Revisão de Artigo:
“*Analysis of computational approaches for motif discovery*”

Li, Nan, and Martin Tompa. *Algorithms for molecular biology* 1.1 (2006): 8.

Introdução / Motivação

- Revisão do artigo “**Analysis of computational approaches for motif discovery**” para a cadeira de “Algoritmos para Bioinformática” @FCUP.

- Estudo do **estado da arte de métodos de descoberta de motifs**.

Objetivos

- Categorização de **funções objetivo** e avaliação das mais promissoras.

- **Features** que tornam um dataset de difícil análise.

- Apresentação de um novo algoritmo, com uma **nova função objetivo** através do uso de uma **nova feature**.

Glossário

- Motif
- Fator de transcrição
- Binding site
- Função objetivo
- Datasets
- Feature
- Problema de pesquisa
- Problema de classificação

Métodos

Abstração dos problemas em:

- Problema de **pesquisa** -- pesquisa de binding sites.

- Problema de **classificação** (mais relevante) -- classificar palavras da sequência em binding site ou não.

Métodos - Funções Objetivo

Log likelihood ratio:

- Não é capaz de capturar eficazmente a natureza dos binding sites.
- Abordando como classificação:
  - não é possível arranjar modelo capaz de distinguir verdadeiros motifs de ruído de fundo.
- **Pior performance** das 3 a serem apresentadas.

Métodos - Funções Objetivo

Z-score

- Baseia-se num **motif** modelo obtido usando **consensus.**
- Não é capaz de incorporar os verdadeiros binding sites.
- Abstração dos fundamentos biológicos.

Sequence specify

- Enfatiza todas as sequências serem ligadas com o fator de transcrição.
- **Melhor performance** das 3 que são analisadas

Métodos - Features do Dataset

Features com maior correlação com a performance do dataset:

- conservação da sequência, ou seja, entropia relativa do alinhamento do binding site (**maior correlação**)
- a posição da distribuição dos binding sites nas sequências (não se distribuem uniformemente)
- o tamanho da sequência presente no dataset

Resultados

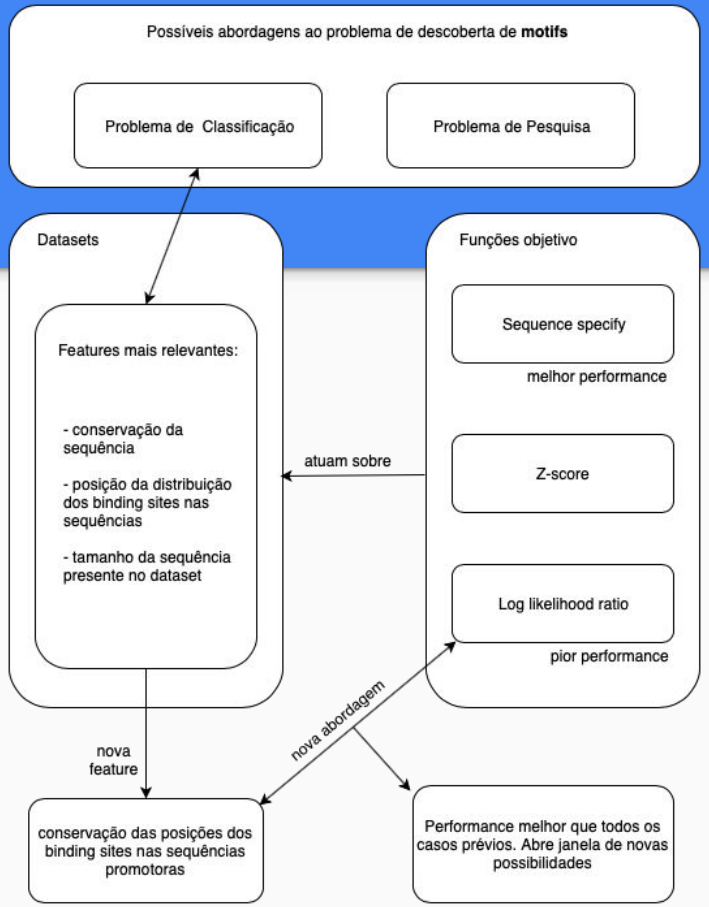
- **Sequence specify** apresenta a melhor performance.

- Introdução de nova feature, **conservação das posições dos binding sites nas sequências promotora**
- regressão linear múltipla
- função objetivo base '**log likelihood ratio**'

nova função objetivo com **performance muito superior** aos métodos estudados.

Conclusão

- Com nova feature e nova abordagem podemos obter performances superiores.
- Tamanho da sequência é mais impactante na performance do modelo do que o tamanho do dataset
- Binding sites não se distribuem uniformemente.



Discussão

- Qualquer questão ou dúvida que queiram esclarecer estejam à vontade para questionar.