CAP 5510: Introduction to Bioinformatics

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS07.html
BLAST & FASTA

- FASTA
  [Lipman, Pearson '85, '88]

- Basic Local Alignment Search Tool
  [Altschul, Gish, Miller, Myers, Lipman '90]
Results of searches using different scoring systems may be compared directly using normalized scores.

If $S$ is the (raw) score for a local alignment, the **normalized** score $S'$ (in bits) is given by

$$S' = \frac{\lambda - \ln(K)}{\ln(2)}$$

The parameters depend on the scoring system.

**Statistically significant normalized score,**

$$S' > \log \left( \frac{N}{E} \right)$$

where $E$-value $= E$, and $N =$ size of search space.
String Matching Problem

Pattern $P$ \rightarrow \text{Set of Locations} $L$

Text $T$ \rightarrow
(Approximate) String Matching

**Input:** Text $T$, Pattern $P$

**Question(s):**
- Does $P$ occur in $T$?
- Find one occurrence of $P$ in $T$.
- Find all occurrences of $P$ in $T$.
- Count # of occurrences of $P$ in $T$.
- Find longest substring of $P$ in $T$.
- Find closest substring of $P$ in $T$.
- Locate direct repeats of $P$ in $T$.

*Many More variants*

**Applications:**
- Is $P$ already in the database $T$?
- Locate $P$ in $T$.
- Can $P$ be used as a primer for $T$?
- Is $P$ homologous to anything in $T$?
- Has $P$ been contaminated by $T$?
- Is $\text{prefix}(P) = \text{suffix}(T)$?
- Locate tandem repeats of $P$ in $T$. 
Input: Text T; Pattern P

Output: All occurrences of P in T.

Methods:

• Naïve Method
• Rabin-Karp Method
• FSA-based method
• Knuth-Morris-Pratt algorithm
• Boyer-Moore
• Suffix Tree method
• Shift-And method
Naive Strategy
Finite State Automaton

Finite State Automaton

ATAQAANANASPVANAGVERANANESISITALVDANANANANAS
## State Transition Diagram

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>S</th>
<th>*</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>AN</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANA</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>ANAN</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ANANA</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>ANANAS</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**Input:** Text $T$; Pattern $P$

**Output:** All occurrences of $P$ in $T$.

### Sliding Window Strategy:

1. Initialize window on $T$;
2. While (window within $T$) do
   - Scan: if (window = $P$) then report it;
   - Shift: shift window to right (by ?? positions)
3. endwhile;
Tries

Storing:
BIG
BIGGER
BILL
GOOD
GOSH

In this figure, the strings either start with B or G. Therefore, the root of the trie is connected to 3 edges called B, G and $.

LEAVES ARE GREEN. THE SYMBOL "$" TERMINATES EACH WORD.
Suffix Tries & Compact Suffix Tries

Store all suffixes of GOOGOL$
Suffix Tries to Suffix Trees

COMPACT TRIE OF SUFFIXES OF THE TEXT: GOOGOL$

Compact Suffix Trie

SUFFIX TREE

Suffix Tree

Key: G O O G O L $

1 2 3 4 5 6 7
Suffix Trees

- Linear-time construction!
- String Matching, Substring matching, substring common to $k$ of $n$ strings
- All-pairs prefix-suffix problem
- Repeats & Tandem repeats
- Approximate string matching
Multiple Alignments

- **Global**
  - ClustalW, ClustalX
  - MSA
  - T-Coffee

- **Local**
  - BLOCKS
  - eMOTIF
  - GIBBS
  - HMMER
  - MACAW
  - MEME

- **Other**
  - Profile Analysis from msa (UCSD)
  - SAM HMM (from msa)
Multiple Alignments: CLUSTALW

