EE381V: Genomic Signal Processing

Basic Information

- **Instructor:** Haris Vikalo
  - E-mail: hvikalo@ece.utexas.edu
  - Phone: (512) 232-7922
  - Office: ACES 3.110
  - Hours: Mon, Wed, 2:00pm-3:00pm

- **Electronic course site:** Blackboard
  - courses.utexas.edu
  - distribution of homework assignments, solutions, and class notes
  - should be able to access it if you have UT EID and are registered

- **Course website:** [http://users.ece.utexas.edu/~hvikalo/ee381v.html](http://users.ece.utexas.edu/~hvikalo/ee381v.html)
  - class notes (mirrored from Blackboard) and suggested reading
  - final project information

- **Lectures location & time:** CLA 0.106, Mon, Wed 11:00-12:30pm
  - may have some guest lectures, not necessarily in the same room
Basic Information

- **Textbook:** none
  - class notes, reading assignments will be distributed via course website, Blackboard

- **Optional reading (on reserve desk in Life Sciences Library):**

- **Homeworks & Exams:**
  - bi-weekly homeworks (algorithmic rigorous thinking, programming assignments)
  - midterm (probably take-home)
  - final project (tackle a research problem, write-up a report)

- **Grading (tentative):**
  - homeworks (30%), midterm (30%), final project (40%)

- **Prerequisites:** EE381J Probability and Stochastic Processes or equivalent
  - exposure to differential equations beneficial
  - no biology background is required
  - familiarity with Matlab beneficial (to carry-out programming assignments)

Goals for the Term

- **Introduction to genomic signal processing**
  - fundamental problems in genomic signal and information processing
  - research directions for active participation in the field

- **Duality: computation and biology**
  - give a biology/technology background to motivate a computational task
  - provide background on the relevant signal processing / computational techniques
  - describe a solution

- **Foundations and frontiers**
  - well defined conventional problems and general methodologies
  - contemporary challenges, future research directions, etc.

- **Scope of the topics**
  - core biotechnologies: modeling and signal processing algorithms
  - cellular systems: algorithmic/computational tools for inferring their structure and understanding how they function
### Signal Processing for Core Biotechnologies

Systems for sequencing and detection:

<table>
<thead>
<tr>
<th>DNA Sequencing</th>
<th>Gene Expression Profiling</th>
<th>DNA Amplification</th>
</tr>
</thead>
</table>
| ![ABI Prism 
310 Genetic Analyzer](image1) | ![Affymetrix GeneChip 
®](image2) | ![Roche LightCycler 
®](image3) |

### Signal Processing for Cellular Systems

- Information flow in a cell (traditional view: Central Dogma):

  - Information (signal) is carried by molecules.
**Signal Processing for Cellular Systems**

- Follow the information flow:

  ![diagram](DNA -> RNA -> Protein)

  - **DNA**
  - **RNA**
  - **Protein**
  - **Mechanisms**

- Moreover, study the temporal changes in the information flow
  - gives insight in regulation mechanisms, biological network structure, etc.
  - Previously mentioned biotechnologies interrupt the information flow and so provide insight into the cellular structure and functions

---

**Computational and Signal Processing Challenges**

- Sequencing and Genome assembly
- Regulatory motif discovery
- Comparative genomics
- Evolutionary theory
- Gene expression analysis
- Cluster discovery
- Gene finding
- DNA
- Sequence alignment
- Database lookup
- Regulatory networks inference
- Protein network analysis
- Emerging network properties

---

**Sequence Alignment**

ACATGCTAT
ACGTGATAA
AGAGGATAT
ATATCATAT
ATATGATTT

**Gene Finding**

**Comparative Genomics**

**Evolutionary Theory**

**Gene Expression Analysis**

**Cluster Discovery**

**Regulatory Networks Inference**

**Protein Network Analysis**
Computational and Signal Processing Challenges

- Sample topics and computational / signal processing tools:
  - Sequencing and sequence analysis
    - modeling with hidden Markov models (HMM)
    - many problems require dynamic programming solutions
  - Technologies (systems) for bio-molecular detection
    - modeling with continuous-time Markov processes (discrete, stochastic),
      often use approximations (continuous-valued, deterministic)
    - estimation techniques for data recovery
  - Gene expression analysis / Network discovery
    - various data mining techniques
  - Network modeling and analysis
    - modeling with (multiple) continuous-time Markov processes, graph models
      (Boolean, Bayesian networks)
    - Monte Carlo simulation techniques, network inference
Computational and Signal Processing Challenges

• Recent IEEE special issues (can be accessed via IEEE Xplore):

• Today’s goal: Molecular Biology Primer
  • will be complemented by a few papers posted to Blackboard

Biological Systems

Organisms are remarkably uniform at the molecular level.

