Técnica de Clusterização não-hierárquica aplicada para a caracterização *in silico* de promotores associados a genes de choque térmico de *Escherichia coli*

Gabriel Dall'Alba[1], Scheila de Avila e Silva[1]

[1] Instituto de Biotecnologia – Universidade de Caxias do Sul (UCS)
Rua Francisco Getúlio Vargas, 1130– 95070-560 – Caxias do Sul – RS – Brazil

{gdalba@ucs.br sasilva6@ucs.br}

Abstract. *Computational techniques play an important role in the post-genomic era, due to the amount of biological data generated and released. In this context, this paper carry out an in silico analysis of σ24 and σ32-dependent promoter sequences. It was made by using the clustering technique with stability values as input data. The content of the clusters was analyzed by average purity obtained for the clusters. In general, all the clusters presented 63% of purity. However, the clusters 7, 10 and 11 obtained, respectively, 85%, 82% and 83% of purity. As conclusions, was possible to identify degrees of degeneration in the sequences and grouping features. After all, this paper contributes to the comprehension of the different biological bacterial promoter's profiles. Furthermore, these results can be applied in the reducing of false positives of in silico promoter predicting tools.*

Resumo. *As técnicas computacionais têm importante papel na era pós-genômica, devido ao crescente número de dados biológicos gerados. Um campo bem desenvolvido envolve as questões relacionadas a regulação gênica. Com base no exposto, este trabalho dedica-se a uma análise in silico dos promotores relacionados aos fatores σ24 e σ32 da bactéria E. coli. Para isso, foi utilizada a técnica de Clusterização e a codificação dos dados nos valores de estabilidade. Os agrupamentos foram analisados com base na pureza obtida. A pureza média obtida foi de 63%, com destaque para os agrupamentos 7, 10 e 11, que obtiveram uma pureza média de 85%, 82% e 83% respectivamente. Foi possível observar diferentes graus de degeneração das sequências e as características que os grupos estudados apresentam. Deste modo, contribui-se para a compreensão dos diferentes perfis biológicos encontrados nos promotores bacterianos. Além disso, os resultados podem auxiliar na redução de falsos positivos em ferramentas de predição de promotores.*

1. Introdução

A era pós-genômica apresenta uma série de desafios para as diferentes áreas de pesquisa, uma vez que a quantidade de dados biológicos chega, atualmente, aos Petabytes [Marx 2013]. Portanto, há a necessidade de aplicar diferentes técnicas que auxiliem no armazenamento e acesso destas quantidades de dados, procurando reduzir o espaço entre a geração de dados e sua análise [Kanehisa et al. 2014]. Neste contexto, a tecnologia computacional permite o manejo de dados e geração de inferências nos diversos segmentos biológicos, pois permite a análise de grandes quantidades de dados, possibilitando o cruzamento de informações e extração de informações que não seriam

possíveis sem o auxílio destas técnicas [Attwood et al. 2011]. Dentro deste contexto, técnicas de aprendizado de máquina, como SVM (*support vector machine*), Redes Neurais Artificiais e Clusterização são aplicadas em diferentes áreas das ciências da vida, como a genômica, proteômica, metabolômica e regulação gênica [de Avila e Silva e Echeverrigaray 2012] [Callebaut 2012].

A regulação da transcrição gênica em seres procariotos desempenha um papel importante na resposta adequada destes seres às mudanças em seu ambiente, possibilitando sua sobrevivência em determinadas condições. A presença de mecanismos de proteção do genoma e formas de seletividade de expressão gênica (a qual ocorre, principalmente, no momento da transcrição) são reflexos da ausência de núcleo [Krebs, Goldstein e Kilpatrick 2014].

O processo de expressão de genes consiste em várias etapas, iniciadas com a interação da enzima chamada RNA polimerase (RNAp) com segmentos de DNA chamados de promotores, os quais antecedem a região codificante. Esta interação possui um valor essencial, uma vez que o reconhecimento dos promotores é fundamental para a transcrição do gene associado a ele. Deste modo, o promotor pode ser considerado um elemento regulatório da expressão gênica [Krebs, Goldstein e Kilpatrick 2014]. No processo de identificação de um promotor, a RNAp possui uma subunidade chamada sigma (σ), a qual auxiliará na identificação de um promotor específico. Por exemplo, em *Escherichia coli*, os fatores σ24 e σ32 são responsáveis pela expressão de genes relacionados ao estresse por choque térmico, já o σ58 possui função relacionada à assimilação de nitrogênio.

