Computational Genomics

http://www.cs.cmu.edu/~02710

Ziv Bar-Joseph
zivbj@cs.cmu.edu
GHC 8006

Chakra Chennubhotla
chakracs@pitt.edu
Suite 3064, BST3
Topics

• Introduction (1 Week)
• Sequence analysis (4 weeks)
• Gene expression (3 weeks)
• RNA and epigenetics (3 weeks)
• Systems biology (3 weeks)
Class overview

- 4 problem sets
- Midterm
- Project (and poster)
- Class attendance and participation
Class grades

• Problem sets (40%)
• Midterm (30%)
• Project (25%)
• Class participation (5%)
High level and brief intro to molecular biology and genomics
Organism, Organ, Cell
Types of Cells

• Eukaryots:
  - Plants, animals, humans
  - DNA resides in the nucleus
  - Contain also other compartments

• Prokaryots:
  - Bacteria
  - Do not contain compartments
Cell signaling

- Cells communication is based on chemical *signals* & *receptors*
  - If you have the correct receptor, respond to signal; no receptor = no response
  - Single-celled organisms receive cues about the environment, status of other individuals
- Process termed the *signal transduction pathway*
  - From signal interacting with receptor to cellular response
Types of Signals

• **Local signaling**: short-distance
  – affect the cells that produce them
  – affect nearby cells (diffuse)

• **Hormonal signaling**: long-distance
  – Typically found in multicellular organisms & use circulatory system for distribution
**Autocrine** signals bind to receptors on the same cell that secretes them.

**Paracrine** signals bind to receptors on nearby cells.

Cells without receptors for a particular signal do not respond to that signal.

**Circulating signals** are transported by the circulatory system and bind to receptors on distant cells.

Target cell

Secreting cell

Circulatory vessel (e.g., a blood vessel)

Target cell
1. **Reception**: signal molecule interacts with receptor

2. **Transduction** typically several steps that involve changes to **responder** molecules and downstream targets

3. **Outcome**: often triggers a cellular response (effect)
Central dogma

DNA

transcription

mRNA

translation

Protein

CCTGAGCCAACTATTTGATGAA

CCUGAGCCACUAAUUGAUGAA

PEPTIDE
Genome

• A genome is an organism’s complete set of DNA (including its genes).
• In humans, less than 3% of the genome actually encodes for genes.
• However, a much larger % of the genome is transcribed (miRNAs, lincRNAs, ...)
• And a large part of the rest of the genome serves as a control regions.
# Comparison of Different Organisms

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size</th>
<th>Num. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>.05*10^8</td>
<td>4,200</td>
</tr>
<tr>
<td>Yeast</td>
<td>.15*10^8</td>
<td>6,000</td>
</tr>
<tr>
<td>Worm</td>
<td>1*10^8</td>
<td>18,400</td>
</tr>
<tr>
<td>Fly</td>
<td>1.8*10^8</td>
<td>13,600</td>
</tr>
<tr>
<td>Human</td>
<td>30*10^8</td>
<td>25,000</td>
</tr>
<tr>
<td>Plant</td>
<td>1.3*10^8</td>
<td>25,000</td>
</tr>
</tbody>
</table>
Genes
What is a gene?

