CS342: Bioinformatics
Lecture 5
Knuth-Morris-Pratt Algorithm

• Solves a version of the basic pattern matching problem.
• Rather than shifting $p$ by one at each iteration (brute-force), use info about $p$ to never go “backwards”.

**Input:** Text $t = t_0 \ldots t_m$ and pattern $p = p_0 \ldots p_n$ (0 index)

**Output:** Index of the first occurrence of $p$ in $t$.

• Step 1: Compute a table $T$ based only on pattern $p$ that tells us where the pattern contains potential repeats.
• Step 2: Use $T$ to search for the first occurrence of $p$ in $t$.

**Computing T**

• $T$ is table of size length of $p$. 
Knuth-Morris-Pratt Algorithm Analysis

Runtime? $O(n + m)$
- Build T? $O(n)$
- Search for match? $O(m)$
The Motif Finding Problem

- Given a random sample of DNA sequences:

  cctgatagacgctatctgctatccacgtacgtaggtcctctgtgcaatctatgcttttccaccat
  agtactcgtgtacattgtatagctacgtagctacgcacgccgaaaccctgaaacaaacgcgagccgacagtaat
  aacgtacgtagcaccctctttctttcttctggttcgctctgcgaacgaggctggtatatagacgaaat
  agcctccgatgtaagtcagctgtaactattacctgccaccccctattacatctttacgtacgtataca
  Cggttatataacgcgtagctgacgattgttttcgctcgctgacgtatctacgagttc

- Find the pattern that is implanted in each of the individual sequences, namely, the motif

- Additional information:
  - Assume the hidden sequence is of length 8
  - The pattern is not exactly the same in each sequence because random point mutations have been introduced
Motif Finding Example

• Finding motifs if there are no mutations
• Probability of a given 8-mer in an infinite sequence is $1/4^8 \approx 1.5 \times 10^{-5}$ (1 every 65Kb)
• Assuming 5 strings of length 68, there are $5 \times (68 - 8) = 300$ distinct 8-mers
• Probability of any one 8-mer is $300/4^8 \approx 0.005$
• So any repeat is rare

```
cctgatagacgctatcttgctatcc acgtacgt aggtcctctgtgcgaatctatgctgtttccaaccat
tagtcggtgtacattttgat acgtacgt acaccggcaacctgaaacaaacgcctcagaaccagagtgac
aaa acgtacgt gcacccctttttctctgcttgctgctgctggcacaagggctgtgatgaagacgaaaatttt
agcctccgatgtaagtcagctgtacacctgacctcactatctttacctgtgctcgtcgacgtc
```

acgtacgt
The Problem Becomes Harder

• Introduce 2 point mutations into each pattern:

\[
\begin{align*}
\text{cctgatagacgctatctgctatcc} & \quad \text{aGgtacT} \\
\text{agtactgtgtacat} & \quad \text{CcAtacgt}
\end{align*}
\]

\[
\begin{align*}
\text{agtcctggtgtacatttgat} & \quad \text{C} \\
\text{actggtgtacatttgat} & \quad \text{C}
\end{align*}
\]

• Our original target pattern no longer appears in any sequence!

Can we still find the motif?
Defining a Motif

• To define a motif, let’s assume that we know where the motif starts in each sequence

• The start positions can be represented as $s = [s_1, s_2, s_3, ..., s_t]$
Motifs: Profiles and Consensus

- Line up the patterns by their start indexes
  
  \[ s = (s_1, s_2, \ldots, s_t) \]

- Construct a matrix profile with the frequencies of each nucleotide in columns

- Consensus nucleotide in each position has the highest score in column

\[
\begin{align*}
\text{Alignment} & \quad \text{A} \quad \text{G} \quad \text{g} \quad \text{t} \quad \text{a} \quad \text{c} \quad \text{T} \quad \text{t} \\
& \quad \text{C} \quad \text{c} \quad \text{A} \quad \text{t} \quad \text{a} \quad \text{c} \quad \text{g} \quad \text{t} \\
& \quad \text{a} \quad \text{c} \quad \text{g} \quad \text{t} \quad \text{T} \quad \text{A} \quad \text{g} \quad \text{t} \\
& \quad \text{a} \quad \text{c} \quad \text{g} \quad \text{t} \quad \text{C} \quad \text{c} \quad \text{A} \quad \text{t} \\
& \quad \text{C} \quad \text{c} \quad \text{g} \quad \text{t} \quad \text{a} \quad \text{c} \quad \text{g} \quad \text{G} \\
\hline
\text{Profile} & \quad \text{A} & \quad 3 & \quad 0 & \quad 1 & \quad 0 & \quad 3 & \quad 1 & \quad 1 & \quad 0 \\
& \quad \text{C} & \quad 2 & \quad 4 & \quad 0 & \quad 0 & \quad 1 & \quad 4 & \quad 0 & \quad 0 \\
& \quad \text{G} & \quad 0 & \quad 1 & \quad 4 & \quad 0 & \quad 0 & \quad 0 & \quad 3 & \quad 1 \\
& \quad \text{T} & \quad 0 & \quad 0 & \quad 0 & \quad 5 & \quad 1 & \quad 0 & \quad 1 & \quad 4 \\
\hline
\text{Consensus} & \quad \text{A} \quad \text{C} \quad \text{G} \quad \text{T} \quad \text{A} \quad \text{C} \quad \text{G} \quad \text{T}
\end{align*}
\]
Consensus

• Think of consensus as an “ancestor” motif, from which mutated motifs emerged

• The *distance* between an actual motif and the consensus sequence is generally less than that for any two actual motifs

• *Hamming distance* is number of positions that differ between two strings
Consensus Properties

• A consensus string has a minimal hamming distance to all source strings.

\[ d = 4 \]

\[ d = 3 \]

\[ d = 2 \]

\[ d = 3 \]

\[ d = 2 \]

\[ d = 4 \]
Defining Some Terms

- **DNA** – array of sequence fragments
- **$t$** - number of sample DNA sequences
- **$n$** - length of each DNA sequence
- **$l$** - length of the motif ($l$-mer)
- **$s_i$** - starting position of an $l$-mer in sequence $i$
- **$s=(s_1, s_2, ..., s_t)$** - array of motif’s starting positions
Illustration of Terms

\[ l = 8 \]

\[ t = 5 \]

\[ n = 69 \]

\[ s_1 = 26 \quad s_2 = 21 \quad s_3 = 3 \quad s_4 = 56 \quad s_5 = 60 \]
Scoring Motifs

- Given $s = (s_1, \ldots s_t)$ and DNA:

$$\text{Score}(s, \text{DNA}) \leq \sum_{i=1}^{t} \text{Max} \ \text{count}(k, i)$$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Consensus: acgtacgt

Score: $3 \times 4 + 4 + 5 + 3 + 4 + 3 + 4 = 30$