CARACTERIZAÇÃO DE SEQUÊNCIAS PROMOTORAS RELACIONADAS AO SIGMA 28 PARA APRIMORAMENTO DA FERRAMENTA BACPP

BIOINFO - UCS

AUTORES: GABRIEL DALL ALBA & DR.ª SCHEILA DE AVILA E SILVA (ORIENTADORA)

MODALIDADE DA BOLSA: BIC/UCS

UCS UNIVERSIDADE DE CAXIAS DO SUL

1. INTRODUÇÃO

Sequências promotoras (ou promotores) são elementos regulatórios fundamentais no processo de transcrição gênica, a qual ocorre através de uma interação entre a enzima RNA polimerase e uma sequência promotora. Estes elementos possuem propriedades estruturais (i.e. como estabilidade e curvatura) que podem servir como característica discriminante entre regiões promotoras e regiões codificantes. Graças à sua aplicabilidade no reconhecimento de padrões, técnicas de aprendizado de máquina (como Redes Neurais Artificiais (ANNs) tem sido constantemente empregadas na predição de promotores. Apesar disso, as características intrínsecas dos promotores ainda apresentam-se como um empecilho para ferramentas computacionais.

2. OBJETIVO

O presente trabalho tem como objetivo estabelecer o perfil que melhor caracteriza sequências promotoras procariotas relacionadas ao Sigma (σ) 28 utilizando a estabilidade da dupla-fita de DNA como parâmetro.

3. METODOLOGIA

Para a realização deste trabalho, 144 sequências (exemplos verdadeiros) relacionadas ao fator σ28 foram extraídas do banco de dados RegulonDB. Adicionalmente, 87 sequências (exemplos falsos) foram geradas via algoritmos em linguagem Python nas probabilidades de: 0.22 para adenina (A) e timina (T), 0.28 para citosina (C) e guanina (G), de acordo com metodologia descrita na literatura. Todas as sequências foram extraídas/geradas com tamanho padrão de 80 nucleotídeos (nt). Os conjuntos de dados foram codificados em valores de estabilidade (Vs) conforme metodologia descrita por Kanhere e Bansal (2005) e scripts em linguagem Python (através da técnica de janela deslizante (moving window), passando por uma janela de um nucleotídeo por vez).

Em seguida, foram obtidos os valores de protótipo de regra do σ28 conforme de Avila e Silva et al. (2014). Um protótipo de regra, de acordo com os autores, são resultantes da extração das regras matemáticas - seguindo uma metodologia específica - do aprendizado de uma ANN. Cada posição de nt em uma sequência de tamanho padrão (80 nt) recebem um valor de acordo com o protótipo de regra (Vr). Foi, então, realizado um cálculo de diferença (Vdif) entre os valores de Vr e Vs de cada sequência (exemplos verdadeiros e falsos), considerando um nt por vez. Para este cálculo, apenas os Componentes Principais das sequências foram considerados de acordo com metodologia descrita em de Avila e Silva et al. (2014). Os autores realizaram uma Análise Multivariada de Componentes Principais em sequências relacionadas aos fatores σ (σ24, σ28, σ32, σ38 e σ54 e σ70), identificando as regiões das sequências que melhor explicam as diferenças discriminatórias entre fatores σ. Essa informação foi selecionada como critério para a abordagem descrita neste trabalho.

APOIO: UCS UNIVERSIDADE DE CAXIAS DO SUL

Com o Vdif para cada sequência, realizou-se a soma de valores para que cada sequência receba um único valor. Por fim, estes valores foram plotados em um histograma, buscando observar a existência de um ponto de corte entre sequências verdadeiras e falsas. A observação de uma diferença entre os valores matemáticos dos exemplos positivos e negativos pode ser estudado e aplicado na ferramenta BacPP, buscando reduzir o número de falsos positivos da ferramenta.

4. RESULTADOS E DISCUSSÃO

Através do histograma gerado com os valores referentes às sequências verdadeiras e às falsas, foi possível observar um possível ponto de corte no intervalo de valores entre -1 a 1 (figura 2).