Promoter                  Protein coding sequence                  Terminator

Genomic DNA
Example of a Gene: Gal4 DNA

ATGAAGCTACTGTCTTCTATCGAACAAGCATGCGATATTTGCGACTTAAAAAGCTCAAG
TGCTCCAAAAGAAAAAACCGAAGTGCACGCTCGGACTAGAACACATCTGACGACAGAAGTGGAATCAAG
CTCTCCAAAACACAAAAGGTCTCCGCTGACTAGGGCACAATCTGACGACAGAAGTGGAATCAAG
CTAGAAAGACTGGAACAGCTATTTTCTACTGATTTTTTCTCCTGAGAAAGACCTTGGACGATGATT
TTGAAAAATGGATTCTTTACAGGATATAAAAAGCATTGTTAAACAGGATTATTTGATACAAGAT
AATGTCGAAATAAGAGATGCCTGGTCACAGATAGATTGGCTCGTCAGTGGAAGACTGATATGCTCTTA
ACATTGAGACAGCATAGAATAAGTGCGACATCATCATCGGAAGAGAGTAGTAGAACAAGAGGT
CAAAGACAGTTGACTGTATCGGACTCGGACTCGGACTCGTATCGATAGATTGGCTCGAAGGATTGACATG
TTGGATTATGCCCAGGGATGTCTCTTATGATTGGTTGCTGTGCAAGGATTGACATG
TCGGATGGGCTTTGCCCTTCCCTGAAAACGAGACCCCAACCATATAATGGGTTCCTTGGCGACGCT
TCTCTCTATGTATTCTCTGATCTATTGGCTTTAAACCGGAAAATACACCGAACCTCTA
GTTAACAGGCTCCCGACCATGATTACCGGATAGATACGCTTGGCTCTAGATCCACAACA
TCCCCCTTACTCTAAGTTATCTCAAATATTTTACCCCTACTGCCCCCTATCGGTCACTCA
CCGACGCTAATGATGTTGTATATAAACCAGATTGGAAATCGCCTCGAAGGATCAATGGCAA
ATCCTTTTTAACTGCTATATTAGCCATTGGGACCTGGGTATAGAGGGGAATCTACTGAT
ATAGATGTTTTTACTATCAAATGCTAAATCTCATTGGACGAGCAAGGTCTCTCAGCTCA
# Genes Encode for Proteins

The table below illustrates the relationship between DNA nucleotide triplets (codons) and corresponding amino acids, which are the building blocks of proteins.

<table>
<thead>
<tr>
<th>1st Letter</th>
<th>2nd Letter</th>
<th>3rd Letter</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>U</td>
<td>A</td>
<td>UGU</td>
</tr>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td>UGU</td>
</tr>
<tr>
<td>U</td>
<td>A</td>
<td>G</td>
<td>UGU</td>
</tr>
<tr>
<td>U</td>
<td>G</td>
<td>G</td>
<td>UGU</td>
</tr>
</tbody>
</table>

- **U** codes for **Leu** (Leucine) or **Phe** (Phenylalanine)
- **C** codes for **Pro** (Proline)
- **A** codes for **Thr** (Threonine) or **Ile** (Isoleucine) or **Met** (Methionine)
- **G** codes for **Val** (Valine)

Additional codons for **Stop** (UAA, UAG, UGA) and **Trp** (UGG) are also shown.

![Genetic Code Table](image)
Example of a Gene: Gal4 AA

MKLLSSIEQACDICRLKKLKCSEKPKCAKCLKNNWECRYSPKTKRSLTRAHLTEVESR
LERLEQLFLLIFPRDLDMILKMDSLQDIKALLTGLFVQDNVNDKAVTDLASVETDMPL
TLRQHRISATSSSSEEENKQQRQLTVSIDSAAHHDNSTIPLDMPRDLHGFWDWSEEDDM
SDGPLFLKTDPNNNGFFGDSLLCILRSIGFKPENYTNSNVNRPLMTIDRYTLSRSTT
SRLQSYLNNFHPYCPIVHSPTLMMLYNNQIEIAKDOQWQLFNCILAIGAWCIEGESTD
IDVFYYQNAKSHLTSKVFESGSIIIVTALHLSSRTQWRQKTNTSYNFLHSFIRMAISLG
LRDLPPSSFSSSILEQRRRIWWSVYWSWEIQSLLYGRISIQLSQTISFPSSVDDVQRTT
TGPTIYHGIETARLLQVFTEKDVTAEKSPICAKKCLMICNEIEEVRQAPKFLQ
MDISTTLALTNLKEPHWLFSFTRFELKWQLSLLYVLRDFTNFTQKKSLEQDQNDHQS
YEVKRCISMLSDAAQRVTMSVSSYMNDNHTVPYFAWNCYLFNAVLMVIPKTLSSNSKS
AENNETAQLLQINTVLMLLKLATFKIQTCKEYIQVLEEVCAPFLLSQCAIPLPHISYN
NSNGSAIKNIVSRTIAQYPTLPEENVNNSIVKYVSPGSGVSPVPLKSGASLDVSKLL
SNRPPSRNPSVTIPRSRSVTPFLGQQQLQLQLVPLTPSALFGGANFNQSGNIADSS
Number of Genes in Public Databases

