SaBios
Revista de Saúde e Biologia

ARTIGOS COMPLETOS/COMPLET ARTICLES

ANÁLISE *IN SILICO* DE PROMOTORES DE *ESCHERICHIA COLI* RECONHECIDOS PELO FATOR σ^{28}

IN SILICO ANALYSIS OF *ESCHERICHIA COLI* PROMOTERS RECOGNIZED BY σ^{28} FACTOR

Gabriel Dall Alba[1], Scheila de Avila e Silva[1,2], André Gustavo Adami[3], Sergio Echeverrigaray[1]

[1] Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul, RS, Brasil.
[2] Campus de Vacaria, Universidade de Caxias do Sul, Vacaria, RS, Brasil.
[3] Centro de Ciências Exatas e da Tecnologia, Universidade de Caxias do Sul, Caxias do Sul, RS, Brasil.
Endereço para correspondência: Instituto de Biotecnologia, Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130, Caxias do Sul, RS, Brasil. Cep:95070-560. Phone +55 54 3218-2149.
E-mail:sasilva6@ucs.br

RESUMO

A regulação da transcrição gênica em seres procariotos desempenha um papel importante para a resposta adequada destes organismos às mudanças ambientais. Neste processo, a especificidade da expressão dos genes se dá por meio da ligação do fator σ na enzima RNA polimerase, e o posterior reconhecimento do promotor. O objetivo deste trabalho foi analisar a composição das sequências promotoras reconhecidas pelo fator σ^{28}(relacionado com mobilidade e patogenicidade bacteriana) e a relação proteína-proteína dos produtos biológicos associados a estas sequências. Os promotores de *Escherichia coli* foram agrupados utilizando a técnica de mineração de dados denominada de clusterização, com o algoritmo *k-means*. O conteúdo dos agrupamentos foi analisado com as ferramentas Weblogo, String-DB e Gene Ontology. Os resultados obtidos mostram que os agrupamentos formados apresentam o conteúdo da sequência divergente ao padrão biológico canônico. Adicionalmente, a análise da interação proteína-proteína indica que a função celular não está relacionada diretamente com a estrutura de nucleotídeos dos promotores, uma vez que este último foi o critério a utilizado pelo algoritmo K-*means* para realizar os agrupamentos.
Palavras-Chave: promotor; clusterização; patogenicidade.

ABSTRACT

The regulation of gene expression in prokaryotes provides the adequate response to environmental changes. The recognition of the promoter sequence plays an important role in the specificity of gene expression, since σ factor binds in RNA polymerase enzyme starting the process. In this context, the aim of this study was to analyze the composition of promoter sequences recognized by σ28 factor of *Escherichia coli* (related to mobility and bacterial pathogenicity) and protein-protein ratio of organic products associated with these sequences. The promoters were grouped by clustering (a data mining technique) with k-means algorithm. The content of clusters was analyzed with Weblogo, String-DB and Gene Ontology tools. The sequence of clusters shows some degree of discrepancy with canonical biologic pattern. Besides, none of the clusters presented metabolic function specificity. Furthermore, the protein-protein interaction analysis indicates that there is no relation between cellular function and nucleotide content, since this was the main criterion used by k-means algorithm in the generation of clusters.
Keywords: promoter; clustering; pathogenicity.

INTRODUÇÃO

O aumento no desempenho e na capacidade de armazenamento dos dispositivos computacionais proporciona novas perspectivas para diversas áreas de pesquisa. Neste contexto, a tecnologia computacional permite analisar uma grande quantidade de dados, buscando cruzar informações e obter relacionamentos que não seriam possíveis sem a aplicação destas técnicas. Assim, a Bioinformática apresenta-se como um ramo de pesquisa que se apropria das técnicas computacionais para a geração de inferências em diversos segmentos biológicos (1).

Seres procariotos possuem uma organização genômica diferenciada em relação a eucariotos. A ausência de núcleo requer, além de mecanismos de proteção do genoma, formas de seletividade na expressão de genes, a qual ocorre, principalmente, no momento da transcrição. Assim, é possível a um organismo apresentar uma resposta adequada aos estímulos ambientais e sobreviver em determinadas condições (2).

