Comparação entre diferentes modelos de cálculo de curvatura do DNA como parâmetro de predição e reconhecimento *in silico* de promotores alternativos de *Escherichia coli*

Scheila de Avila e Silva[1], Rafael Coelho[2], Priscila Portela[1], Jeane Paz[1], Sergio Echeverrigaray[1]

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

[2] Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul - Câmpus Farroupilha
Av. São Vicente, 785 – 95180-000 – Farroupilha – RS – Brazil

{sasilva6@ucs.br,rafael.coelho@farroupilha.ifrs.edu.br,pri-portela@hotmail.com,jeane_paz@yahoo.com.br, selaguna@ucs.br}

Abstract. *Artificial intelligence approaches have been applied in biology due the growing of data available in specific repositories. In this context, it is possible to provide relevant insights in gene regulation context by carrying out machine learning methodologies. This paper describes the application of artificial neural networks in prediction and recognition of E. coli alternative σ-dependent promoter sequences. Additionally, the DNA curvature was used in the repertoire of neural networks as a classification parameter. This both procedures are underexploited in the related literature. For each data set, the accuracies obtained with this approach were: σ^{24}: 72,36%; σ^{28}: 67,17%; σ^{32}: 72,67%; σ^{38}: 75,45%. These results are an indicate of DNA curvature potential as discriminative feature of promoters. Besides, this characteristic can be combined with other promoter information in order to improve promoter prediction tools.*

Resumo. *As tecnologias de inteligência artificial ganham aplicabilidade nas áreas científicas, como biologia, devido ao aumento de dados disponíveis em repositórios específicos. Considerando este contexto, é possível obter inferências relevantes com a aplicação destas técnicas em questões relacionadas à regulação gênica. Assim, este trabalho descreve a aplicação de redes neurais artificiais no reconhecimento e predição de promotores associados a fatores σ alternativos de E. coli. Além da separação por fator σ, este trabalho apresenta como diferencial, a exploração da curvatura da molécula de DNA como parâmetro de classificação. A exatidão obtida com a aplicação desta metodologia, para os diferentes fatores σ foi de: σ^{24}: 72,36; %; σ^{28}: 67,17%; σ^{32} : 72,67; σ^{38} :75,45%. Estes resultados indicam o potencial da curvatura como uma característica a ser incorporada nas ferramentas de predição de promotores a fim de diminuir o número de falsos positivos.*

1. Introdução

Um dos maiores desafios da era pós-genômica é a determinação de quando e como os genes são “ligados” e “desligados”. A diferença entre duas espécies está muito mais relacionada com a transcrição de seus genes do que com a estrutura destes em si. Deste modo, o estudo da regulação gênica contribui para a construção do conhecimento a

respeito da funcionalidade dos genes, entre outras questões [Howard e Benson 2003]. Dentre as sequências de DNA que atuam como reguladoras da expressão gênica estão as regiões promotoras. De uma maneira simplificada, pode-se dizer que estas localizam-se anteriormente à região codificante, interagem com a enzima RNA polimerase (RNAp), desencadeando o processo de transcrição [Lewin 2008]. Fazendo uma analogia, as regiões codificantes representam a memória de um computador e os promotores, os programas que atuam nesta memória. Assim, o estudo dos promotores pode prover modelos sobre a constituição do "programa" e de como este opera [Howard e Benson 2003].

Em organismos procarióticos, a RNAp é formada por cinco subunidades e uma subunidade adicional (que se liga de forma transitória) chamada fator sigma (σ). A coleção de diferentes σ é responsável pela expressão de genes específicos de resposta às mudanças ambientais. Assim, os genes associados com promotores reconhecidos pelos diferentes σ atuam: *(i)* σ^{32} e σ^{24} na resposta ao estresse por choque térmico; *(ii)* σ^{28} na expressão de cílios e flagelos; *(iii)* σ^{54} no metabolismo de nitrogênio e; *(iv)* σ^{70} na maioria dos genes de atuação geral [Lewin 2008].

