Introduction

02-714 / 02-514
<table>
<thead>
<tr>
<th>Topic</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Exact string matching (Z-algorithm, Knuth-Morris-Pratt, Boyer-Moore, Rabin-Karp)</td>
</tr>
<tr>
<td>2</td>
<td>Advanced inexact matching (edit distance, alignment in linear space, Four-Russians' speedup, approximation algorithms for MSA, whole genome alignment)</td>
</tr>
<tr>
<td>3</td>
<td>Parallel string search</td>
</tr>
<tr>
<td>4</td>
<td>Suffix trees and arrays and their applications; Ukkonen's suffix tree construction algorithm</td>
</tr>
<tr>
<td>5</td>
<td>Subword graphs and their construction</td>
</tr>
<tr>
<td>6</td>
<td>String compression algorithms</td>
</tr>
<tr>
<td>7</td>
<td>Compressed self-indices (data structures that support fast searching and complete reconstruction of the full text in sublinear space). Burrows-Wheeler transform, the FM-index.</td>
</tr>
<tr>
<td>8</td>
<td>Read mapping (matching huge collections of substrings to reference strings); tools for doing this such as Bowtie, BWA, TopHat</td>
</tr>
<tr>
<td>9</td>
<td>Compressive genomics (i.e. doing analyses directly on compressed data): searching compressed collections of sequences; compariative assembly from compressed read databases</td>
</tr>
<tr>
<td>10</td>
<td>Genome assembly; Shortest superstring problem, Celera assember, de Bruijn-graph-based assembly; mixed assembly.</td>
</tr>
<tr>
<td>11</td>
<td>Isoform / transcript assembly (e.g. Cufflinks, Trinity)</td>
</tr>
<tr>
<td>12</td>
<td>Gene and isoform expression quantification (RSEM, eXpress, Jellyfish, Cufflinks)</td>
</tr>
</tbody>
</table>
String Algorithms in Biology

1. Genome assembly
2. Gene discovery
3. Understanding the origin of swine flu
4. Gene Expression
Genome assembly
Genome of the Cow

a sequence of 2.86 billion letters

enough letters to fill a million pages of a typical book.

TATGGAGCCAGGTGCCTGGGGCAACAAGACTGTGGTCACCTGAATTCATCCTTCTTGTCTAACAGAGAACATAG
AACTGCAATCCATCCTTTTTTGCATCTTCTCTCTTTTGCTATGTGATCACAGTCGGGGGCAACTTGAGTATCCTGG
GCCGCCATCTTTTGTGAGCCACATTCACCACACCCCCATGTACTACTTCCTGGGGAACCTTTCTCTGCTGGACAT
TGGGTCATCAGTGTCCACATTTCTCCTCCCACTGCTGACCTGTTGACCTGACGTCCAGGGAGCCCTGGTGGCCGTCT
GCTGCTCCATCTCCTTCATCAATGCTCTGACCCACAGTGGCTGTGTCTGTGCTGGACTTCTGCCCTAACGTGGT
CAACCACTTCTACTGTGACCTCCCGCCCCTTTTCCAGCTCTCCTGCTCCAGCATCCACCTCACGGGCAGCTACTTT
TCGTGGGGGCCACCTTCATGGGGGTGGTCCCCATGGTCTTCATCTCGGTATCCTATGCCACGTGGCAGCCGCAGTC
CGGATCCGCTCGGCAGAGGGCAGGAAGAAAGCCTTCTCCACGTGTGGCTCCCATCACCGTGGTCTGCATCTTTTAT
GGGAACCGGCTTCTTCAGCTACATGCGCCTGGGCTCCGTGTCCGCCTCA
GACAAGGACAAGGGCATTGGCATCCTCAACACTGTCATCAGCCCCATGCTGAACCCACTCATCTACAGCCTCCG
GAACCCTGTGCTGACGGGCAGGGCCTGCTGCTGCCCTCCTCTGCTTCTCTTCTTCATG

ACGCTACCTTGCCATTTGCCAGCCCCCTACCTATAGCATCAGCTGGACCTCCCCTCGCTCCTCTCGCTACCAGCAC
ACTACCTCTCTTCATGCTGTCTTGAGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG

ACGCTACCTCTCACAGCATGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG
The Cat only grinned when it saw Alice. It looked good-natured, she thought: still it had very long claws and a great many teeth, so she felt that it ought to be treated with respect.

