Computational Genomics

Selections from the Menu

Lenwood S. Heath

Department of Computer Science
Virginia Tech

May 2, 2015
Outline

Some Biology

Alternative Splicing

Genome Alignment

Life Identification Numbers

Viral Genomics
Some Biology

Alternative Splicing

Genome Alignment

Life Identification Numbers

Viral Genomics
Genomes

- **Genome**: the total DNA sequence of an organism
- **Nucleotide**: character in the sequence
- **Part of the human genome**: 

  ```
  ...ATGGCGACGAAGGCCGTGTCGTGCTGAAGGGAAGGGCGACGGCCCATCAATTTCGAGC
  AGAAGGCAAGGGCTGAGGACGGAAGGCTTTGCTGGA
  GGCGTGCTCCACCACCCCGCTCGTCCCGGCCACGGAC
  GGCCGCTCCCAGTGCAGGGCATCATCAATTTCGAGC
  AGAAGGCAAGGGCTGAGGACGGAAGGCTTTGCTGGA
  GGCGTGCTCCACCACCCCGCTCGTCCCGGCCACGGAC
  GGCCGCTCCCAGTGCAGGGCATCATCAATTTCGAGC
  AGAAGGCAAGGGCTGAGGACGGAAGGCTTTGCTGGA
  GGCGTGCTCCACCACCCCGCTCGTCCCGGCCACGGAC
  GGCCGCTCCCAGTGCAGGGCATCATCAATTTCGAGC
  AGAAGGCAAGGGCTGAGGACGGAAGGCTTTGCTGGA
  GGCGTGCTCCACCACCCCGCTCGTCCCGGCCACGGAC
  GGCCGCTCCCAGTGCAGGGCATCATCAATTTCGAGC
  ```
Central Dogma of Molecular Biology

- **Gene**: DNA sequence that is *transcribed* to **messenger RNA** (mRNA)
- **Protein**: mRNA is *translated* to protein according to the genetic code

Gene (DNA) $\xrightarrow{\text{transcribe}}$ mRNA $\xrightarrow{\text{translate}}$ protein
Table of Contents

Some Biology

Alternative Splicing

Genome Alignment

Life Identification Numbers

Viral Genomics
Open reading frame of a gene is made up of **exons** and **introns**

After transcription, introns are **spliced** out of the mRNA

**Alternative splicing**: exons and introns may change from one transcription to the next!
Control of Alternative Splicing

- **Splicing regulatory elements**
  - Specific short sequences in exons or introns
  - Typically 4–20 nucleotides long

```
GCGATCTGGGTGTCATAGTGGA
GGAAGA
TCCAGCTCACACCACCACAT
ATGACGCCCTCTAGTGGCTCTAGA
```

- Bound by proteins called **splicing factors**
- Binding may enhance or silence splicing
Minigene Approach

- **Experimental data generated:** Ke et al. 2011
- Experiments for all $4^6 = 4096$ length 6 DNA sequences (6-mers)
- Measured level of inclusion of exons in mRNA
- **Scored and ranked these 6-mers** with respect to inclusion
- **Top ranked 6-mers** are good candidates for enhancing splicing regulatory elements
De Bruijn Graph Representation

- Node set consists of all 6-mers
- Edges connect “adjacent” 6-mers

Build subgraph consisting of top-ranked splicing regulatory elements
- GenSRE algorithm uses graph search to find longer enhancers
Results

- Discovered 53,984 potential exonic splicing enhancers
- Lengths 6–87 nucleotides
- Word count enrichment analysis reduced number to 1500–2000
- Many sequences confirmed by independent experimental databases
Table of Contents

Some Biology

Alternative Splicing

Genome Alignment

Life Identification Numbers

Viral Genomics
Evolution Acts on Genomes

- Single nucleotide mutations
  
  CGACGGCCCAGTGCAGGGCATCATCAATTTCGAGC
  CGACGGCCCAGTACAGGGCATCATCAATCTCGAGC

- Insertions and deletions
  
  CGACGGCCCAGTGCAGGGCATCATCAATTTCGAGC
  CGACGGCCCAGTGCAGGGCAT---CATCAATTTTCGAGC

- Repeats
  
  CCGGTCGGTGCCCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTGCGTGCGGTGCC
  CCGGTCGGTGCCCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTGCGTGCGGTGCC

- Gain and loss of genes

- Genomic rearrangements
Comparing Two Genomes

- Simple sequence similarity — sequence alignment
  
  CGACGGCACAAGT--AG-GCATCATCAAT--TCGAGC
  CGACGGCCCAGTGCAGGGCAT---CAATTTCGATC

- Gene content

- Average nucleotide identity (ANI)

- Challenge of rearrangements
Our Approach to Whole Genome Alignment

- **Graph representation:** nodes are nucleotides

- Adjacency edges within a genome (blue)
- Edges demonstrating **common descent**, both **forward** and **reverse**
Results

• Breakable arrangements: ability to recursively break alignments into blocks
• Dynamic programming algorithm: generates alignment graph from two genomes
• Algorithm is optimal for breakable arrangements
• Demonstrated that many real-world examples are breakable
Table of Contents

Some Biology

Alternative Splicing

Genome Alignment

Life Identification Numbers

Viral Genomics
Species Naming

- **Linnaean taxonomy**: hierarchy of life on earth
- Genus-species binomial naming, e.g., *Homo sapiens*
- Naming a new species requires time for acceptance and publication
- Does not address variation below the species level, e.g., variation among individuals
Life Identification Numbers™ (LINs™)

- Leverages increasing availability of fully sequenced genomes for individual organisms
- Uses average nucleotide identity (ANI) for comparison
- Creates a hierarchy based on ANI percentage identity between a pair of genomes
LIN™ Example

- Based on 20%, 40%, 60%, 80%, 90%, 95%, 99% identity

\[
\begin{align*}
0_A2_B57_C14_D5_E29_F4_G \\
0_A2_B57_C14_D5_E29_F6_G \\
0_A2_B57_C14_D5_E29_F4_G \\
0_A2_B57_C8_D31_E7_F11_G
\end{align*}
\]

- **Advantages:**
  - A LIN™ can be generated quickly for a newly sequenced genome
  - LINs™ are meaningful in the context of other genomes
Validation of LINs™

- Prototype LIN™ assignment pipeline developed in Java
- Found to match well to phylogenetic relationships for genomes already investigated:
  - γ proteobacteria genomes
  - Human mitochondrial genomes
  - Ebola virus genomes
Commercialization

- **Startup TGL, Inc.**
- Founders: **Heath** and **Vinatzer**
- Plan **Web service** to provide LINs™ for biologists and medical professionals, e.g., epidemiologists
Nature of Viral Genomes

- Can be DNA or RNA
- Small genomes, carry just a few genes
- Interact with the replication machinery of host cells
- Genomes of RNA viruses are particularly susceptible to mutations
- **Approximately one nucleotide mutation per genome per replication**
- In an individual host, there are many viral genome variations
Quasispecies and Hypercubes

- Collection of viral genome variations in a population is called a **quasispecies**
  - Basic unit of evolution for viruses
- **Fitness**
  - Relative ability to survive and replicate
  - Changes with mutations — most variations die out
- **Model:** $d$-dimensional 4-ary hypercube
  - Nodes are the $4^d$ DNA sequences of length $d$
  - Edge between two nodes if they differ in one position
- **Evolution on a hypercube, moderated by fitness**
Proposed Research

• **Conserved sequences** among viral species
• **Rapid identification** of viral genomes with LINs™
• Characterize **fitness landscapes** in Ebola genome
• Characterize **transmission paths** of the evolving Ebola genome
Acknowledgements

- Eman Badr
- Nahla Belal
- Haitham Elmarakeby
- Alex Kallam
- Eric Lovell
- Caroline Monteil
- National Science Foundation
- Hanaa Torkey
- Boris Vinatzer
- Alex Weisberg
References