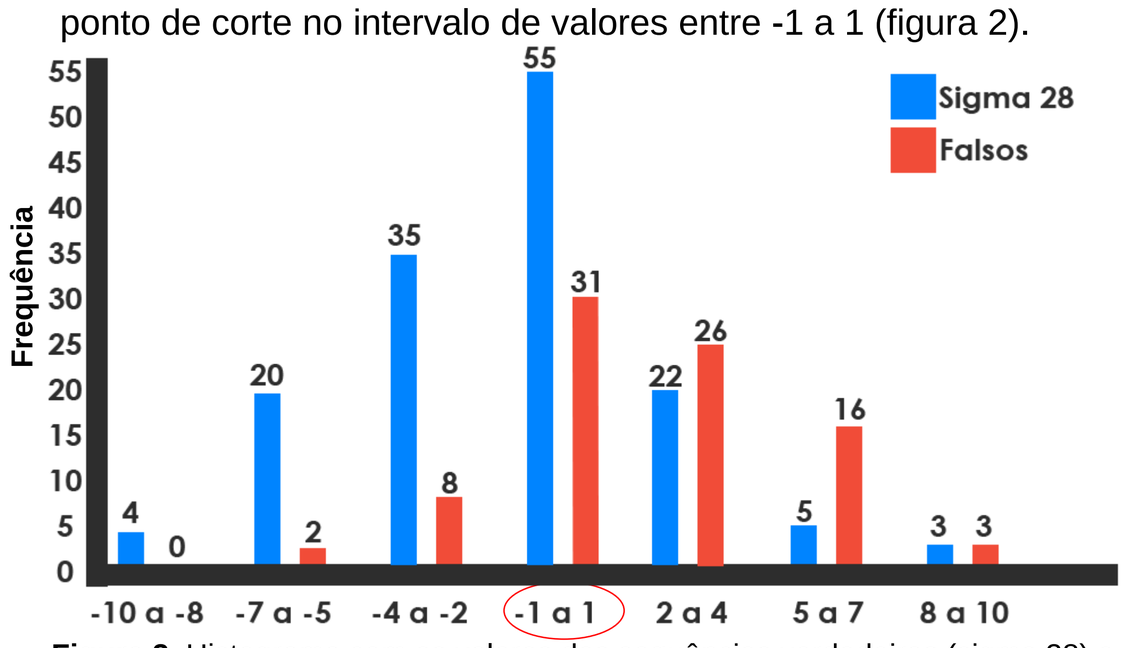

Figura 2. Histograma com os valores das sequências verdadeiras (sigma 28) e falsas.

Conforme a tabela 1, explica-se a razão de atribuir o ponto de corte ao intervalo -1 a +1: 41% das sequências verdadeiras encontram-se em valores inferiores ao intervalo, enquanto que 52% das sequências falsas encontram-se em valores superiores. É possível que, buscando intervalos menores entre -1 e +1, as sequências verdadeiras e falsas sejam melhor separadas, aumentando a eficácia da metodologia e, posteriormente, da ferramenta BacPP.

Tabela 1. Porcentagem de sequências nos intervalos menores e maiores que o ponto de corte.

Qtd. de sequências nos intervalos:	Sigma 28	Falsas
-10 a -1	40,97%	11,62%
-1 a +1	38,19%	36,04%
+1 a +10	20,83%	52,32%

5. CONSIDERAÇÕES FINAIS

Neste trabalho foi demonstrado que: (i) considerando os componentes principais das sequências promotoras e (ii) a diferença entre a regra de aprendizado da ferramenta BacPP e valores de estabilidade das sequências é possível determinar um ponto de corte entre sequências verdadeiras e falsas. Pretende-se refinar e expandir esta análise para os demais fatores sigma. Por fim, pretende-se aprimorar a ferramenta BacPP, bem como auxiliar na resolução de problemáticas relacionadas à predição de promotores bacterianos.

6. REFERÊNCIAS

DE AVILA E SILVA, S.; GERHARDT, G. J. L. & ECHEVERRIGARAY, S. Rules extraction from neural networks applied to the prediction and recognition of prokaryotic promoters. **Genetics and Molecular Biology**, 34(2), 2011. 353-360 p.

DE AVILA E SILVA, S et al. (2014). DNA duplex stability as discriminative characteristic for Escherichia coli s54- and s28- dependent promoter sequences. **Biologicals**, 42, 22-18.

KANHERE, A., BANSAL, M. A. (2005b). Novel method for prokaryotic promoter prediction based on DNA stability. **BMC Bioinformatics** 6(1).