Em bactérias, o processo responsável pela expressão de genes em um determinado momento celular, inicia-se com o reconhecimento de segmentos de DNA chamados de promotores. Nesta região, antecedente à região codificante, liga-se uma proteína chamada RNA Polimerase (RNAp). Esta interação é essencial, uma vez que o reconhecimento do promotor é fundamental para a transcrição do gene associado a ele. Deste modo, pode-se afirmar que o promotor atua como um elemento regulatório da expressão gênica (2).

Para auxiliar na identificação de um promotor, a RNAp possui uma subunidade sigma (σ), a qual irá identificar um promotor específico. Por exemplo, em *Escherichia coli*, os genes reconhecidos pelo σ^{28} estão relacionados à mobilidade e patogenicidade bacteriana (2,3).O reconhecimento de uma determinada sequência como promotora se dá em regiões específicas, chamadas de motivos consensuais. Estes estão localizados em duas regiões do promotor, denominadas-10 e -35, os quais se referem ao primeiro nucleotídeo transcrito, que recebe numeração +1. A região -35 atua como sinal para que a RNAp reconheça a sequência, enquanto que na região -10 ocorre a abertura da fita de DNA (2,3).A composição de nucleotídeos para as regiões consensuais dos diferentes promotores bacterianos é apresentada na Tabela 1.

Tabela 1. Fatores σ de *E. coli*.

Fator σ	Gene	Função	Consenso -35 / -10
28	fliA	Mobilidade celular e patogenicidade	CTAAA 15 pb GCCGATAA
32	rpoH	Estresse por choque térmico	CCCTTGAA 13-15 pb CCCGATNT
38	rpoS	Resposta a estresse	TTGACA 16-18 pb TATACT
54	rpoN	Assimilação de Nitrogênio	CTGGNA 16-18 pb TTGCA
24	rpoE	Estresse por choque térmico	GGAACTT 15 pb GTCTAA
70	rpoD	Sigma constitutivo	TTGACA 16-18 pb TATAAT

Fonte: DE AVILA E SILVA E ECHEVERRIGARAY (3)

Apesar de serem chamados de sequência consenso, percebe-se que a composição de nucleotídeos é diferente entre os promotores reconhecidos pelos fatores σ.

Por exemplos, motivos consensuais dos promotores relacionados a genes de reposta ao choque térmico diferem das sequências consenso de genes relacionados

a patogenicidade, o que explica seu reconhecimento por diferentes fatores σ. Além disso, estas diferenças justificam o estudo dos promotores de forma individual, ou seja, realizado considerando as peculiaridades de cada grupo de sequências.

Considerando o exposto, este trabalho dedica-se a análise *in silico*(computacional) dos promotores reconhecidos pelo fator σ^{28}. O objetivo principal foi aplicar a técnica de aprendizado de máquina denominada clusterização, a fim de verificar se os agrupamentos resultantes apresentam relações biológicas, tanto de caráter estrutural quanto funcional. A escolha deste fator σ^{28} em particular, deve-se a sua relação com o reconhecimento dos genes de mobilidade bacteriana, o que influencia no processo de patogenicidade (4). Com o auxílio da quimiotaxia (interação entre o organismo com uma substância química no meio, causando a aproximação ou o afastamento do organismo a esta substância, obrigatório em organismos móveis),a bactéria é capaz de promover respostas específicas ao meio, fundamentais a sua sobrevivência (5). Nestes casos, elementos como a velocidade ou morfologia do flagelo bacteriano variam, tornando-se funcionais para a situação em que ele se encontra.

METODOLOGIA

O presente trabalho utilizou uma abordagem computacional para a obtenção dos objetivos propostos. Assim, a metodologia utilizada foi composta pelas etapas representadas na Figura 1. Todas as etapas da metodologia foram implementadas em linguagem de programação C# em plataforma desktop com a interface com o usuário apresentada na Figura 1 (6). Após as simulações, os agrupamentos resultantes foram analisados com o auxílio das ferramentas WebLogo (7) e dos bancos de dados String (8) e Gene Ontology (9).

A primeira etapa consistiu na coleta e seleção dos dados a serem analisados. As 144 sequências promotoras utilizadas neste trabalho foram retiradas do banco de dados RegulonDB (10) em sua versão de Junho de 2014. Para a aplicação do algoritmo de clusterização, foi necessário a codificação das letras constituintes da sequência em valores numéricos.

