Data structure: Lookup Table

Application: BLAST
The Look-up Table Data Structure

Definitions:
- A \textit{k-mer} is a string of length \( k \).
- A \textit{lookup table} is a table of size \(|\Sigma|^k\) that stores the location of all \( k \)-mers in the input sequence(s)

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\Sigma = \{\text{A,C,G,T}\} & k = 2 \\
S_1: \quad \text{C A G T C C T C} \\
\rightarrow 4^2 (=16) \text{ entries in lookup table}
\end{array}
\]

![Lookup table diagram](image-url)
Lookup Table - Application

- How to build lookup tables?
  - (see lecture notes)

- BLAST: basic local alignment search tool
Need for a Fast Alignment Method

- What to do with a newly found gene candidate, $s_{new}$?
- Locate “similar” genes in GenBank

**One Approach: (database search)**

1. Concatenate all sequences in our genomic database into one sequence, say $s_d$
2. Compute the local alignment between $s_{new}$ and $s_d$
3. Report all “significant” local alignments

Run-time: $O(|s_d|.|s_{new}|)$

Very long query time!!
Basic Local Alignment Search Tool (BLAST)

- Altschul et al. (1990) developed a program called BLAST to quickly query large sequence databases

**Input:**
- Query sequence $q$ and a sequence database $D$

**Output:**
- List of all significant local alignment hits ranked in increasing order of $E$-value (aka $p$-value, which is the probability that a random sequence scores more than $q$ against $D$).
The BLAST algorithm

0. **Preprocess:** Build a *lookup table* of size $|\Sigma|^w$ for all $w$-length words in $D$

\[ \Sigma = \{A,C,G,T\} \]

$w = 2$

$4^2 = 16$ entries in lookup table

Preprocessing is a one time activity
BLAST Algorithm …

1. **Identify Seeds:** Find all $w$-length substrings in $q$ that are also in $D$ using the lookup table

2. **Extend seeds:** Extend each seed on either side until the aggregate alignment score falls below a threshold
   - **Ungapped:** Extend by only either matches or mismatches
   - **Gapped:** Extend by matches, mismatches or a limited number of insertion/deletion gaps

3. **Record** all local alignments that score more than a certain statistical threshold

4. **Rank and report** all local alignments in non-decreasing order of $E$-value
Illustration of BLAST Algorithm

Ungapped Extension

Gapped Extension (over a band of diagonals)
## Different Types of BLAST Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Query</th>
<th>Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein/peptide</td>
<td>protein/peptide</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide</td>
<td>protein/peptide</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein/peptide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
</tbody>
</table>

Selected Bibliography for Alignment Topics

**Papers**


**Books**

Selected Bibliography for BLAST Related Topics

Serial BLAST

HPC BLAST
NCBI BLAST - Web Resources

- NCBI BLAST Webpage:

- For a comprehensive list of BLAST related references: