Projeto 3 de Algoritmos para Bioinformática: Revisão de Artigo

Edgar Carneiro
Faculdade de Ciências da Universidade do Porto
(Dated: 12 de Maio de 2019)

I. INTRODUÇÃO

No âmbito da disciplina de Bioinformática, disciplina esta que aborda temáticas atuais e relevantes na área da investigação, será feita a revisão do artigo entitulado ***"Analysis of computational approaches for motif discovery"*** [1].

Este artigo foi publicado em 2006 por *Li, Nan* e *Martin Tompa* no jornal *Algorithms for molecular biology 1.1.*

Neste artigo, é realizado um estudo do **estado da arte de métodos de descoberta de *motifs***.

II. OBJETIVOS

O artigo em questão tem como objetivo realizar a categorização das funções objetivo dos métodos em análise, bem como uma avaliação de quais as funções objetivo mais promissoras a nível de possíveis posteriores otimizações.

Tem como objetivo também identificar quais as *features* que tornam um *dataset* de difícil análise para os métodos em questão.

Em jeito de conclusão, tem como objetivo desenhar um novo algoritmo para descoberta de *motifs*, apresentando uma nova função objetivo que incorpore uma importante *feature* que até ao momento não teria sido utilizada pelas ferramentas existentes.

III. GLOSSÁRIO

Primeiramente, torna-se fulcral clarificar o significado de diversos termos técnicos, pertinentes à area da Bioinformática, usados com frequência ao longo do artigo:

- ***motif***: Um padrão sequencial de nucleotídeos ou aminoácidos que tem, ou é conjeturado que tenha, significado biológico [2].
- **fator de transcrição**: Proteína que controla a frequência de transcrição de informação genética, de DNA para RNA.
- ***binding site***: Região de uma macro-molécula, como uma proteína, que se liga a outra molécula específica. [4]
- **função objetivo**: A função que representa o objetivo do problema em questão e que se pretende maximizar, de forma a obter os melhores resultados.
- **datasets**: Conjuntos de dados com a informação referente ao problema.
- ***feature***: Característica presente em todas as entradas do conjunto de dados, com informação mensurável relativamente ao dataset [3].
- **problema de pesquisa**: problema que tem como objetivo encontrar uma determinada sequência, elemento ou padrão num dado conjunto de dados.
- **problema de classificação**: problema que tem como objetivo agrupar data de um dataset segundo um critério em particular [5].

IV. MÉTODOS

O problema de descoberta de *motifs* é frequentemente abstraído como um **problema de pesquisa**, em que pesquisamos por *binding sites* candidatos, nas sequências do dataset, que otimizem a função objetivo. Uma função objetivo ideal deve ser capaz de avaliar com *score* ótimo aos verdadeiros *binding sites* do *motif* e a mais lado nenhum.

O mesmo problema pode também ser modelado como **problema de classificação**: classificar todas as palavras, de um determinado tamanho, de uma sequência como sendo *binding sites* ou não.

A. Funções Objetivo

Caso a previsão feita seja inferior ao verdadeiro valor - ***under-representation*** -, então a diferença entre os dois valores mostra qual a extensão para a qual a função objetivo é incapaz de detectar os valores ótimos.

Caso a previsão seja superior ao verdadeiro valor - ***over-representation*** -, então a função objetivo não é precisa o suficiente para modelar os verdadeiros *binding sites*.

De seguida, são avaliadas três funções objetivo popularmente utilizadas:

1. ***Log likelihood ratio***: Para a maioria dos datasets ocorre *over-representation*. Conclui-se que que até um algoritmo que garanta a solução global ótima para a função objetivo irá falhar o verdadeiro *binding site* pois esta não é capaz de capturar eficazmente a natureza dos *binding sites*. Abordando este

problema como um problema de classificação, não foi possível encontrar um balanço entre a sensibilidade e a especificação da classificação que tornasse o *Log likelihood ratio* capaz de distinguir os verdadeiros *motifs* do ruído de fundo. Ferramentas que usam: *MEME* [6].

2. ***Z-score***: Mede a significância das previsões baseando-se nas suas *over-representations*. Define um *motif* modelo obtido usando *consensus* e encontra os candidatos com maior *Z-score*, sendo no entanto por vezes criticado por não ser capaz de incorporar os verdadeiros *binding sites* no modelo criado e de se abstrair dos fundamentos biológicos. Ferramentas que usam: *YMF*.

3. ***Sequence specify***: Função objetivo que enfatiza o facto de todas as sequências serem ligadas com fator de transcrição. Previsões que consideram um número equilibrado de *binding sites* são as mais significantes. Nos datasets usados esta função objetivo foi a que obteve a melhor performance. Ferramentas que usam: *ANN-spec* e *Weeder*.

Nenhuma função objetivo satisfaz o padrão de que todos os *motifs* têm de ter um *score* pelo menos tão alto como os das previsões.

B. *Features* do Dataset

Quando o problema é abordado como um problema de classificação, e é utilizada regressão linear múltipla, são encontradas oito *features* correlacionadas com a performance do dataset. Naturalmente, os resultados provam também que as *features* não são independentes.

O subconjunto de *features* que apresenta maior correlação com a performance do dataset é:

- conservação da sequência, ou seja, entropia relativa do alinhamento do *binding site*;
- a posição da distribuição dos *binding sites* nas sequências;
- o tamanho da sequência presente no dataset.

Apesar de estas serem as que maior impacto têm na performance do modelo, as restantes *features* não podem ser descartadas. A conservação da sequência revela-se, das três, a mais impactante.

V. RESULTADOS

Das três alternativas de função objetivo previamente apresentadas ***sequence specify*** é a que apresenta melhor performance nos datasets utilizados.

No entanto, a introdução de uma nova *feature* - conservação das posições dos *binding sites* nas sequências promotoras - e a formulação de uma nova função objetivo, permite obter resultados em muito superiores aos resultados previamente obtidos, usando regressão linear múltipla e como base a função objetivo ***log likelihood ratio***.

VI. REPRESENTAÇÃO ESQUEMÁTICA DO ARTIGO

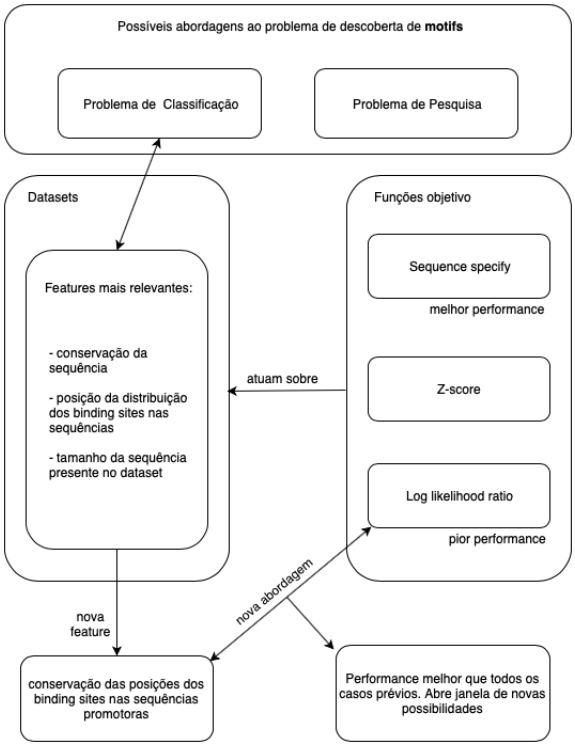

Figura 1. Esquema gráfico das temáticas apresentadas no artigo [1]

VII. CONCLUSÕES

Das três funções objetivo previamente apresentadas a que obtém melhores resultados é a *sequence specify* e pior é a *log likelihood ratio*.

Da análise dos datasets é possível concluir que o tamanho da sequência é mais impactante na performance do modelo do que o tamanho do dataset e que os *binding sites* não se distribuem uniformemente.

No entanto, usando o método que pior performance tinha obtido com uma nova abordagem e com relevância numa nova *feature*, gera-se uma nova função objetivo que obtém resultados superiores, abrindo assim toda uma gama de novas possibilidades promissoras, através da utilização de outras ferramentas com este método.

VIII. BIBLIOGRAFIA

[1] Li, Nan, and Martin Tompa. "Analysis of computational approaches for motif discovery."Algorithms for molecular biology 1.1 (2006): 8.

[2] CTI Reviews. "Principles of Virology, Molecular Biology: Biology, Biology."Cram101 Textbook Reviews, 2016

[3] feature & feature engineering, http://www.datascienceglossary.org

[4] Alberts B, Johnson A, Lewis J, et al. "Molecular Biology of the Cell. 4th edition."New York: Garland Science, 2002

[5] Classification problem, https://brilliant.org/wiki/classification

[6] The MEME Suite, http://meme-suite.org