O reconhecimento de um determinado promotor se dá em regiões específicas da sequência, denominadas motivos consensuais. Estes motivos estão localizados em duas regiões distintas do promotor, chamadas de região -10 e região -35, em referência ao primeiro nucleotídeo transcrito, que recebe a numeração +1. Ambas as regiões possuem função distinta, a região -35 atua como sinal para o seu reconhecimento pela RNAp, enquanto a região -10 situa-se na região que ocorre a abertura da fita de DNA [Krebs, Goldstein e Kilpatrick 2014] [De Avila e Silva e Echeverrigaray 2012]. Apesar de levarem o nome de "sequências consenso", o grau de conservação varia tanto entre promotores de um fator σ como entre sequências reconhecidas por diferentes fatores σ. Deste modo, evidencia-se a dificuldade em executar uma análise global dos dados a partir deste critério. As divergências observáveis entre os motivos consensuais justificam seu reconhecimento por diferentes fatores σ, considerando as peculiaridades de cada grupo de sequências. Na tabela 1, é possível visualizar a composição de nucleotídeos dos motivos consensuais para cada fator σ em *Escherichia coli*.

Tabela 1. Composição de nucleotídeos dos motivos consensuais para os diferentes fatores σ em *Escherichia coli*. Onde "pb" representa a distância em pares de base.

Fator σ	*Função*	*Consenso -35 / -10*
24	Estresse por choque térmico	GGAACTT 15 pb GTCTAA
28	Mobilidade celular e patogenicidade	CTAAA 15 pb GCCGATAA
32	Estresse por choque térmico	CCCTTGAA 13-15 pb CCCGATNT
38	Resposta a estresse	TTGACA 16-18 pb TATACT

54	Assimilação de Nitrogênio	CTGGNA 16-18 pb TTGCA
70	Sigma constitutivo	TTGACA 16-18 pb TATAAT

Fonte: DE AVILA E SILVA E ECHEVERRIGARAY, 2012.

Além do conteúdo de nucleotídeos, os promotores também possuem características estruturais próprias, como a estabilidade e a curvatura, oferecendo a possibilidade de diferentes abordagens para o estudo dos promotores [Kanhere e Bansal 2005]. Trabalhos previamente publicados já demonstraram que as regiões promotoras são menos estáveis que as regiões gênicas [Kanhere e Bansal 2005] [Ramprakash e Schwarz 2007] [Jáuregui *et al.* 2003]. Deste modo, justifica-se o potencial do uso desta característica como critério de classificação das sequências promotoras. Uma metodologia desenvolvida em Kanhere e Bansal (2005) e aprimorada por Rangannan e Bansal (2007) mostra, em seus resultados, que a estabilidade apresenta-se como uma medida mais eficaz que os motivos conservados para diferenciar regiões promotoras e não-promotoras [Rangannan e Bansal 2007]. A metodologia é baseada nas diferenças de estabilidade (ΔG^o) entre as regiões promotoras e codificantes.

Com base no que foi apresentado, este trabalho dedica-se a análise *in silico* (computacional) dos promotores reconhecidos pelos fatores σ24 e σ32 em *Escherichia coli*, aplicando a técnica de aprendizado de máquina denominada clusterização e utilizando o critério de estabilidade para a codificação dos dados. A escolha destes dois fatores σ deve-se ao fato de ambos estarem descritos com a mesma função (estresse por choque térmico). O estresse por choque térmico pode alterar alguns processos biológicos da bactéria como, por exemplo, o envelopamento, fixação de proteínas e alterações na forma e estrutura do DNA [Lim et al. 2013].

2. Metodologia

Os dois conjuntos de sequências promotoras da bactéria *E. coli* foram obtidos no banco de dados RegulonDB [Salgado et al. 2013]: (*i*) 521 sequências reconhecidas pelo σ24; (*ii*) 324 sequências reconhecidas pelo σ32. A estabilidade do DNA pode ser expressada através de sua energia livre (ΔG), a qual depende da composição de mononucleotídeos e dinucleotídeos. Portanto, a estabilidade do DNA pode ser predita a partir de sua sequência e das interações com cada vizinho mais próximo. A contribuição de cada dinucleotideo está descrito por SantaLucia e Hicks (2004). Para aplicação da técnica utilizando a energia livre (ΔG), ΔG foi calculada utilizando a seguinte equação, conforme SantaLucia e Hicks (2004) e Kanhere e Bansal (2005).

$$\Delta G^0 = \Delta G_{ij} \quad (1)$$

Onde ΔG^0_{ij} é a variação padrão de energia livre para os dinucleotídeos de tipo *ij*. A equação original, descrita por Kanhere e Bansal (2005), foi adequada para os objetivos propostos neste trabalho. Os valores foram obtidos por meio da técnica de janela deslizante (move window), passando por uma janela de um nucleotídeo por vez. Tanto a preparação dos dados, quanto a ferramenta de execução do algoritmo K-means foram implementadas pelos autores em linguagem de programação python e C#.