Molecules  Macromolecules  Tissue/Cell  Organ  Organism
Biological sciences study different morphology levels of biological systems.

### Biological Systems

- **Molecules**
- **Macromolecules**
- **Tissue/Cell**
- **Organ**
- **Organism**

#### Biophysical Levels

- **Biophysics**
- **Genomics**
- **Molecular Biology**
- **Biochemistry**
- **Cell Biology**

### Biological Systems: DNA Molecules

- In eukaryotes, DNA is tightly packaged into the structures called chromosomes, inside the nucleus of a cell.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>pea plant</td>
<td>14</td>
</tr>
<tr>
<td>sunflower</td>
<td>34</td>
</tr>
<tr>
<td>cat</td>
<td>38</td>
</tr>
<tr>
<td>puffer fish</td>
<td>42</td>
</tr>
<tr>
<td>human</td>
<td>46</td>
</tr>
<tr>
<td>dog</td>
<td>78</td>
</tr>
</tbody>
</table>

- Potato has 48 of them, goldfish 94
- Prokaryotes (e.g., bacteria): only a single loop of stable chromosomal DNA
Structure of DNA

- Four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T)
- Forms a double helix – each strand is linked via sugar-phosphate bonds, strands are linked via hydrogen bonds
- sugar-phosphate bonds are strong, hydrogen weak

Structure of DNA: Nucleotides

- Structure of adenine (one of the nucleotides):
  - Symbolically:
  - Base pairing (A with T, C with G):
Structure of DNA: Backbone

- What about the backbone?
  - Sugar-phosphate backbone is directional (5'→3' or 3'→5')
  - Why it matters: enzymes typically “care” about the direction

Structure of DNA

- Human Genome has 3.2 billion DNA base pairs
- 3.2 billion \((3.2 \times 10^9)\) symbols:
  - 200 (1000 pages each) NYC phone books
  - 800Mb (roughly, a data CD)
  - a person typing 60 words/minute for 8 hours/day, would take more than 50 years to type the entire human genome sequence
  - placed end-to-end the DNA in one human cell extends almost 6 feet
  - if all the DNA in a body were connected this way, it would stretch approx. 67 billion miles!
    - 150k round trips to the moon, 70 to the sun
- DNA stores hereditary information
  - copied/replicated during the cell reproduction process
DNA Replication

- During the cell reproduction process, DNA replicates
  - The twisted, compacted double helix of DNA has to unwind and separate its two strands
  - Each strand becomes a pattern, or template, for making a new strand, so the two new DNA molecules have one new strand and one old strand
  - The copy is done by a cellular protein machine called DNA polymerase, which reads the template DNA strand and stitches together the complementary new strand
  - The process of replication is astonishingly fast and accurate, although occasional mistakes, such as deletions or duplications, occur.

Agent of DNA Replication: DNA Polymerase

- DNA polymerase adds free nucleotides to the 3' end of the new strand
  - so, the new strand grows in a 5'-3' direction
  - requires a “primer” (short sequence) to initiate extension

- New strands grow in 5’-3’ direction
Mistakes in DNA Replication

- A built-in proof reading system (mismatch-pair system) catches and corrects nearly all of these errors
  - DNA replication: 1 error per 1 billion bases
- Mistakes that are not corrected can lead to diseases such as cancer and certain genetic disorders
  - Fanconi anemia, early aging diseases, etc.
- A sidenote: many drugs used to treat cancer work by attacking DNA replication
  - chemotherapy drugs disrupt the DNA copying process, which goes on much faster in rapidly dividing cancer cells than in other cells
  - side-effect: most of these drugs do affect normal cells that grow and divide frequently, such as cells of the immune system and hair cells

Central Dogma

Stated by Francis Crick in 1958, re-stated in a *Nature* paper published in 1970:

"The central dogma of molecular biology is based on the principle that the flow of genetic information travels from DNA to RNA and finally to the translation of proteins."

