



## Introdução

O estresse desencadeado pelo choque térmico em uma bactéria provoca rápidas mudanças em alguns processos biológicos, por exemplo, o envelopamento e a fixação de proteínas. A resposta biológica a este estímulo é regulada pela ligação dos fatores sigma ( $\sigma$ ) 24 e 32 na enzima RNA polimerase (RNAP) e o posterior reconhecimento da região promotora. A identificação dos pares de base constituintes do promotor auxiliam o estudo da regulação gênica. Deste modo, amplia-se a compreensão da capacidade e dos mecanismos de sobrevivência utilizados pelos organismos.

## Objetivo

Com base nestes aspectos, o objetivo deste trabalho foi analisar a composição de nucleotídeos das regiões promotoras reconhecidas pelos fatores  $\sigma$ 24 e  $\sigma$ 32 de *Escherichia coli*.

## Metodologia

### Etapa 1



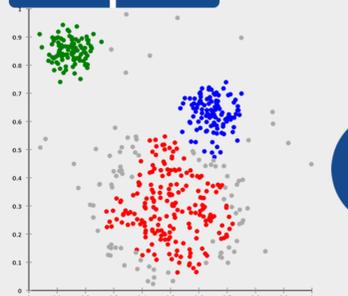
Coleta de dados  
(Banco de Dados)



Script em  
linguagem Python  
para transformação  
e formatação de  
dados

Dados  
transformados e  
formatados

### Etapa 2



Algoritmo de clusterização  
K-means

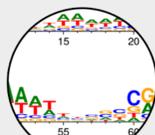
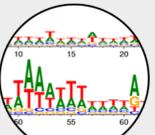
### Etapa 3



Análise de  
pureza  
dos *clusters*



Análise de composição  
de nucleotídeo:  
WebLogo



Análise na  
literatura



Transformação  
dos resultados

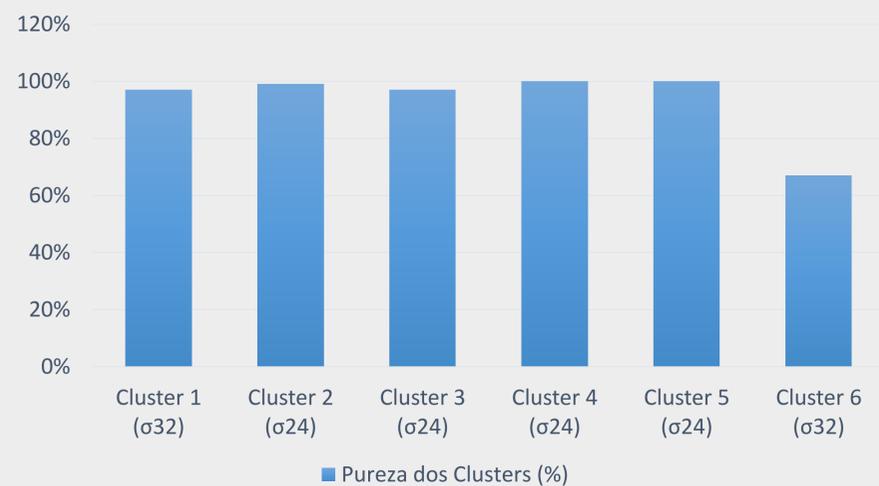
### Resultados

Cluster 1—Cluster 4  
Cluster 2—Cluster 5  
Cluster 3—Cluster 6

## Resultados e Discussão

Os agrupamentos foram analisados em relação à pureza do cluster, ou seja, quantas sequências de um mesmo  $\sigma$  foram agrupados em um mesmo cluster. Como a metodologia formal para a obtenção do número ótimo de k é empírica, foram realizadas simulações com diferentes valores de k: 3, 6 e 8. Com a análise dos resultados obtidos, verificou-se que a simulação com k=6 apresentou agrupamentos com uma pureza média de 93%

Pureza dos Clusters (%)



Assim, foi possível observar que a região canônica denominada -10 apresentou conteúdo similar ao consenso biológico previamente descrito em todos os clusters, com exceção de um agrupamento (cluster 5). Já para a região -35, não foi possível estabelecer relação com o consenso estabelecido na literatura. Estas observações são fundamentadas pela presença da região -10 estendida para os promotores do  $\sigma$ 32, ou seja, a informação região -35 foi realocada nesta região, a fim de proporcionar o funcionamento da RNAP. Outra característica evidente foi a prevalência dos nucleotídeos A e G na região +1 (início da transcrição). Conforme relatos prévios, a presença do nucleotídeo G na posição +1 é responsável pela pausa na transcrição, a qual trata-se de um importante mecanismo para reparo de erros de mutação ao longo da sequência da região codificante.

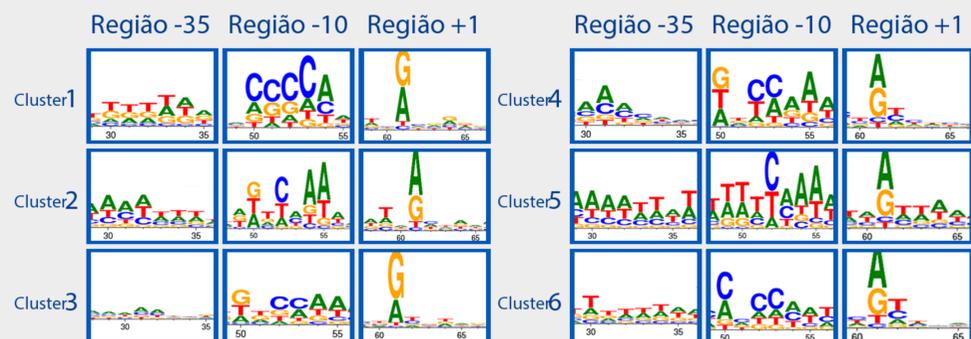


Figura 2. Frequência de nucleotídeos na ferramenta WebLogo para os respectivos *clusters*, evidenciadas as regiões-35, -10 e +1.

## Considerações Finais

Percebe-se que um certo grau de degeneração das regiões consenso é tolerado no processo de transcrição gênica, porém evidenciam o desafio computacional. Assim, estabelecer os distintos perfis encontrados auxilia na redução do número de falsos positivos em ferramentas *in silico* relacionados à predição de promotores

## Referências

KREBS, J.; GOLDSTEIN, S.; KILPATRICK, S. T. **Genes XI**. 11. ed. Sudbury, Massachusetts: Jones and Bartlett, 2014. 930 p.  
LIM, Bentley et al. Heat Shock Transcription Factor  $\sigma$ 32 Co-opts the Signal Recognition Particle to Regulate Protein Homeostasis in *E. coli*. **Plos Biology**. Public Library of Science (PLoS). v 11, n 12, p 1-15, 2013.

## Apoio