is_a and part_of relationships

useful for indexing, searching, pattern matching
Pathway Databases

- model the biochemical roles that genes play in cellular processes

useful for simulation and prediction
Research Question

• Can we combine the Gene Ontology and pathway databases to create a functional model with both of their advantages?
• used pathway databases to add *in_process* relationships to the Gene Ontology and improve its pattern matching capabilities

automated process for creating new relationships (wrote our own)

automated process for pattern matching (used Progol)
Linking Molecular Function and Biological Process

Molecular Function

- kinase activity
- hydrolase activity
- tagatose-6-phosphate kinase activity
- lactase activity
- LacZ

Turned on

pfkB

New links from pathway data

Biological Process

- lactose catabolism via tagatose-6-phosphate
- lactose catabolism
- lactose hydrolysis

Two functions which would be unrelated in Badea’s experiment are now seen to be part of the same Biological Process
Summer 2004

Continue experiments with adding \textit{in\_process} relationships to the Gene Ontology

- evaluate quality of the new relationships
- evaluate quality of existing gene annotations
- improve process for adding new relationships
Summer of 2004

Create Functional Models for other kinds of gene functions, such as transcription regulators

- Add more kinds of relationships to the Gene Ontology
- Create pathway-like databases for simulation and prediction
Fall 2004

Introduction to Bioinformatics, Take Two

• Co-taught with Dr. Hledin (Biology/Chemistry)
• 2 cr. lab, Tuesdays 8 to 11 am
• Prerequisite: CSCI 225 or General Biology