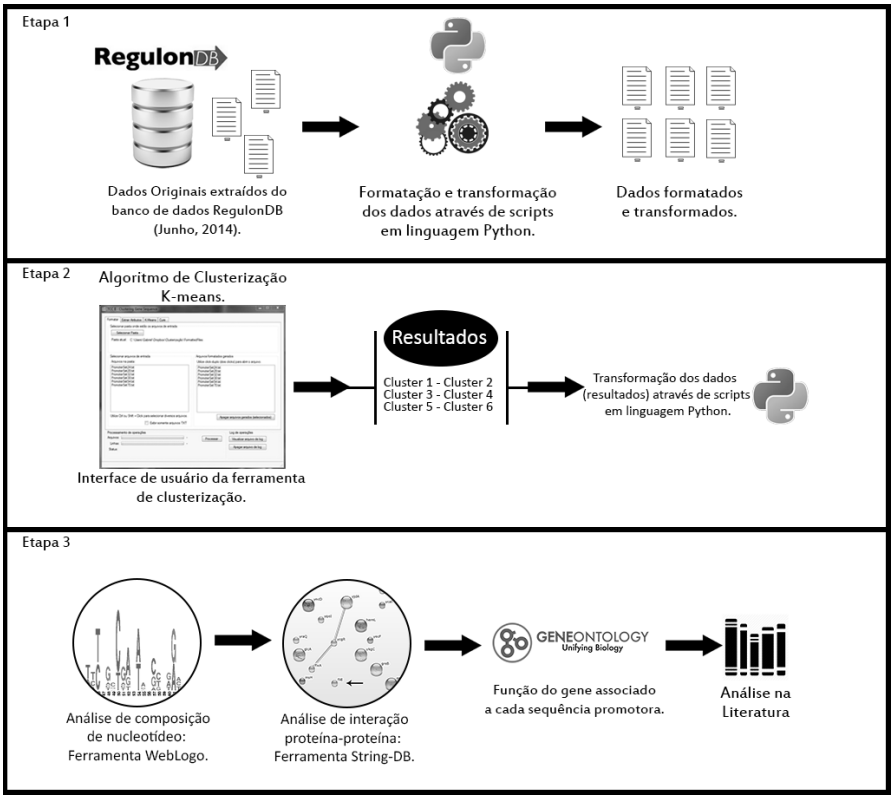

Figura 1. Representação esquemática da metodologia utilizada.

Neste trabalho, foi utilizada a codificação proposta por JI (2008), a qual propõe a construção de um vetor de atributos de acordo com a frequência de ocorrências dos elementos da sequência promotora. A construção do espaço vetorial iniciou-se com a seleção de um grupo de atributos de acordo com o critério estatístico da co-ocorrência. Este grupo é denotado por $\boldsymbol{T = \{t_1, t_2, \ldots t_m\}}$, representando o conjunto de termos para quantificação. Dado um objeto X, utiliza-se a representação Frequência de Termo – Inverso da Frequência no Documento (*Term Frequency – Inverse Document Frequency TF * IDF)* para formatar o conteúdo do objeto em um vetor, formulado por:

$$X = (x_1, x_2, \ldots, x_m)$$

$$= (tf_1 . idf_1, tf_2 . idf_2, \ldots tf_m . idf_m),$$

onde tf_i é a frequência de termos (TF) do termo t_i no objeto X, e idf_i é o inverso da frequência de objeto (IDF) de t_i, em relação a todo conjunto de dados, definido por:

$$idf_i = log \frac{df_i}{N}$$

Onde df_i é a frequência de objeto de t_i, ou seja, o número de objetos que contém t_i, e N é o total de objetos no conjunto de dados (11).

Neste trabalho, optou-se por aplicar a técnica de seleção de atributos em janelas de 2 nucleotídeos, uma vez que as sequências têm tamanho reduzido (81 nucleotídeos). A aplicação da técnica de seleção de atributos com uma janela com mais nucleotídeos acarretaria na diminuição do conteúdo informativo da sequência promotora.