Após esta etapa, as sequências estão adequadas para a aplicação da técnica computacional de clusterização, uma técnica de classificação não supervisionada que pode ser dividida em dois métodos: o hierárquico e o não-hierárquico. O algoritmo *K-Means* encontra-se dentro da categoria não-hierárquica, sendo capaz de separar um conjunto de dados em agrupamentos de acordo com um critério de distância predefinido, sendo o critério de cálculo de distância Euclidiano o mais usado. O algoritmo requer um valor numérico *K* (atribuído pelo usuário da ferramenta) equivalente ao número de agrupamentos desejados.

A técnica de clusterização apresenta-se eficaz para o reconhecimento de padrões escondidos dentro de um conjunto de dados. Apesar de ser um método simples e efetivo, cada cluster possui uma sensibilidade apurada quanto ao seu centroide inicial, ou seja, arranjos totalmente diferentes podem surgir a partir de uma nova escolha randômica do centroide inicial [Witten e Frank 2005].

Devido à metodologia para obtenção do número ótimo de K ser empírica, as simulações foram realizadas utilizando valores de K = 8 até 20, conforme trabalhos prévios realizados pelos autores, os quais estão em processo de publicação. Os agrupamentos foram analisados em relação à sua pureza, ou seja, quantas sequências de um mesmo fator σ foram agrupados em um mesmo cluster. Posterior a esta etapa, foi utilizada a ferramenta WebLogo [Crooks 2004] para visualização dos motivos consensuais encontrados em cada agrupamento, a qual está disponível on-line.

3. Resultados e Discussão

3.1 Agrupamentos resultantes

Com a análise dos resultados obtidos, foi possível verificar que a simulação com K = 12 obteve uma pureza média de 63%. A pureza de cada agrupamento gerado, pode ser observada na figura 1.

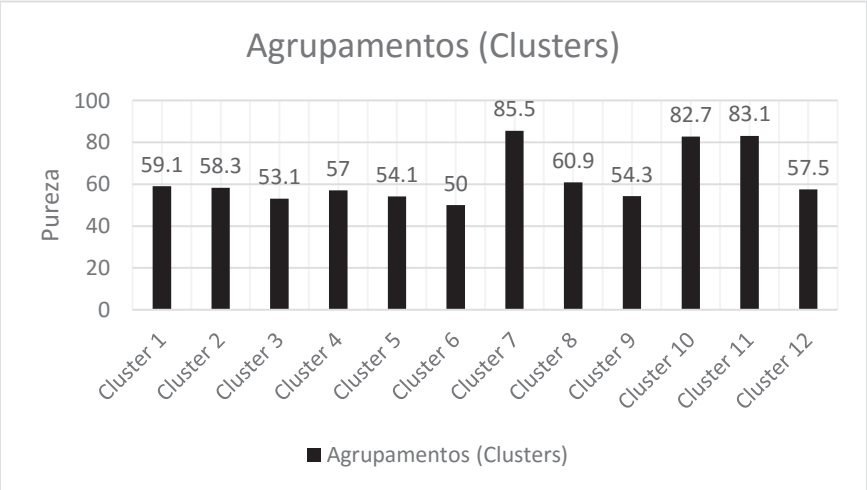

Figura 1. Pureza dos agrupamentos gerados na simulação com valor de K = 12.

Com base na pureza obtida, uma investigação mais aprofundada foi realizada somente para os clusters que obtiveram uma pureza média acima de 80%, ou seja, os clusters 7, 10 e 11, os quais tiveram predominância do σ24. Com o auxílio da ferramenta WebLogo [Crooks 2004], foi possível visualizar os seguintes motivos

consensuais para cada um dos clusters, conforme a figura 2.