Há duas regiões biologicamente importantes nos promotores, conhecidas como região -35 e região -10. Os promotores possuem similaridades na composição e localização dos nucleotídeos que compõe estas regiões, por esta razão, elas são chamadas de consensuais. No entanto, o grau de conservação varia entre os promotores reconhecidos por um determinado fator σ ou entre as sequências reconhecidas por diferentes fatores σ [Lewin 2008]. Esta falta de conservação limita a análise global dos dados, sendo a separação dos promotores conforme o fator σ, um critério importante.

Os motivos consensuais são sinais importantes, mas não a única fonte de informação no genoma para a regulação da expressão gênica [Olivares-Zavaleta et al. 2006]. Além destes locais, os promotores possuem características estruturais próprias, tais como deformabilidade, estabilidade e curvatura [Kanhere e Bansal 2005], [Kozobay-Avraham et al. 2008].

A curvatura do DNA está envolvida em processos biológicos como transcrição, recombinação e replicação do DNA. Em organismos procariontes (como as bactérias), a curvatura apresenta-se mais acentuada na região antecedente ao promotor, mas é possível encontrá-la na própria região promotora [Kozobay-Avraham et al. 2008]. Para realizar o cálculo da curvatura do DNA são analisados parâmetros como os ângulos de *roll* (enovelamento), *twist* (torção) e *tilt* (inclinação) dos nucleotídeos. Existem modelos disponíveis para obtenção destes valores, como os propostos por *(i)* Bolshoy et al. (1991), o qual determinaram os ângulos da curvatura por meio dos resultados obtidos com eletroforese em gel de agarose; *(ii)* Ulanovsky e Trifonov (1987), que atribuem a curvatura do DNA ao ângulo dos dinucleotídeos AA e, *(iii)* Calladine et al. (1988), que formularam seu modelo com base em resultados de cristalografia de raio-x.

As mais variadas abordagens computacionais têm sido empregadas para reconhecer e predizer se uma região é ou não promotora de um gene. Embora haja progressos na predição e análise de promotores, estes ainda estão longe de possuir uma alta acurácia, já que os resultados apresentam um alto índice de falsos positivos [Rani et al. 2007]. Assim, a análise de outras características (como a curvatura do DNA) torna-se importante para o aprimoramento das técnicas *in silico* de predição de promotores. Deste modo, o presente trabalho tem o objetivo de predizer e reconhecer promotores reconhecidos por fatores σ alternativos de *Escherichia coli* por meio de uma abordagem de redes neurais artificiais. Para isso foram utilizados como parâmetro de entrada os

valores de curvatura, ângulo da curvatura e ângulo da maleabilidade obtidos com três modelos para cálculo da curvatura estática do DNA. Deste modo, foi possível avaliar a eficiência destes modelos na metodologia aplicada neste trabalho.

2. Metodologia

2.1. Preparação dos dados

As sequências promotoras da bactéria *E. coli* foram obtidas do banco de dados RegulonDB [Salgado et al. 2013], totalizando 1162 exemplos positivos distribuídos nos diferentes fatores σ que reconhecem a sequência: σ^{24} (515 exemplos); σ^{28} (141 exemplos); σ^{32} (285 exemplos); σ^{38} (129 exemplos) e σ^{54} (92 exemplos). Os exemplos negativos foram extraídos aleatoriamente de regiões intergênicas não-promotoras em mesmo número dos exemplos positivos.

O cálculo da curvatura estática do DNA foi realizado utilizando a ferramenta *DNA Curvature*, disponível *on-line* [Gohlke 2014]. Após a geração dos valores, estes foram suavizados com um filtro passa-baixa. Para encontrar o número de interações que ameniza as irregularidades do sinal sem descaracterizá-lo, o processo de suavização foi realizado com 3, 5, 8 e 12 graus.