Após a preparação dos dados, foi aplicada a técnica de clusterização, a qual busca padrões para a organização de agrupamentos. A técnica procura identificar objetos com semelhanças e dissemelhanças entre os demais agrupamentos gerados, sendo que cada agrupamento é denominado *cluster* (12).O algoritmo escolhido foi o K-*means*, o qual recebe este nome por apresentar uma variável a ser definida pelo pesquisador: o denominado *k*, com função de identificar o número de *clusteres* desejados para o trabalho (12). Posterior a esta etapa, escolheu-se de um centroide aleatório, o qual foi o modelo para o cálculo da distância entre o centroide e o conjunto de dados. A escolha por esta técnica baseou-se no princípio de que se trata de um método eficiente para encontrar padrões não visíveis de um conjunto de dados. No entanto, sabe-se que esta técnica seja sensível em relação à escolha do centroide, uma vez que agrupamentos diferentes podem ocorrer através de pequenas mudanças neste parâmetro (12). Para a obtenção dos *clusteres*, foram realizadas simulações com os valores seis e doze para a variável *k*, a fim de obter agrupamentos com menor quantidade de sequências constituintes.

RESULTADOS E DISCUSSÃO

Em ambas as simulações (com seis e doze *clusteres*), apesar de haver divergência em relação às sequências constituintes em cada grupo, os padrões biológicos foram similares. Assim, optou-se por discutir apenas a simulação que formou seis *clusteres*, uma vez que a menor quantidade destes facilita a discussão dos resultados. A quantidade de sequências promotoras constituintes de cada grupo é apresentada na Tabela 2. Dentre estes *clusteres*, observa-se dois com maior quantidade de sequências (*clusteres* 3 e 5), ambos com mais de 30 sequências componentes. Independentemente da quantidade de sequências apresentada em cada agrupamento, não foi possível definir uma especificidade biológica nas funções apresentadas pelos genes reconhecidos pelos promotores de cada *cluster*.

Tabela 2. Conteúdo dos *clusteres* gerados com valor de *k* = 6.

Cluster	**Quantidade de Sequências**	**Porcentagem em relação ao total**
1	9	6,25%
2	18	12,5%
3	36	25,0%
4	18	12,5%
5	40	27,8%
6	23	15,95%

Em relação à composição de nucleotídeos (Figura 2), os *clusteres* 1, 3 e 5 apresentaram o conteúdo similar ao consenso biológico, enquanto que os demais apresentaram um conteúdo divergente a este padrão. Uma característica comum a todas as sequências, de todos os *clusteres*, é a posição do sinal biológico encontrado para a posição-35. Independente da similaridade ao padrão biológico, o sinal encontra-se próximo à posição -30. Já para o sinal da região -10, apenas as sequências dos *clusteres* 5 e 6 apresentam este sinal centrado no nucleotídeo -8. Assim, percebe-se que as sequências promotoras podem apresentar um grau de divergência em seu conteúdo, mas ainda permanecerem com sua função biológica. A semelhança com o padrão biológico pré-estabelecido na literatura é maior na região -10, o que pode ser explicado pelo papel desta região na transcrição gênica, uma vez que ela estabiliza a ligação da RNAp ao promotor. Além disso, a região -35 não se apresenta em aproximadamente 20% de promotores bacterianos conhecidos (13). Considerando que os sistemas de controle genético apresentam múltiplas sequências componentes, com distâncias variadas, a ligação da RNAp pode ser afetada pelo conteúdo da sequência. Assim, analisar a composição e a localização de nucleotídeos relacionados com o processo de transcrição gênica contribui para trabalhos experimentais relacionados com a regulação da expressão gênica, já que o conteúdo interfere na conformação espacial de curvatura, maleabilidade e estabilidade da sequência de DNA (14). Além disso, auxilia em projetos de anotação genômica e outros de biologia teórica e computacional.