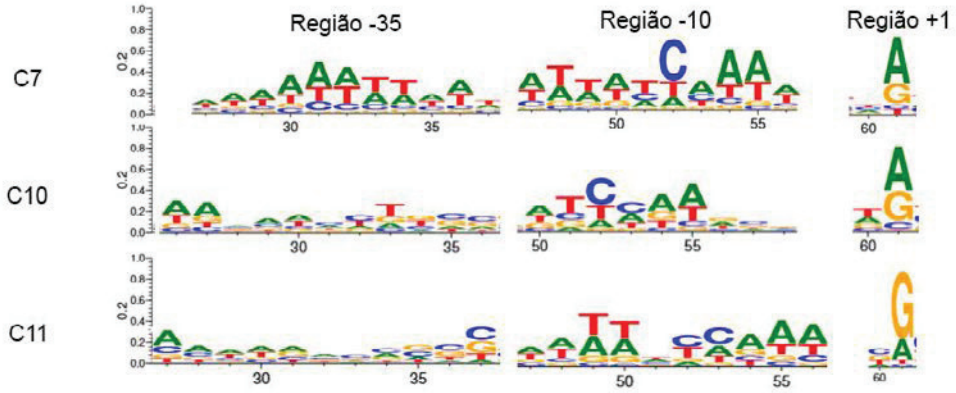

Figura 2. Composição de nucleotídeos para as 2 regiões consensuais, -35 e -10, e para a região +1 encontradas nos clusters com pureza acima de 80%.

É possível observar que a região canônica denominada -10 apresentou similaridades ao consenso biológico previamente descrito na literatura conforme tabela 1 (De Avila e Silva e Echeverrigaray, 2012). Por outro lado, não foi possível verificar relação entre o consenso da região -35 estabelecido na literatura e o consenso obtido. Essas observações são fundamentadas pela presença da região -10 expandida, principalmente para os promotores do σ32, ou seja, a informação da região -35 foi realocada para a posição -10, a fim de proporcionar um melhor funcionamento da RNAp [Lim et al. 2013] [Koo et al. 2009].

Nos 3 clusters analisados, há a prevalência de nucleotídeos A e G localizados na região +1 (início da transcrição). Conforme relatos prévios na literatura, a presença do nucleotídeo G nesta posição pode estar relacionada com um mecanismo de pausa e varredura do gene a ser transcrito, com o objetivo de reparar erros e, consequentemente, diminuir a incidência de mutações. Os mecanismos de resposta ao estresse por choque térmico são importantes para a sobrevivência de um organismo, apresentando-se de forma rápida e eficaz para reduzir as chances de possíveis danos ao genoma [Herbert et al. 2006] [Houten e Kisker 2014] [Vvedenskaya et al. 2014].

Dentro do contexto biológico, percebe-se que um certo grau de degeneração das regiões consenso é tolerado no processo de transcrição gênica, ou seja, a transcrição ocorre de forma apropriada apesar das alterações na composição de nucleotídeos (Figura 1). Porém, computacionalmente, os desafios para a predição de promotores bacterianos são evidenciados. A pureza média obtida para a simulação com k = 12 foi de 63%, o que reforça a dificuldade em separar os dois fatores σ estudados.

Uma das possíveis explicações para a dificuldade apresentada encontra-se nos motivos consensuais do fator σ32. A maior incidência da região -10 extendida é relatada para este fator. Utilizando a ferramenta Weblogo [Crooks 2004] para os agrupamentos que resultaram em uma pureza abaixo de 80%, foi possível observar que, em sua grande maioria, as sequências relacionadas ao fator σ32 possuem o motivo consensual da região -10 degenerado ou similar ao consenso -10 relacionado ao fator σ24. Portanto, a falta de especificidade nesta região para o σ32 pode explicar a dificuldade em separar computacionalmente os dois fatores

3.2 Trabalhos Relacionados

Os trabalhos relacionados à análise *in silico* de promotores bacterianos possuem perfis distintos, com o uso de abordagens diferentes. Kaushik et al. (2016) explica as distintas características estruturais do DNA, como a formatação cruciforme do DNA, a curvatura, cadeias paralelas de DNA e quádruplos de Guanina. Todas as características descritas possuem um amplo potencial para a pesquisa, por exemplo, os quádruplos de Guanina estão ligados com a transcrição / regulação gênica e a formatação cruciforme com seu impacto direto no enrolamento do DNA, podendo sobrepor sítios de ligação de proteínas e afetar as interações que promovem a regulação gênica. A importância das características estruturais do DNA na regulação gênica bem como na necessidade de compreender os mecanismos de controle estrutural em múltiplas posições do genoma é reforçada. O conhecimento aprofundado de tais características pode auxiliar na criação de alvos terapêuticos e no desenvolvimento de estratégias eficientes para o combate de doenças, tendo como alvo direto a expressão gênica.