2.2. Simulação de redes neurais

As redes neurais são adequadas para o problema de predição de promotores devido a sua capacidade de identificar padrões degenerados, imprecisos e incompletos, como os apresentados por estas sequências [Burden et al. 2005], [de Avila e Silva e Echeverrigaray, 2012]. A arquitetura de rede neural utilizada foi a *multilayer perceptron* (MLP) com três camadas: uma camada de entrada, uma camada oculta e uma camada de saída. Para cada conjunto de promotores reconhecidos por um determinado fator σ, diferentes arquiteturas, com a combinação cruzada dos seguintes parâmetros foram testadas:

1. Número de neurônios na camada de entrada: 80 neurônios.
2. Número de neurônios na camada oculta: 1 até 8.
3. Número de neurônios na camada de saída: 1.

Estas combinações foram realizadas utilizando como parâmetro de entrada os valores de curvatura, ângulo da curvatura e ângulo da maleabilidade (separadamente) obtidos com três modelos diferentes, aqui denominados: Calladine [Calladine et al. 1988], Bolshoy [Bolshoy et al. 1991] e AA-Wedge [Ulanovsky e Trifonov 1987]. Todas as simulações foram realizadas no ambiente R [R Development Core Team, 2008] com o algoritmo *back-propagation*. A metodologia de validação cruzada (*k-fold cross-validation*) foi escolhida por proporcionar resultados válidos estatisticamente. Assim, a cada iteração um dos *k* subconjuntos foi usado para testar o aprendizado obtido a partir dos outros subconjunto, os quais formaram o conjunto de treinamento. O erro médio obtido nas k tentativas foi computado [Polate e Günes, 2007]. O valor de k foi diferente para cada conjunto de dados, sendo este determinado conforme o número de sequências promotoras disponíveis. Assim, os valores de k atribuído foram de: 2 (σ^{54}), 3 (σ^{28}, σ^{32} e σ^{38}) e 5 (σ^{24}).

O desempenho da rede neural foi avaliado pelos valores de exatidão (A), especificidade (S) e sensibilidade (SN), conforme as fórmulas (1), (2) e (3), respectivamente [Sonego et al. 2008].

$$A = \frac{TP + TN}{TN + TP + FN + FP} \quad (1)$$

$$S = \frac{TN}{TN + FP} \quad (2)$$

$$SN = \frac{TP}{TP + FN} \quad (3)$$

onde, VP, VN, FP, e FN representam, respectivamente: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Esses valores indicam sequências promotoras classificadas como promotoras (VP); sequências promotoras classificadas como não-promotoras (VN); sequências não-promotoras classificadas como promotoras (FP) e sequências não-promotoras classificadas como não-promotoras (FN).

Os valores de ponto de corte para a classificação de uma determinada sequência foram obtidos por meio da análise da curva ROC (*Receiver Operator Characteristic Curve*). Para uma sequência ser considerada promotora, o valor atribuído a ela pela rede neural deveria ser maior que o ponto de corte. Informações adicionais sobre a curva ROC pode se encontrada em [Sonego et al. 2008].

3. Resultados e Discussão

O resultado inicial de um treinamento de rede neural é a definição da arquitetura que apresenta o menor valor de erro médio quadrático durante o processo de treinamento e teste. Portanto, conforme este critério, para cada conjunto de promotores reconhecidos por um determinado fator σ, foram selecionadas as arquiteturas apresentadas na Tabela 1, dentre um total de 154350 simulações diferentes. Para o conjunto de promotores reconhecidos pelo fator σ^{54}, não houve nenhuma arquitetura que fornecesse valores de erro médio quadrático para o conjunto de teste menor que 0,5.

Tabela 1: Melhor arquitetura de rede neural obtida em cada conjunto de dados

Fator σ	Modelo utilizado	Parâmetro de entrada	Número de iterações no Low121	Camada de entrada	Camada oculta	Camada de saída
σ^{24}	AA Wedge	Curvatura	5	80	7	1
σ^{28}	Calladine	Ângulo da Maleabilidade	12	80	4	1
σ^{32}	AA Wedge	Ângulo da Maleabilidade	5	80	4	1
σ^{38}	AA Wedge	Ângulo da Maleabilidade	5	80	6	1

Após a análise das arquiteturas, os valores de corte foram estabelecidos a fim de estabelecer um limiar de classificação para as medidas de performance analisadas. Desta

maneira, os limites foram 0,5 para os promotores reconhecidos pelo σ^{28} e 0,6 para os demais fatores σ. Os valores das medidas de performance exatidão, sensibilidade e especificidade para as arquiteturas de redes neurais selecionadas (Tabela 1) são apresentados, na mesma ordem, na Tabela 2. O acréscimo de neurônios na camada oculta não aumentou os valores das medidas de performance.