The Cat only grinned when it saw Alice. It looked good-natured, she thought, still it had very long claws and a great many teeth, so she felt that it ought to be treated with respect.

Fast algorithms are needed to piece the story together.
The Cat only grinned when it saw Alice.
     It looked good-natured, she thought: still
     It looked good-natured, she thought, still
     it had very long claws and a great many teeth, so she

     a greet many so she felt that it ought

     ought to be treated
     be treated with respect.

It’s a jigsaw puzzle ... ...except with 35 million pieces
Recent Genomics (DNA)

- First genome sequenced in 1995 (the bacteria *H. influenzae* with a genome of 1,830,140 letters).
- 1st draft of human genome finished in 2001 (~ 3 billion letters)
- Now: Over 1100 bacterial genomes
- Hundreds of higher-order genomes done or in progress.
- Several complete human genomes finished.
Researchers at many institutions are putting together the genomes of many animals.

Help understand how to make animals and plants more hardy, resistant to disease, and understand their biology.

New technologies and larger genomes require new algorithms and faster computers.
Other Sequenced Genomes

Arabidopsis thaliana

Saccharomyces cerevisiae (baker’s yeast)

Apis mellifera (honey bee)

Bos torus (cow)

Callithrix jacchus (marmoset)

Canis lupus familiaris (dog)

Drosophila melanogaster (fruit fly)

Equus caballus (horse)

Sequenced Eukaryotic Chromosomes

and many more...
Gene discovery
Obviously, computers are needed to understand what this means.

Where are the genes encoded in this sequence?
What causes each gene to be turned on or off?
How does the genome produce observed traits?
Two ways to find genes...

1. Search for sequence of DNA similar to known gene:

   Human version of gene

   Similar sequences in other organisms

   ⇒ need to search for strings really fast (allowing for errors)

2. Search the genome sequence for patterns of letters that “look like genes”.
Some Human Genes

http://useast.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000139618;r=13:32889611-32973805
Estimates for the # of Human Genes

Before human genome sequence was available, many (but not all) estimates for # of genes were high (> 80,000).

Now estimates are < 23,000.
• Subsequences of DNA are “genes” that encode (mostly) for proteins.

• # of genes in various organisms still not definitely know (because finding genes in the sequence is a hard problem that we will talk about).

• But there are reasonably good estimates.
Tracking the evolution of influenza
Influenza Virus

Rapidly evolving (it’s genome is mutating):
that’s why you have to get a different flu shot every year

3 strains must be selected each year to include in the vaccine.

So, the evolution of the virus is must be predicted.
Evolutionary Trees

http://www.amnh.org/exhibitions/darwin/idea/treelg.php
Gene Expression
Alternative Splicing & Isoform Expression

- Expression of genes can be measured via RNA-seq (sequencing transcripts)
- Sequencing gives you short (35-300bp length reads)
The Isoform Expression Estimation Problem

- RNA-Seq now standard for gene and isoform expression estimation.
- A main use for transcriptome sequencing is estimating gene and isoform abundance.
- This leads to the following computational problem:

**Given:**
- Collection of RNA-Seq reads
- A set of known transcript sequences

**Estimate:**
- The relative abundance of each transcript
Performance on Universal Human Brain Tissue

92,524,365 reads, each 35bp long

Bullard et al, 2010; MACQ Consortium, 2006
Big Genomic Data

Sequence Read Archive at the NCBI stores next generation sequence data.

It currently contains 1,587 terrabases of total data.