As diferentes técnicas de inteligência artificial, como a *Support Vector Machine* (SVM) e redes neurais artificiais, são utilizadas para a predição de promotores bacterianos. Não foram encontrados trabalhos utilizando o algoritmo K-means em sequências promotoras. Os resultados não são diretamente comparáveis, visto as diferenças técnicas, mas é possível contextualizar os resultados obtidos neste trabalho. Além disso, há uma baixa quantidade de trabalhos que analisam outros fatores σ além do σ70. Gordon et al. (2003), utilizando um *kernel* de alinhamento de sequências (através da abordagem de SVM), obtiveram valores para a exatidão de 84%, especificidade de 84% e sensibilidade de 82% para sequências reconhecidas pelo fator σ70. Utilizando redes neurais artificiais, Rani et al. (2007) obteve valores acima de 90% para exatidão, especificidade e sensibilidade, utilizando as sequências de promotores reconhecidos pelo fator σ70.

Os trabalhos de Gordon et al. (2003) Rani et al. (2007) utilizaram a composição de nucleotídeos como parâmetros para o treinamento das redes neurais. Em contraste, autores como Askary et al. (2009) e de Avila e Silva et al. (2014) utilizaram os valores de estabilidade da sequência como parâmetro. No caso de Askary et al. (2009), foram utilizadas sequências com tamanho de 413 nucleotídeos, em um total de 467 sequências, todas com seus Sítios de Início de Transcrição estabelecidos experimentalmente. O valor obtido para a exatidão foi de 94%. Em uma abordagem mais ampla, utilizando 6 fatores σ de *E. coli*, de Avila e Silva et al. (2014) apresenta a estabilidade como característica que distingue os promotores reconhecidos pelos diferentes fatores σ. No trabalho, ficou evidenciada a distinção entre os fatores σ28 e σ54, cada um com valores de exatidão de 80,2% e 78,8%, respectivamente. Para os fatores σ24 e σ32, foi encontrada uma exatidão de, respectivamente, 58% e 64%. Uma das diferenças do trabalho aqui proposto em relação aos demais apresentados é o uso da técnica de clusterização em conjunto com um parâmetro de estrutura física do DNA. Além de trabalhar com dois fatores σ que compartilham a mesma função, evidenciando as dificuldades em trabalhar *in silico* com tais conjuntos de dados.

4. Considerações Finais

O presente trabalho dedicou-se à análise *in silico* dos promotores relacionados ao estresse por choque térmico, regulados pelos fatores σ24 e σ32. A análise foi realizada utilizando a técnica da clusterização, com os dados codificados em seus valores de estabilidade. Foram realizadas diversas simulações com um valor de K entre 8 até 20 agrupamentos. A

pureza média obtida para a simulação de K = 12 foi de 63%, com o destaque de alguns agrupamentos com pureza acima de 80% (*clusters* 7, 10 e 11). A análise posterior à utilização da clusterização foi realizada com o auxílio da ferramenta WebLogo [Crooks 2004]. Dentro dos grupos com maior pureza, apenas a região -35 apresentou-se degenerada, enquanto a região -10 não divergiu do consenso descrito na literatura. Foi possível observar que o fator σ32 apresenta uma maior incidência da região -10 extendida, descrita por Lim et al. (2013) e Koo et al. (2009). Adicionalmente, não é possível encontrar similaridades com o consenso descrito na literatura para as duas regiões consensuais desde fator, a falta de especificidade neste conjunto de dados pode explicar uma baixa pureza na geração dos agrupamentos.

Biologicamente, as estratégias para funcionamento da RNAp e a degeneração encontrada nos motivos consensuais não acarreta em dificuldades para a transcrição gênica. Porém, evidenciam as dificuldades em analisar computacionalmente os dados biológicos. Deste modo, os diferentes agrupamentos apresentados neste trabalho contribui-se para a compreensão dos diferentes perfis biológicos encontrados nos promotores bacterianos. Além de evidenciar o uso da estabilidade como parâmetro para análises *in silico* e auxilia na redução de falsos positivos em ferramentas de predição de promotores, como a ferramenta BacPP [de Avila e Silva, et al., 2011]. Com base no que foi exposto, pretende-se expandir a pesquisa para os demais fatores σ alternativos, ampliando os parâmetros para codificação dos dados, a fim de utilizar a curvatura como critério de classificação.