Tabela 2: Resultados das medidas de performance obtidos para a melhor arquitetura de redes neurais de cada conjunto de dados

Fator σ	Exatidão (%)	Sensibilidade (%)	Especificidade (%)
σ^{24}	72,36	74,43	76,28
σ^{28}	67,17	64,81	69,44
σ^{32}	72,67	72,34	72,99
σ^{38}	75,45	74,77	76,09

Os valores de exatidão para as arquiteturas apresentadas neste trabalho variaram entre 67,17% (σ^{28}) até 75,45% (σ^{24}). Outra característica importante está relacionada com os valores de sensibilidade e especificidade para cada conjunto de dados. Em todas as arquiteturas, estas duas medidas não foram discrepantes entre si. Isso significa que a rede neural consegue reconhecer, equiparadamente, promotores e não-promotores. Em muitos trabalhos de reconhecimento de promotores (ver seção trabalhos relacionados), a capacidade de reconhecer promotores é muito maior do que a capacidade de reconhecer não-promotores [de Avila e Silva e Echeverrigaray, 2012].

Os fatores σ com menor valor nas medidas de performance foram os σ^{28} e σ^{54} (o qual não apresentou nenhum resultado de classificação maior que 0,5). Estes valores podem estar associados ao menor conteúdo AT apresentado por estas sequências em relação aos demais conjuntos de dados. Isto é particularmente relevante, uma vez que esta informação relaciona-se com o algoritmo que apresentou os maiores valores de medida de performance. É possível perceber a prevalência do modelo AA-Wedge, o qual atribui maior peso aos nucleotídeos AA. Assim, como os promotores possuem maior conteúdo AT, esta característica tornou-se evidente para o aprendizado da rede neural [de Avila e Silva et al. 2014].

Adicionalmente, estes resultados são complementares aos apresentados por de Avila e Silva et al. (2014), que mostrou a estabilidade como um parâmetro que distingue, justamente, os fatores σ^{28} e σ^{54} (conjuntos de dados com menor exatidão). Assim, percebe-se o potencial de utilização deste parâmetro como uma característica distintiva dos promotores e, uma análise posterior (como extração de regras das redes neurais treinadas) pode contribuir na diminuição do número de falsos positivos das ferramentas de predição de promotores [de Avila e Silva et al. 2011].

3.1. Trabalhos relacionados

A maioria dos trabalhos relacionados à predição e reconhecimento de promotores é aplicada apenas às sequências reconhecidas pelo fator σ^{70} utilizando com parâmetro de classificação as informações dos nucleotídeos da molécula. Diferentemente dos demais trabalhos, este artigo utilizou a abordagem de redes neurais artificiais na predição e reconhecimento de regiões promotoras de acordo com o fator σ que reconhece as sequências. Além disso, os valores de entrada utilizados para a realização do

treinamento (curvatura da molécula de DNA) também são pouco explorados.

Por meio de uma abordagem de redes neurais artificiais, Rani et al. (2007) utilizaram a informação dos pares de nucleotídeos de promotores reconhecidos pelo fator σ^{70} como parâmetro de entrada. Os autores apresentaram resultados de exatidão de 96%, especificidade de 98% e sensibilidade 93%. Os exemplos negativos para este conjunto de dados foram sequências aleatórias com 60% de nucleotídeos A-T.Outra ferramenta baseada na composição dos nucleotídeos é o BacPP [de Avila e Silva et al. 2011], originado a partir da ponderação das regras extraídas de redes neurais treinadas. Esta ferramenta apresenta os seguintes valores de exatidão para os diferentes fatores σ de *E. coli*: 86.9% (σ^{24}); 92.8% (σ^{28}); 91.5% (σ^{32}); 89.3% (σ^{38}); 97.0% (σ^{54}) e 83.6% (σ^{70}). Na análise de promotores de outras bactérias Gram-negativas, a ferramenta apresenta uma acurácia de 76%.