* identical
: conserved substitutions
. semi-conserved substitutions

<table>
<thead>
<tr>
<th>gi</th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>2213819</td>
<td>CDN-ELKSEAIEHLCASEFALR----------MKIKEVKKENGDKK</td>
<td>223</td>
</tr>
<tr>
<td>12656123</td>
<td>----ELKSEAIEHLCASEFALR----------MKIKEVKKENG-</td>
<td>31</td>
</tr>
<tr>
<td>7512442</td>
<td>CKNKNDNDNDMETLCKNFALK-----------IKVKEITYINRDYK</td>
<td>211</td>
</tr>
<tr>
<td>1344282</td>
<td>QDECKFDYVEVYETSSGAFSLLGRFCGAEPPPHLVSSHHHELAVLFRTDH</td>
<td>400</td>
</tr>
</tbody>
</table>

Red: AVFPMLW (Small & hydrophobic)
Blue: DE (Acidic)
Magenta: RHK (Basic)
Green: STYHCNGQ (Hydroxyl, Amine, Basic)
Gray: Others
### Multiple Alignments

- **Family alignment for the ITAM domain (Immunoreceptor tyrosine-based activation motif)**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
<th>Sequence 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD3D_MOUSE/1-2</td>
<td>EQLYQP1RDR</td>
<td>EDTQ-YSRLG</td>
<td>GN</td>
<td></td>
</tr>
<tr>
<td>Q90768/1-21</td>
<td>DQLYQPIGER</td>
<td>NDGQ-YSQLA</td>
<td>TA</td>
<td></td>
</tr>
<tr>
<td>CD3G_SHEEP/1-2</td>
<td>DQLYQP1KER</td>
<td>EDDQ-YSHLR</td>
<td>KK</td>
<td></td>
</tr>
<tr>
<td>P79951/1-21</td>
<td>NDLYQPLGQR</td>
<td>SEDT-YSHLN</td>
<td>SR</td>
<td></td>
</tr>
<tr>
<td>FCEG_CAVPO/1-2</td>
<td>DGITQG1STR</td>
<td>NQET-VETLK</td>
<td>HE</td>
<td></td>
</tr>
<tr>
<td>CD3Z_HUMAN/3-0</td>
<td>DGLYQG1STA</td>
<td>TKDT-YDALH</td>
<td>MQ</td>
<td></td>
</tr>
<tr>
<td>C79A_BOVIN/1-2</td>
<td>ENLYEGLNLD</td>
<td>DCSM-YEDIS</td>
<td>RG</td>
<td></td>
</tr>
<tr>
<td>C79B_MOUSE/1-2</td>
<td>DHTVEGLNID</td>
<td>QTAT-YDIV</td>
<td>TL</td>
<td></td>
</tr>
<tr>
<td>CD3H_MOUSE/1-2</td>
<td>NQLYNE1NLG</td>
<td>RREE-YDVLE</td>
<td>KK</td>
<td></td>
</tr>
<tr>
<td>CD3Z_SHEEP/1-2</td>
<td>NPVYNE1NVG</td>
<td>RREE-YAVLD</td>
<td>RR</td>
<td></td>
</tr>
<tr>
<td>CD3E_HUMAN/1-2</td>
<td>NPDYEPIRK</td>
<td>QRDL-YSGLN</td>
<td>QR</td>
<td></td>
</tr>
<tr>
<td>CD3H_MOUSE/2-0</td>
<td>EGVYNAQKD</td>
<td>KMAEAYSEIG</td>
<td>TK</td>
<td></td>
</tr>
<tr>
<td>Consensus/60%</td>
<td>-.LYpsLspc</td>
<td>pcsp.YspLs</td>
<td>pp</td>
<td></td>
</tr>
</tbody>
</table>

Simple Modular Architecture Research Tool
Multiple Alignment

Motif

```
xxxMxxxxxx
xxxxxMxx
xxxxxMxx
xMxxxxxxx
xxxxxxxxx
Mxxxxxxx
xxxMxxx
xMxxxxxx
xxxxxxM
```

Random start positions chosen
Location of motif in each sequence provides first estimate of motif composition
How to Score Multiple Alignments?

