Computational Genomics and Molecular Biology

Dannie Durand
Fall 2004
Lecture 1

Outline
A whirlwind review of molecular biology
An overview of computational molecular biology
New problems in genomics

Genes Encode Proteins

GTGCACCTGACTCCTGAG...

A gene is a DNA sequence

V H L T P E...

A protein is an amino acid sequence

A protein folds into a 3D structure

DNA forms a double stranded helix

A → T
G → C

DNA replication

DNA
Protein Synthesis

DNA

Protein Synthesis: Transcription

DNA

RNA transcription

...GTTCGCTGGACTCTTGGCTACCCGAG...
...CACAGCGACTGAGGAACCGATGGGCTC...

GTGCAC

RNA transcription

open DNA helix

GTGTCGGTGCACCTGACTCCTGAG

GTGCACCTGACTCCTGAG

CACGTGGACTGAGGACTC...

CACaAGCa

CACGTG aGACTGAGGAACCGATGGGCTC...

RNA transcription

CTGACTCCTTGGCTACCCGAG...

TCG

GTG

CAC

AGC

GACTGAGGAACCGATGGGCTC...

RNA

- Adenine, Guanine, Cytosine, Uracil (AGCU)

...CUGACUCCUUGCUACCCGAG...

...GACTGAGGAACCGATGGGCTC...

- Single stranded
- Secondary structure
RNA Secondary Structure

CCGUGAA CCGUACC CGA UUUUAU UCCC... 

Protein Synthesis: Translation

Protein Translation

amino acid sequence

transfer RNA

UGAGGA

messenger RNA

CUC

ribosome

Protein Translation

amino acid sequence

transfer RNA

UGAGGA

messenger RNA

CUC

ribosome

Protein Translation

amino acid sequence

tertiary protein structure

transfer RNA

messenger RNA
Gene Regulation

If I have the same set of genes in every cell, how come my liver cells look so different from my skin cells?

Only a small number of genes are being translated into protein at any one time.
Gene Regulation Controls When a Gene Is Transcribed

Translating Genes into Proteins: Multicellular organisms

Gene Regulation

In single cell organisms, gene regulation orchestrates
- Responses to changing environment
- Cell cycle
In multicellular organisms, gene regulation orchestrates
- Tissue type differentiation
- Development from embryo to adulthood

Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
- Sequence comparison
- Reconstruct evolutionary history
- Gene prediction
- Predict structure from sequence
  - RNA
  - Proteins
New problems in genomics

This course
Computational Structural Biology 15-879

The Origins of Computational Biology

ARPANET 1970
First royal email

USENET newgroups

TCP/IP 1980
Internet

World Wide Web, Gopher

NCSA Mosaic

Pizza Hut goes online.
Growth of sequence data during the ’90’s


Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
  – Sequence comparison
  – Reconstruct evolutionary history
  – Gene prediction
  – Predict structure from sequence
    • RNA
    • Proteins
New problems in genomics

Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
  – Sequence comparison
  – Reconstruct evolutionary history
  – Gene prediction
  – Predict structure from sequence
    • RNA
    • Proteins
New problems in genomics

Sequence Comparison

Applications
  • Database searching
  • RNA structure
  • Evolutionary tree reconstruction
  • Gene finding
  • Sequence assembly

Why sequence data is so powerful:
Sequences are related!

Sequence similarity => functional similarity

BLAST
Reconstructing Evolutionary History

Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
- Sequence comparison
- Reconstruct evolutionary history
- Gene prediction
- Predict structure from sequence
  - RNA
  - Proteins
New problems in genomics
RNA Secondary Structure

Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
- Sequence comparison
- Reconstruct evolutionary history
- Gene prediction
- Predict structure from sequence
  - RNA
  - Proteins

New problems in genomics

Structure Determination is Hard

Predicting Protein Structure

- Tertiary structure:
  - Detailed physical models
  - Find the configuration of minimum energy
- Secondary structure:
  - e.g., Which amino acids participate in a helix?
- Secondary structure motifs:
  - e.g., Is this a coiled-coil protein?
- Threading:
  - Estimate protein structure from a related protein with known structure and similar sequence.
Predicting Protein Structure

Complete physical model of tertiary structure

Predicting Protein Structure

• Tertiary structure:
  – Detailed physical models
  – Find the configuration of minimum energy

• Secondary structure:
  – e.g., Which amino acids participate in a helix?

• Secondary structure motifs:
  – e.g., Is this a coiled-coil protein?

• Threading:
  – Estimate protein structure from a related protein with known structure and similar sequence.

Predicting Protein Structure

• Tertiary structure:
  – Detailed physical models
  – Find the configuration of minimum energy

• Secondary structure:
  – e.g., Which amino acids participate in a helix?

• Secondary structure motifs:
  – e.g., Is this a coiled-coil protein?

• Threading:
  – Estimate protein structure from a related protein with known structure and similar sequence.

Predicting Protein Structure

Complete physical model of tertiary structure

Too hard!

Predicting Protein Structure

Secondary structure:
Does this amino acid participate in an alpha-helix?

Protein Secondary Structure

alpha helix
beta sheet

Is this a coiled-coil protein?
Predicting Protein Structure

- Tertiary structure:
  - Detailed physical models
  - Find the configuration of minimum energy
- Secondary structure:
  - e.g., Which amino acids participate in a helix?
- Secondary structure motifs:
  - e.g., Is this a coiled-coil protein?
- **Threading**:
  - Estimate protein structure from a related protein with known structure and similar sequence.

Threading

**Structure?**

Estimate protein structure from a related protein with known structure and similar sequence.

Structural Genomics

- A world-wide consortium to determine novel protein structures.
- Which proteins are likely to be novel?

Outline

A whirlwind review of molecular biology
An overview of computational molecular biology
Thurs: Sequencing technology and whole genome sequence assembly
Next Tuesday: New problems in genomics