A utilização de parâmetros de entrada para as redes neurais, que não diretamente a sequência de nucleotídeos, foi realizada por Askary et al. (2009) e de Avila e Silva et al. (2014). Nestes casos, foram utilizados os valores de estabilidade da sequência. A metodologia de Askary et al. (2009) baseou-se na análise de sequências com tamanho de 413 nucleotídeos, a fim de determinar a posição do primeiro nucleotídeo da sequência codificadora. As predições realizadas mostraram uma exatidão de 94%. Por outro lado, de Avila e Silva et al. (2014), utilizando sequências promotoras com 80 nucleotídeos, mostrou que a estabilidade é uma característica que distingue, principalmente, os promotores reconhecidos pelo fator σ^{28} e σ^{54}. Neste trabalho, as simulações de redes neurais foram realizadas utilizando 6 fatores de *E.* coli, sendo que os valores para as medidas de performance foram melhores para os fatores σ^{28} e σ^{54}. Nestes casos, respectivamente, a exatidão obtida foi de 80,2% e 78,8%.

Outras técnicas de inteligência artificial, como *support vector machines* (SVM) também são aplicadas na predição de promotores. Utilizando um *kernel* de alinhamento de sequências, Gordon et al. (2003), obtiveram exatidão de 84%, especificidade de 84% e sensibilidade de 82%, para promotores reconhecidos pelo fator σ^{70}. Adicionalmente, Polate and Günes (2007) descrevem a aplicação de l*east-square support vector machines* com exatidão de 84,6%, sensibilidade de 90,9% e especificidade de 80%. Além destes trabalhos, Song (2012) utilizou o método chamado Z-curve para extrair características dos promotores procariontes. Estas foram utilizadas como dados de entrada em uma *partial least squares technique*. O autor apresenta como resultados os seguintes valores de exatidão, conforme o conjunto de exemplos positivos e negativos usados como entrada: 96,05% (promotores σ^{70} e regiões codificantes); 90,44% (promotores σ^{70} e regiões não-codificantes); 92,13% (promotores de outros fatores σ e regiões codificantes) e 92,50% (promotores de outros fatores e regiões não codificantes). As exatidões obtidas por Song (2012) são maiores que as obtidas individualmente neste artigo. No entanto, estes resultados não podem ser diretamente comparados uma vez que o autor não especifica quais conjuntos de promotores foram utilizados.

Além destas abordagens, Ranganann e Bansal (2007) desenvolveram seu próprio método de predição de promotores. Os autores utilizaram seu os valores de estabilidade da molécula de DNA como parâmetro classificatório. Como resultado, eles obtiveram uma exatidão de 52,2% e uma sensibilidade de 99%. Estes resultados mostram que o número de falsos positivos ainda é um problema na predição de promotores.

Com exceção de Avila e Silva et al. (2014) e de Avila e Silva et al. (2011), os trabalhos relacionados tratam apenas das sequências reconhecidas pelo fator σ^{70}, os

quais não foram tratados neste trabalho. Assim, devido as peculiaridades da composição de nucleotídeos dos promotores reconhecidos por fator σ alternativos, não é possível realizar uma comparação direta dos valores de exatidão dos trabalhos relacionados com resultados deste trabalho. O diferencial aqui apresentado está na utilização da curvatura como parâmetro classificador e na preocupação em realizar a predição em promotores reconhecidos por fator σ alternativos, os quais desempenham importantes papeis em questões como adaptação ao ambiente e patogenicidade.