- Genes carry hereditary information but do not do any actual work in cells
  - they serve as instruction books for making functional molecules such as ribonucleic acid (RNA) and proteins
  - two steps: transcription and translation
Transcription

- In transcription, the information coded in DNA is copied into RNA
  - the RNA nucleotides are complementary to those on the DNA
  - note: RNA pairs a uracil (U), instead of a T, with an A on the DNA

RNA Polymerase

- Like DNA polymerase, RNA polymerase adds free nucleotides to the 3’ end of the new strand
  - only one DNA strand is copied
  - the new strand grows in a 5’-3’ direction

- The resulting mRNA sequence:

- This direction is preferred for energy reasons…
Transcription

- All cells contain the same DNA
  - so, what makes a nerve cell different from a red blood cell?

- Each cell "turns on," or expresses, only a subset of genes

- Activity of RNA polymerase is affected by a number of proteins
  - these proteins vary in different cell types throughout the body

- Note: gene expression levels are affected by diseases

Translation

- In translation, messenger RNA (mRNA) is mapped to a specific protein (string of amino acids) according to the rules specified by the genetic code

- A four-letter alphabet is mapped to a 20-letter alphabet
  - there is an embedded redundancy
Translation

• During translation, a four-letter alphabet is mapped to a 20-letter alphabet:

DNA Sequence Length = 324 bases  RNA Sequence Length = 324 bases  Amino acid length = 108 bases

DNA  RNA  Protein

Translation

• Simplified version of the translation mechanism:
This is a fairly good description of the translation process in prokaryotes (give or take a few details omitted for simplicity).

However, the process is much more complicated in eukaryotes.

In eukaryotes, the region of DNA coding for a protein is usually not continuous. This region is composed of alternating stretches of exons and introns:
- exons: pieces of coding sequence
- introns: regions between exons

To get an mRNA molecule which is mapped to a working protein, the intron sections are trimmed and exon pieces stitched together: RNA splicing.

If inaccurate, splicing may lead to an abnormal protein or no protein at all:
- a form of Alzheimer’s disease is caused by this.
**Translation in Eukaryotes**

- **Alternative splicing**: arranging exons in different patterns
  - enables cells to make different proteins from a single gene.

![Gene structure diagram]

- Leads to possibility of creating many proteins from a single gene
  - e.g., 25k human genes can make hundreds of thousands of different proteins

---

**Translation in Prokaryotes**

Detailed description of the translation process in prokaryotes is complicated:
Translation in Eukaryotes

Detailed description of the translation process in eukaryotes is even more complicated:

Central Dogma Revisited

- The traditional view:

  - DNA \rightarrow RNA \rightarrow Protein

  1. Very stable molecule
  2. Very large length
  3. Only 4 building blocks

- However, there is feedback, creating a control system:

  - DNA \rightarrow RNA \rightarrow Protein

  1. Unstable molecules
  2. A fragment of DNA length
  3. Only 4 building blocks

  1. Complex molecules with huge variety of shapes and forms
  2. Proportional to RNA size
  3. Has 20 building blocks

External Stimuli
Central Dogma Revisited

- Feedback, creating a control system:

![Diagram of Central Dogma]

- We are interested in information/signals in these complex, nonlinear, and probabilistically described biomolecular systems with feedback.

Central Dogma Revisited

- The signal/information in these systems is carried by bio-molecules
  - so, we may be interested in their structure, amounts, interaction…

![Diagram of Central Dogma with additional elements]
Cell as a Control System

- Information/signals are carried by molecules:
  - Transcription
    - DNA
    - RNA Polymerase
    - RNA
    - Ribosome
  - Translation
    - Protein
  - Biochemical Biological Functionality

- Signals are controlled via feedback
  - control allows adaptability to varying conditions
  - again, molecules facilitate the control

Controlling Transcription Mechanism

- May be rather complex, involves transcription factors (proteins) which bind to promoter regions (upstream of a gene):
  - Selective binding of enzyme to promoter region
  - Synthesizing RNA based on the DNA template
  - Enzyme detachment at termination region

- Transcription can be upregulated/downregulated
Controlling Transcription Mechanism

Transcription factors bind to promoters and recruit RNA polymerase:

- Since located close to the beginning of a gene, promoter regions are indicative of genes locations on DNA

End of a (brief) molecular biology primer

Next: DNA Sequencing