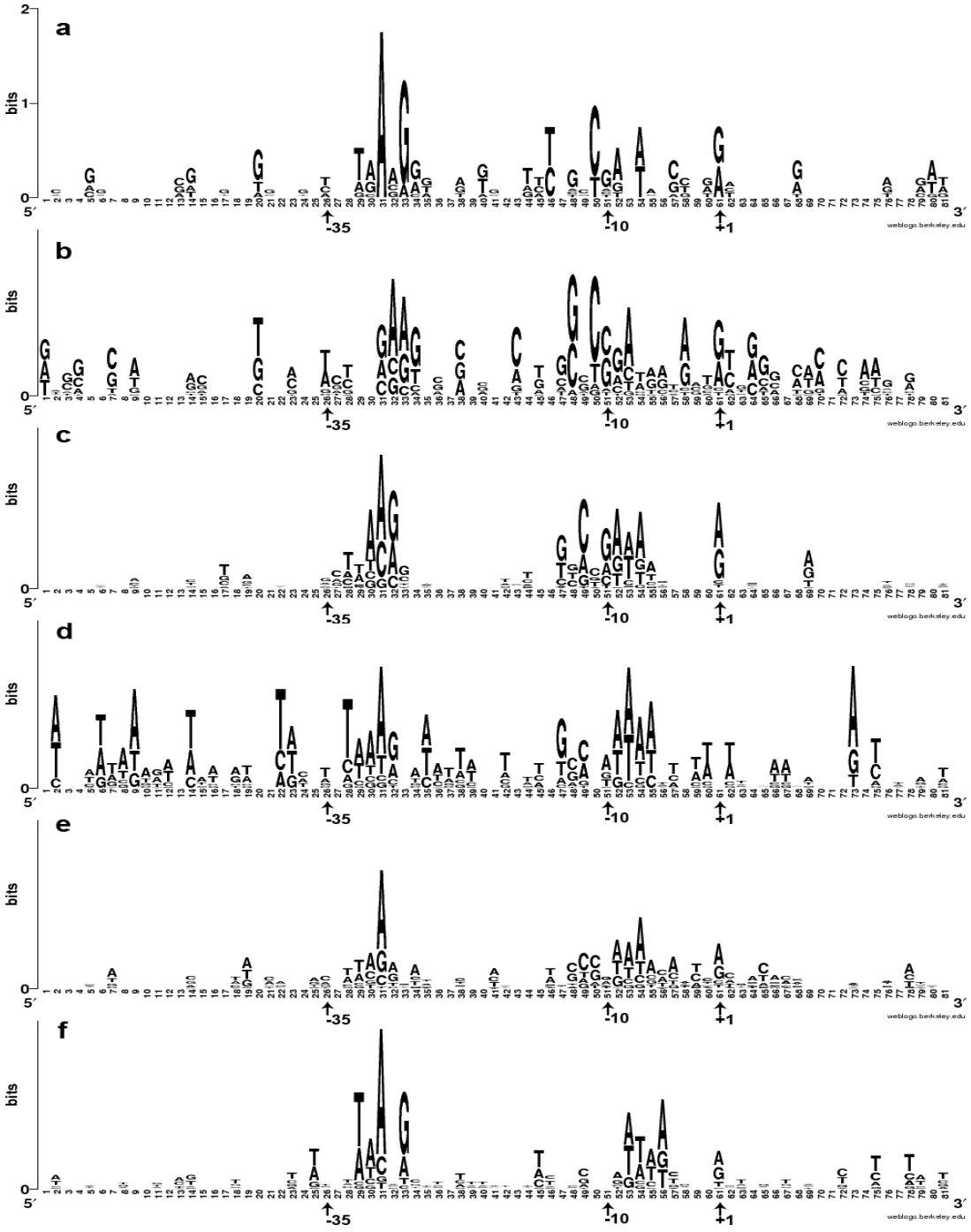

Figura 2. Frequência de nucleotídeos na ferramenta Weblogo para: (a) *Cluster* 1; (b) *Cluster* 2; (c) *Cluster* 3; (d)*Cluster* 4; (e) *Cluster* 5; (f) *Cluster* 6. As posições 26, 51 e 61 referem-se, respectivamente, às regiões -35, -10 e +1.

Após a análise de conteúdo, procurou-se identificar o gene associado a cada sequência promotora constituinte de um cluster. Assim, foi possível verificar a relação proteína-proteína apresentada pelos genes relacionados (Figura 3).