4. Considerações finais

O presente trabalho apresentou uma abordagem de redes neurais artificiais aplicadas na predição de promotores de *E. coli* de acordo com o fator σ que reconhece a sequência. A separação do conjunto de dados facilita o processo de aprendizado da rede neural, o que aumenta a robustez da classificação. Além disso, esta abordagem extrapola o contexto de promotores reconhecidos apenas pelo fator σ^{70}, uma vez que os fatores σ alternativos estão amplamente distribuídos entre as bactérias. Outro aspecto importante é a utilização de parâmetros de entrada além da composição de nucleotídeos diretamente. Sabe-se que os promotores possuem sinais próprios que os diferenciam das demais regiões genômicas. Assim, estes parâmetros podem ser explorados a fim de contribuir com novas inferências para o problema em questão.

As exatidões obtidas neste trabalho foram, para cada conjunto de dados: σ^{24}: 72,36%; σ^{28}: 67,17%; σ^{32} : 72,67%; σ^{38} :75,45%. Estes valores indicam dois diferenciais principais: o uso da curvatura da sequência como parâmetro classificatório e a análise dos fatores σ alternativos. Assim, pode-se perceber o potencial de aplicação da curvatura em combinação com outras informações estruturais dos promotores Deste modo, contribui-se para a redução do número de falsos positivos nas técnicas de predição atuais. Portanto, como continuação deste trabalho, pretende-se realizar o processo de extração de regras das redes neurais treinadas utilizando os pesos dos neurônios da camada oculta. Após, estas regras serão algoritmizadas e implementadas na ferramenta de predição de promotores BacPP [de Avila e Silva et al. 2011].

References

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.

Bolshoy, A., McNamara, P., Harrington, R. E. e Trifonov, E. N. (1991). Curved DNA without A–A: experimental estimation of all 16 DNA wedge angles. In *Procedings of National Academic of Science of United States of America* (88), páginas 2312–2316.

Burden, S., Lin, Y.-X., e Zhang, R. (2005). Improving promoter prediction for the NNPP2.2 algorithm: a case study using Escherichia coli DNA sequences. In *Bioinformatics* (21) (5), páginas601-607.

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. 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., 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.

Calladine, C. R., Drew, H. R. e McCall, M. J. (1988). The intrinsic curvature of DNA in solution. In *Journal of Molecular Biology* (201), páginas 127–137.

Gohlke, C. (2014) "Dnacurve.py", http://www.lfd.uci.edu/~gohlke/code/dnacurve.py.html , Abril

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

Howard, D., Benson, K. (2003). Evolutionary computation method for pattern recognition of cis-acting sites. In *BioSystems* (72) (1/2), páginas19-27.

Kanhere, A. e Bansal, M. (2005). Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. In *Nucleic Acids Research (*33) (10), páginas 3165-3175.

Kozobay-avraham, L., Hosid, S., Volkovich, Z. e Bolshoy, A. (2008). Prokaryote Clustering based on DNA curvature distributions. In *Discrete Applied Mathematics* (11), páginas 2378-2387.

Lewin, B. (2008), Genes IX, Artes Médicas, 9ª edição.

Olivares-Zavaleta, N., Jáuregui, R. e Merino, E. (2006). Genome analysis of *Escherichia coli* promoters evidences that DNA static curvature plays a more important role in gene transcription than has previously been antecipated. In *Genomics* (87), páginas 329-337.

Polate, K. e Günes, S. (2007). A novel approach to estimation of *E. coli* promoter gene sequences: Combining feature selection and least square support vector machine (FS_LSSVN). In *Applied Mathematics and Computation* (190), páginas 1574-1582.

R Development Core Team. (2008). "R: A language and environment for statistical computing". URL: http://www.R-project.org, Abril.

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* 32 (5), 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), páginasD203-D213.

Ulanovsky, L. E. e Trifonov, E. N. (1987). Estimation of wedge components in curved DNA. In *Nature* (326), páginas720–722.

Sonego, P., Kocsor, A. e Pongor, S. (2008). ROC analysis: applications to the classification of biological sequences and 3D structures. In *Briefings in Bioinformatics* (9) (3), páginas 198-209.

Song, K. (2012). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method, In *Nucleic Acids Research* (40) (3), páginas963-971.