Growth of GenBank
(1982 - 2008)

Number of entries in UniProtKB/TrEMBL

Sequences (millions)
Basé Pairs
Sequences

Base Pairs of DNA (kilome)
500,000
Structure of Genes in Mammalian Cells

• Within coding DNA genes there can be un-translated regions (Introns)
• Exons are segments of DNA that contain the gene’s information coding for a protein
• Need to cut Introns out of RNA and splice together Exons before protein can be made

• Alternative splicing increases the potential number of different proteins, allowing the generation of millions of proteins from a small number of genes.
Comparative genomics

Mouse and Human Genetic Similarities

Mouse chromosomes

Human chromosomes

Courtesy Lisa Stubbs
Oak Ridge National Laboratory
Regulatory Regions
Promoter

The promoter is the place where RNA polymerase binds to start transcription. This is what determines which strand is the coding strand.
DNA Binding Motifs

• In order to recruit the transcriptional machinery, a transcription factor (TF) needs to bind the DNA in front of the gene.

• TFs bind in to short segments which are known as DNA binding motifs.

• Usually consists 6 – 8 letters, and in many cases these letters generate palindromes.

• Note however that TF binding requires an open chromatin (a set of proteins that pack the DNA). Several factors are general chromatin modifiers. ‘Chicken and egg’ problem.
Epigenetics
Messenger RNAs (mRNAs)
RNA

Four major types (one recently discovered regulatory RNA).

• mRNA – messenger RNA
• tRNA – Transfer RNA
• rRNA – ribosomal RNA
• shRNA, microRNA – RNA interference
Messenger RNA

• Basically, an intermediate product
• Transcribed from the genome and translated into protein
• Number of copies correlates well with number of proteins for the gene.

• Unlike DNA, the amount of messenger RNA (as well as the number of proteins) differs between different cell types and under different conditions.
Complementary base-pairing

- mRNA is transcribed from the DNA
- mRNA (like DNA, but unlike proteins) binds to its complement

Activators

Transcription apparatus

mRNA

Gene

RNAPII

mRNA

label

hybridization

AUGC

UACG
Hybridization and Scanning—Microarrays

RNASeq using next generation sequencing methods
Perturbation

• In many cases we would like to perturb the systems to study the impacts of individual components (genes).
• This can be done in the sequence level by removing (knocking out) the gene of interest.
• Not always possible:
  - higher organisms
  - genes that are required during development but not later
  - genes that are required in certain cell types but not in others
Perturbations: RNAi
Proteins
From RNA to proteins: The Ribosome

• Decoding machine.

• Input: mRNA, output: protein

• Built from a large number of proteins and a number of RNAs.

• Several ribosomes can work on one mRNA
The Ribosome
Proteins

- Proteins are polypeptide chains of amino acids.
- Four levels of structure:
  - Primary Structure: The sequence of the protein
  - Secondary structure: Local structure in regions of the chain
  - Tertiary Structure: Three dimensional structure
  - Quaternary Structure: multiple subunits
Secondary Structure: Alpha Helix
Secondary Structure: Beta Sheet
Protein Structure
Domains of a Protein

- While predicting the structure from the sequence is still an open problem, we can identify several domains within the protein.

- Domains are compactly folded structures.

- In many cases these domains are associated with specific biological function.
Protein Interaction

In order to fulfill their function, proteins interact with other proteins in a number of ways including:

• Regulation

• Pathways, for example A -> B -> C

• Post translational modifications

• Forming protein complexes
Putting it all together: Systems biology
High throughput data

Time-series measurements
- gene expression
- epigenetics

Static data sources
- sequence
- motif
- CHIP-Seq
- PPI

RNA-Seq
- mRNA
- Short reads

RNA-Seq
- motif
- CHIP-Seq
High throughput data

- We now have many sources of data, each providing a different view on the activity in the cell
  - Sequence (genes)
  - DNA motifs
  - Gene expression
  - Protein interactions
  - Image data
  - Protein-DNA interaction
  - Etc.

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology
Reverse engineering of regulatory networks

Segal et al. *Nature Genetics* 2003

- Gene expression
- Protein-DNA and gene expression


Figure 1: Rich media gene modules network
Dynamic regulatory networks

Protein-DNA, motif and time series gene expression data