Quick-Motif: An Efficient and Scalable Framework for Exact Motif Discovery

Yuhong Li, Leong Hou U, Man Lung Yiu, Zhiguo Gong
Department of Computer and Information Science, University of Macau, Macau
Department of Computing, The Hong Kong Polytechnic University, Hong Kong

Motif Discovery: Given a sequence object \( s \) and the targeted motif length \( \ell \), the motif discovery is to return a pair of subsequences \( (s_i, s_j) \), where the normalized Euclidean distance of \( s_i \) and \( s_j \) is minimum among all nontrivial subsequence pairs.

**Problem Definition**

**Time Complexities of Methods**

<table>
<thead>
<tr>
<th>Methods</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brute force</td>
<td>( O(m^2 - \ell) )</td>
</tr>
<tr>
<td>SBF</td>
<td>( O(m^3) )</td>
</tr>
<tr>
<td>MK</td>
<td>( \frac{1}{2}m^2 (1 - p_0) \cdot R + \frac{1}{2}m^2 (1 - p_0) \cdot \ell )</td>
</tr>
<tr>
<td>PAA MBR Construction</td>
<td>( \frac{1}{2}m^2 (1 - p_0) \cdot \phi + \frac{1}{2}m^2 (1 - p_0) - \ell \cdot w )</td>
</tr>
</tbody>
</table>

**Methods**
- Brute force: \( O(m^2 - \ell) \)
- SBF: \( O(m^3) \)
- MK: \( \frac{1}{2}m^2 (1 - p_0) \cdot R + \frac{1}{2}m^2 (1 - p_0) \cdot \ell \)
- PAA MBR Construction: \( \frac{1}{2}m^2 (1 - p_0) \cdot \phi + \frac{1}{2}m^2 (1 - p_0) - \ell \cdot w \)

**Methods**
- Brute force: \( O(m^2 - \ell) \)
- SBF: \( O(m^3) \)
- MK: \( \frac{1}{2}m^2 (1 - p_0) \cdot R + \frac{1}{2}m^2 (1 - p_0) \cdot \ell \)
- PAA MBR Construction: \( \frac{1}{2}m^2 (1 - p_0) \cdot \phi + \frac{1}{2}m^2 (1 - p_0) - \ell \cdot w \)

**Naïve solution:**
- Compute the distance LB for every \( w \)-MBR pair
- Time complexity is \( O(m/w) \cdot \phi \cdot \phi \), \( \phi \) is the PAA dimensionality

**How to efficiently find the surviving \( w \)-MBR pairs?**
- Enable batch pruning \( \rightarrow \) Hierarchical tree structure
- Discovery the true motif as soon as possible to improve the pruning ability \( \rightarrow \) Locality-based search strategy

**Lazy Group Refinement**

**Objective:** lazily refine the surviving \( w \)-MBR pairs by batch to maximize the reusability of running cross-sums.

**Advantages:**
- Prune \( w^2 \) subsequences pairs in a batch
- Enable incremental \( O(1) \) distance computations

**Experiments**

Project page: http://deggroup.cis.umac.mo/quickmotifs/