- **Sum of Pairs Score (SP)**
  - Optimal alignment: $O(dN)$ [Dynamic Prog]
  - Approximate Algorithm: Approx Ratio 2
    - Locate Center: $O(d^2N^2)$
    - Locate Consensus: $O(d^2N^2)$

*Consensus char*: char with min distance sum

*Consensus string*: string of consensus char

*Center*: input string with min distance sum
Multiple Alignment Methods

- **Phylogenetic Tree Alignment** *(NP-Complete)*
  - Given tree, task is to label leaves with strings

- **Iterative Method(s)**
  - Build a MST using the distance function

- **Clustering Methods**
  - Hierarchical Clustering
  - K-Means Clustering
Multiple Alignment Methods (Cont’d)

- **Gibbs Sampling Method**

- **Hidden Markov Model**
Multiple Sequence Alignments (MSA)

- **Choice of Scoring Function**
  - Global vs local
  - Gap penalties
  - Substitution matrices
  - Incorporating other information
  - Statistical Significance

- **Computational Issues**
  - Exact/heuristic/approximate algorithms for optimal MSA
  - Progressive/Iterative/DP
  - Iterative: Stochastic/Non-stochastic/Consistency-based

- **Evaluating MSAs**
  - Choice of good test sets or benchmarks (BAliBASE)
  - How to decide thresholds for good/bad alignments
Progressive MSA: CLUSTALW

Figure 1. Limits of the progressive strategy.

This example shows how a progressive alignment strategy can be misled. In the initial alignment of sequences 1 and 2, ClustalW has a choice between aligning CAT with CAT and making an internal gap or making a mismatch between C and F and having a terminal gap. Since terminal gaps are much cheaper than internals, the ClustalW scoring schemes prefers the former. In the next stage, when the extra sequence is added, it turns out that properly aligning the two CATs in the previous stage would have led to a better scoring sum-of-pairs multiple alignment.

C. Notredame, Pharmacogenomics, 3(1), 2002.
Software for MSA

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Website</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>Exact</td>
<td><a href="http://www.ibr.wustl.edu/ibr/msa.html">http://www.ibr.wustl.edu/ibr/msa.html</a></td>
<td>[28]</td>
</tr>
<tr>
<td>OMA</td>
<td>Iterative DCA</td>
<td><a href="http://bibiserv.techfak.uni-bielefeld.de/oma">http://bibiserv.techfak.uni-bielefeld.de/oma</a></td>
<td>[61]</td>
</tr>
<tr>
<td>ComAlign</td>
<td>Consistency-based</td>
<td><a href="http://www.daimi.au.dk/~ocaprani">http://www.daimi.au.dk/~ocaprani</a></td>
<td>[75]</td>
</tr>
<tr>
<td>Praline</td>
<td>Iterative/progressive</td>
<td><a href="mailto:jhering@nimr.mrc.ac.uk">jhering@nimr.mrc.ac.uk</a></td>
<td>[48]</td>
</tr>
<tr>
<td>HMMER</td>
<td>Iterative/Stochastic/HMM</td>
<td><a href="http://hmmer.wustl.edu/">http://hmmer.wustl.edu/</a></td>
<td>[68]</td>
</tr>
<tr>
<td>GA</td>
<td>Iterative/Stochastic/GA</td>
<td><a href="mailto:czhang@watnow.uwaterloo.ca">czhang@watnow.uwaterloo.ca</a></td>
<td>[52]</td>
</tr>
</tbody>
</table>

MSA: Conclusions

- Very important
  - Phylogenetic analyses
  - Identify members of a family
  - Protein structure prediction

- No perfect methods

- Popular
  - Progressive methods: CLUSTALW
  - Recent interesting ones: Prrp, SAGA, DiAlign, T-Coffee

- Review of Methods [C. Notredame, Pharmacogenomics, 3(1), 2002]
  - CLUSTALW works reasonably well, in general
  - DiAlign is better for sequences with long insertions & deletions (indels)
  - T-Coffee is best available method