Bioinformatics
A Communication/Signal Processing Perspective

Khalid Sayood
Occult Information Lab
Department of Electrical and Computer Engineering
University of Nebraska-Lincoln
Outline

- Beginning
- Middle
- End
My roots

• Signal Processing
• Communication
• Information Theory
• Data Compression
My roots

• Data Compression - The art or science of finding compact representations of information in data
• Speech ➔ Images ➔ Video ➔ DNA
Prokaryotes

Analogy to Communications: Start, Message, Stop

Eukaryotes

DNA
RNA transcript with cap and tail
mRNA

DNA molecule

NUCLEUS
CYTOPLASM

Gene 1
Gene 2
Gene 3

Protein-coding sequence
Stop Signal
Intergenic Sequence
Promoter

Addition of cap and tail

Introns removed

Exons spliced together

Coding sequence

Transcription

Start, Message, Stop
Definitions of Bioinformatics

• Original definition: Study of informatic processes in biotic systems – Pauline Hogeweg and Ben Hesper
Definitions of Bioinformatics

• Original definition: Study of informatic processes in biotic systems – Pauline Hogeweg and Ben Hesper

• Bioinformatics is conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying “informatics” techniques (derived from disciplines such as applied math, CS, and statistics) to understand and organize the information associated with these molecules, on a large-scale. (Mark Gerstein, 1999)
Bioinformatics is a management and analysis information system for life sciences.

- **Data Storage and Management**
  - Protein Structure Prediction
    - Protein/RNA tertiary structure
    - Docking
    - Drug Design
  - Molecular Sequence Analysis
    - Homology Search
    - Phylogeny Construction
    - Whole Genome Sequencing
    - Gene Finding
  - Functional Genomics and Proteomics
    - Microarrays
    - Biomarker Discovery
- **Data Analysis**
- **Interpretation of Results**
- **Systems Biology**
  - Pathways
  - Network based wholistic approach
From RNA-seq reads to differential expression results

Alicia Oshlack, Mark D Robinson and Matthew D Young
Preprocessing:
  *FastQC*
  *Trimmomatic*

Indexing the Reads:
  *Bowtie 2*

Aligning to hg19/KSHV
  *TopHat 2*
    Local alignment

Extracting Features
  *CuffDiff*
    P < 0.01
    FPKMs values

Differentially Expressed Genes
So what is the problem?

• What is actually going on.
Is there hope (for me)?

Hope springs eternal

Digital Camera (1975)

Genome Sequencer (2018)
DON'T WORRY. TECHNOLOGY
WILL SAVE YOU.
What do we mean by a Communication Theory Perspective
What do I mean by a Communication Theory Perspective

Information exists in the form of a stochastic process
How do we deal with stochastic processes?

• Look at the signal using different basis sets – frequency domain processing.
• Look at correlation structures.
• Look at models.

All these involve averaging of some sort

All these result in the discovery of underlying structure

They can also result in dimensionality reduction.
Realizations of a stochastic process
Frequency profile
Statistical Profile
Models

Model Order 3

Model Order 5

Model Order 11

Model Order 21
But we have sequences – we don’t have numbers

GAGACATTCAAGTG

G
C
T
A

GAGACATTCAAGTG
But we have sequences – we don’t have numbers

```
GAGACATTCAGT
A
C
G
T
```
But we have sequences – we don’t have numbers

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Identification of Protein Coding Regions Using the Modified Gabor-Wavelet Transform

But we have sequences – we don’t have numbers

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>T</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
But we have sequences – we don’t have numbers

\[ I_k = \sum_{X \in A} \sum_{Y \in A} p_k(X,Y) \log \left( \frac{p_k(X,Y)}{p(X)p(Y)} \right) \]
AMI Profile for Human Chromosome 1
Human chromosomes
Mouse Chromosome
C. elegans Chromosome

Graphs showing the average mutual information for various chromosomes and the combination of all chromosomes.
• Phylogeny for *Candida* and *Saccharomyces* clades based on multiple sequence alignment of 706 orthologous genes

• Posterior probabilities shown

• WGD: Whole Genome Duplication

• CTG: Translation of CTG codons as serine rather than leucine
• Ribosomal DNA (rDNA) is commonly used to evaluate species relatedness

• The rDNA gene complex contains 3 genes, each of which are ribosomal components once transcribed

• Internal transcribed spacer (ITS) 1 and ITS2 separate these genes

• ITS regions have 2 benefits:
  1. Easy to design primers (ribosome genes highly conserved, many copies)
  2. Spacers diverge more quickly than ribosome genes
• Distance matrix $D$ generated by calculating pairwise distance $d_{ij}$ between AMI profiles $x_i$ and $x_j$

• Distance defined in two ways:
  1. Correlation distance (angle between profiles)
     \[ d_{ij} = 1 - \cos \theta = 1 - \frac{x_i \cdot x_j}{\|x_i\| \|x_j\|} \]
  2. Euclidean distance
     \[ d_{ij} = \|x_i - x_j\| \]

• Phylogenetic trees generated using PHYLIP (neighbor joining)
Candida albicans
Candida dubliniensis
Candida tropicalis
Candida parapsilosis
Lodderomyces elongisporus
Candida guilliermondii
Debaryomyces hansenii
Candida lusitaniae
Saccharomyces cerevisiae
Saccharomyces paradoxus
Saccharomyces mikatae
Saccharomyces bayanus
Candida glabrata
Saccharomyces castellii
Kluyveromyces lactis
Ashbya gossypii
Kluyveromyces waltii
Yarrowia lipolytica

C. glabrata
S. castellii
S. cerevisiae
S. paradoxus
S. bayanus
S. mikatae
K. lactis
A. gossypii
C. tropicalis
C. albicans
C. dubliniensis
C. lusitaniae
C. parapsilosis
L. elongisporus
D. hansenii
Y. lipolytica
GO Prediction

"High Abundance" GO terms

"Low Abundance" GO terms

BP: Biological Processes, CC: Cellular Component, MF: Molecular Function
Slow progressor populations

Rapid progressor populations
Metagenomics

Who is in there?

How are they doing it?

What are they doing?
Similarities between genomes + single read processing = detection errors
Learn channel statistics to estimate true abundance given the observed detection.

\[ p_{i,j} = \frac{|G_i \cap G_j|}{|G_j|} \]
RRMSE for the simulated metagenomes corresponding to a mixture of 10, 20, 50, and 100 randomly selected organisms for 0.01X and 0.1X average genome coverages.
DAY 44:
STILL STRANDED, WITH
NOTHING BUT FLAT EMPTY
WATER AS FAR AS THE
EYE CAN SEE.
Once, I thought I was wrong but it turns out...I was mistaken.