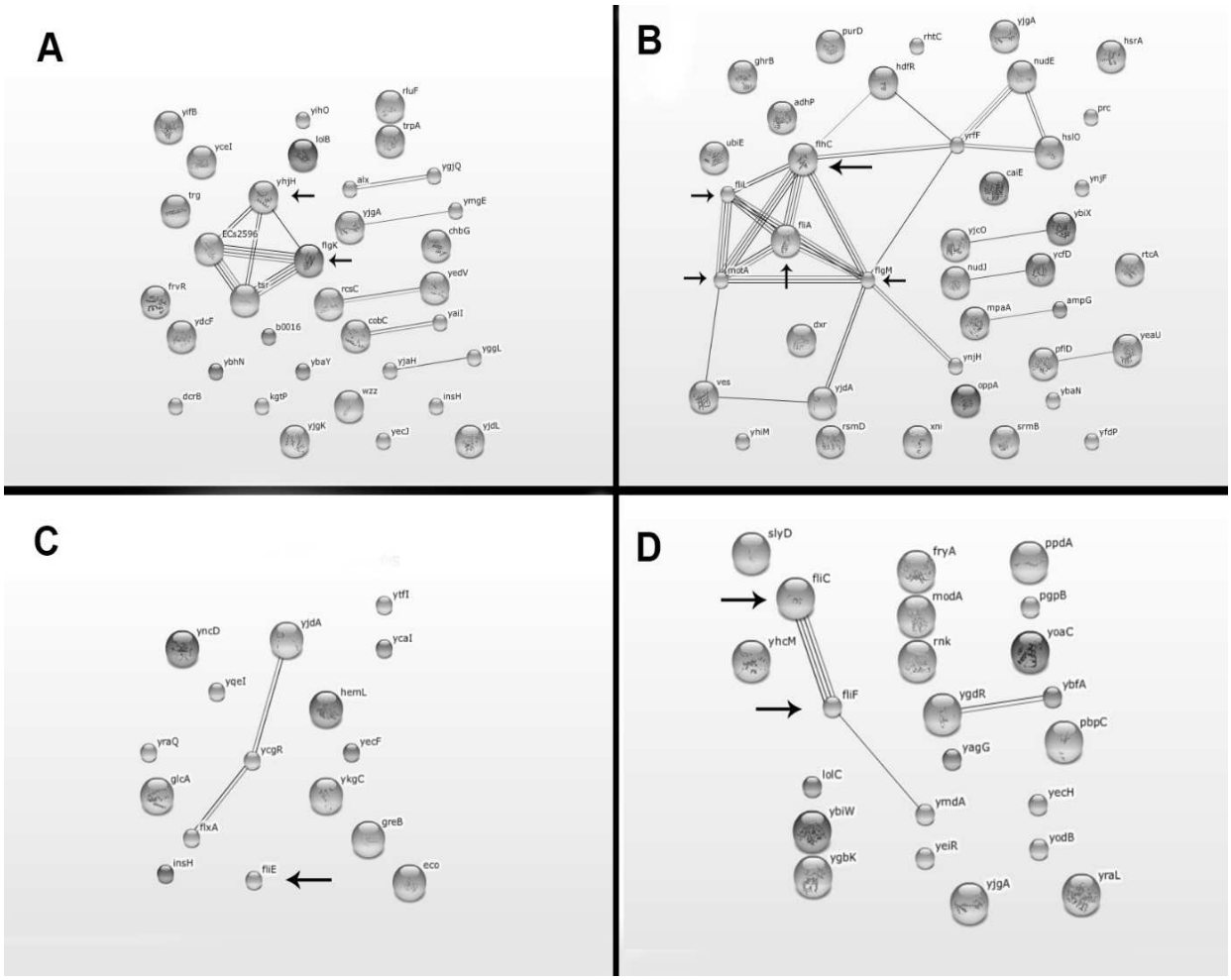

Figura 3. Resultados da ferramenta String-db para: (a) *Cluster* 3; (b) *Cluster* 5; (c) *Cluster* 4; (d) *Cluster* 6. Os genes de maior importância e com maior número de conexões estão marcados pelas setas.