4. Referências Bibliográficas

Askary, A., Masoudi-Nejad, A., Sharafi, R., Mizbani, A., Parizi, S. N. e Purmasjedi, M. (2009). N4: A precise and highly sensitive promoter predictor using neural network fed by nearest neighbors. In *Genes & Genetic Systems* (84) (6), páginas 425-430.

Attwood, T. K. et al. (2011). Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective, In Trends and Methodologies, Editado por Mahmood A. Mahdavi, InTech, Croácia.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. In *Studies in History and Philosophy of Biological and Biomedical Sciences*, páginas 69-80.

Crooks, G. E. et al. (2004). WebLogo: A Sequence Logo Generator. In *Genome Research* (14) (6), páginas 1188-1190.

de Avila e Silva, S. e Echeverrigaray, S. (2012). "Bacterial Promoter Features Description and Their Application on *E. coli* in silico Prediction and Recognition Approaches, In Bioinformatics, Editado por Horácio Pérez-Sánchez, InTech, Croácia.

de Avila e Silva, S., et al. (2014). DNA duplex stability as discriminative characteristic for Escherichia coli σ54- and σ28- dependent promoter sequences. In *Biologicals* (42) (1), páginas 22-28.

de Avila e Silva, S., Echeverrigaray, S. e Gerhardt, G. J. L. (2011). BacPP: Bacterial promoter prediction - A tool for accurate sigma-factor specific assignment in enterobacteria. In *Journal of Theoretical Biology* (287), páginas 92-99.

Gordon, L., et al. (2003). Sequence alignment for recognition of promoter regions. In *Bioinformatics* (19) (15), páginas 1964-1971.

Herbert., M. K. et al. (2006). Sequence-Resolved Detection of Pausing by Single RNA Polymerase Molecules. In *Cell* (125) (6), páginas 1083-1094.

Houten, B. V. e Kisker, C. (2014). Transcriptional pausing to scout ahead for DNA Damage. In *Proceedings of the National Academy of Sciences* (111) (11), páginas 3905-3906.

Jáuregui, R. et al. (2003). Conservation of DNA curvature signals in regulatory regions of prokaryotic genes. In *Nucleic Acids Research*, páginas 6770-6777.

Kanehisa, S. et al. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *In Nucleic Acids Research* (42), páginas D199-D205.

Kanhere, A. e Bansal, M (2005). A Novel method for prokaryotic promoter prediction based on DNA stability. In *Bioinformatics* (6) (1), páginas 1-10.

Kaushik, M. et al. (2016). A bouquet of DNA structures: Emerging diversity. In *Biochemistry and Biophysics Reports* (5), páginas 388-395.

Koo, B. M. et al. (2009). Dissection of recognition determinants of *Escherichia coli* σ32 suggets a composite -10 region with na 'extended -10' motif and a core -10 element. In *Molecular Microbiology* (72) (4), páginas 815-829.

Krebs, J.,Goldstein, S. e Kilpatrick, S. T. (2014) Genes XI, ed. Sudbury, Massachusetts: Jones and Bartlett, 930 p.

Lim, B. et al. (2013). Heat Shock Transcription Factor σ32 Co-opts the Signal Recognition Particle to Regulate Protein Homeostasis in *E. coli*. In *PLOS Biology* (11) (12), páginas 1-15.

Marx, V. (2013). The Big Challenges of Big Data. In *Nature* (498), páginas 255-260.

Ramprakash, J. e Schwarz, F. P. (2007) Identification and annotation of promoters regions in microbial genome sequences on the basis of DNA stability. In *Journal of Biosciences* (32), páginas 851-862.

Rangannan, V. e Bansal, M. (2007). Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. In *Journal of Biosciences*, páginas 851-862.

Rani, T. S., Bhavani, S. D. e Bapi, R. S. (2007). Analysis of *E. coli* promoter recognition problem in dinucleotide feature space. In *Bioinformatics* (23), páginas 582-588.

Salgado, H. et al. (2013). RegulonDB v. 8: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and morre. In *Nucleic Acids Research* (41), páginas D203-D213.

SantaLucia J and Hicks D (2004). The thermodynamics of DNA structural motifs. *Annual review of biophysics and biomolecular structure* (33), páginas 415-440.

Vvedenskaya, I. O. et al. (2014). Interactions between RNA polymerase and the "core recognition element" counteract pausing. In *Science* (344), páginas 1285-1289.

Witten, I. H. e Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Ed. San Francisco: Morgan Kaufman, 560p.