L-SME: a system for mining loosely structured motifs

Fabio Fassetti[1] Gianluigi Greco[2] Giorgio Terracina[2]

f.fassetti@deis.unical.it;{ggreco, terracina}@mat.unical.it
[1]ICAR-CNR & [2]Dept. of Mathematics., University of Calabria

ECML-PKDD 2011

Introduction

- **Singling out the regions that are over-represented in suitably selected sets of DNA sequences provides us with insights on the biological functions played by the corresponding macromolecules.**
- **These regions are called motifs in the literature.**

Motif Discovery Problem

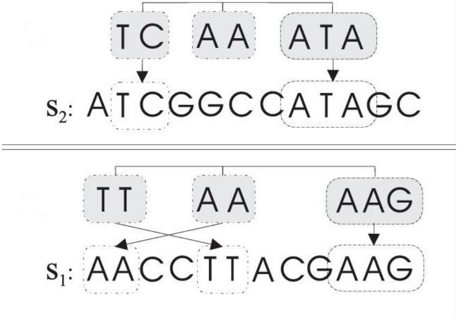
- A motif template $\hat{p}$ is a tuple $\langle l_1, d_1, l_2, d_2, \ldots l_{r-1}, d_{r-1}, l_r \rangle$.
  - l_i: length of the i-th box ($l_i = [\min\_l_i, \max\_l_i]$)
  - d_j: length of the gap between j-th and $(j+1)$-th box ($d_i = [\min\_d_i, \max\_d_i]$)
- A pattern instance p for $\hat{p}$ is a string $b_{l_1} X(d_1) b_{l_2} X(d_2) \ldots b_{l_{r-1}} X(d_{r-1}) b_{l_r}$
  - b_{l_i} is a string with length in the range $[\min\_l_i, \max\_l_i]$
  - $X(d_j)$ is a sequence of “don’t care” symbols with length in the range $[\min\_d_j, \max\_d_j]$
- A pattern instance p occurs in a DNA sequence s if there is a substring s' of s if s' matches p
- The motif discovery problem over a set of DNA sequences is to find all the instances for $\hat{p}$ that occur in at least $\mathbb{Q}$ of them
  - $\mathbb{Q}$ is the quorum considered appropriate by the biologist

Supported-Templates Perspective

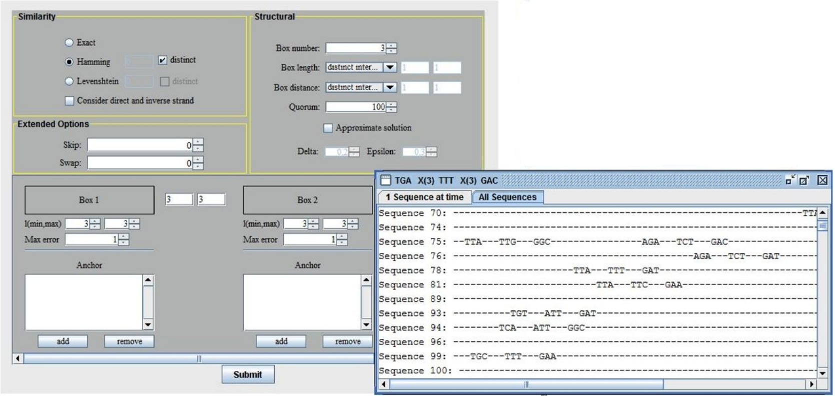
L-SME is a tool for motif discovery supporting various innovative functionalities, under various different perspectives

- L-SME allows the user to specify any kind of model template
- L-SME deals with other relevant variabilities in pattern matching, in particular, it supports
  - both Hamming and Levenshtein distance
  - box skips: a user-definable number of boxes is not matched at all
  - box swaps: a user-definable number of inversions between adjacent boxes

Supported-Templates Perspective



Interfacing Perspective



Similarity

○ Exact
● Hamming ☑ distinct
○ Levenshtein □ distinct
□ Consider direct and inverse strand

Extended Options

Skip: 0
Swap: 0

Structural

Box number: 3
Box length: distinct inter... 1 1
Box distance: distinct inter... 1 1
Quorum: 100
□ Approximate solution
Delta: 0.2 Epsilon: 0.3

Box 1: 3 3
l(min,max) 3 3
Max error 1

Box 2
l(min,max) 3 3
Max error 1

TGA X(3) TTT X(3) GAC

```
Sequence 70: -------------------------------------------------------------TTA
Sequence 74: ----------------------------------------------------------------
Sequence 75: --TTA---TTG---GGC----------------AGA---TCT---GAC----------------
Sequence 76: ------------------------------------------AGA---TCT---GAT-------
Sequence 78: ---------------------TTA---TTT---GAT----------------------------
Sequence 81: -------------------------TTA---TTC---GAA------------------------
Sequence 89: ----------------------------------------------------------------
Sequence 93: ----------TGT---ATT---GAT---------------------------------------
Sequence 94: --------TCA---ATT---GGC-----------------------------------------
Sequence 96: ----------------------------------------------------------------
Sequence 99: ---TGC---TTT---GAA----------------------------------------------
Sequence 100: ---------------------------------------------------------------
```

Computation Perspective

- Issues:
    - Motif discovery is a computationally intensive task
- Solution:
    - L-SME is designed to incrementally produce results
        - Each request is immediately answered with an url
        - There, the results are visualized as soon as they are discovered
        - The results remain available for some days.

Algorithmic Perspective

- Issues:
  - The system needs to handle wide classes of templates
  - The system must guarantee scalability over genome-wide applications
- Solution:
  - The system supports search via randomization with a-priori guaranteed quality
  - The user is allowed to tune two normalized coefficients δ and ϵ for setting time/space requirements

Conclusion

- **System available at:**
  - `http://siloe.deis.unical.it/l-sme/`
- For additional information:
  - `terracina@mat.unical.it`