Os agrupamentos que mostraram proteínas com mais de três interações estavam presentes nos *clusteres* 3 e 5. A análise do *cluster* 3 (Figura 3A) apresentou relação de coexpressão e co-ocorrência compatíveis com as funções biológicas de quimiotaxia, motilidade e virulência. As quatro proteínas que estabeleceram três conexões entre si foram: *(i)* ECs2596 (quimiotaxia); *(ii)* tsr (quimiotaxia); *(iii)* flgK (produção de flagelo); *(iv)* yhjH (virulência). Conforme TAMAYO et al. (2009), infra regulação da proteína yhjH (a qual regula a quantidade de diguanilato cíclico) inibe parte da capacidade móvel de *E. coli.* Por outro lado, a supra regulação desta proteína ocasiona o processo inverso. Deste modo, há uma interação entre patógeno-hospedeiro, visto que a motilidade é fundamental para os primeiros processos de colonização no organismo infectado (15).O Cluster 5 (Figura 3B) apresentou relações de vizinhança, co-ocorrência e coexpressão entre as proteínas relacionadas à motilidade flagelar (FliA, FliL, MotA, FlhC e FlgM), fatores de transcrição (FlhC, HdfR) e mecanismos de defesa (HslO). No total, foram obtidas 23 conexões, das quais 19 apresentaram ligação com mais de duas proteínas. As proteínas FlhC e FlgM foram as que apresentaram maior quantidade de conexões (6 conexões). Este resultado pode ser explicado pelo papel biológico apresentado por estas proteínas, uma vez que estas regulam a atividade da proteína FliA (4).

A Figura 3 mostra que as proteínas relacionadas à motilidade flagelar não foram agrupadas todas em um único *cluster*. Por exemplo, as proteínas relacionadas à

motilidade, FliC e FliF, foram agrupadas no *cluster* 6 (Figura 3D). Estas proteínas, além da conexão entre si, apresentaram uma conexão com a proteína YmdA (proteína predita, relacionada à expressão flagelar). Já a proteína FliE aparece no *cluster* 4 (Figura 3C), sem nenhuma conexão com os demais componentes de seu agrupamento. Assim, pode-se inferir que a similaridade na composição da sequência promotora, não implica, necessariamente, em uma única função biológica.

Segundo a literatura, o fator σ^{28} FliA e o fator antissigma FlgM apresentam papel importante na expressão flagelar. Cerca de 25 proteínas constituem o flagelo bacteriano (16). Considerando a relevância destas proteínas no contexto metabólico da produção de cílios e flagelos, procurou-se visualizar a rede de interação destas proteínas (4). Assim, a figura 4 apresenta as relações de algumas proteínas de *E. coli* com a proteína FliA, a qual mostra que genes reconhecidos por outros fatores σ traduzem proteínas que interagem com a FliA durante o processo de formação de cílios e flagelos. Assim, percebe-se a complexidade das interações relacionadas à interação patógeno-hospedeiro (17).

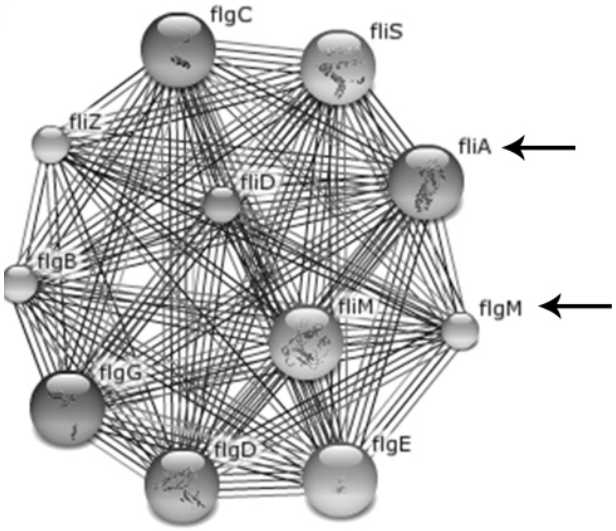

Figura 4. Rede de conexões a partir da proteína FliA.

CONCLUSÃO

A compreensão dos mecanismos de expressão e regulação dos genes inclui múltiplas variáveis, visto que os seres vivos interagem continuadamente com o meio que os cercam. Assim, a análise *in silico* dos elementos regulatórios, bem como de seus genes e produtos biológicos associados, contribui para o levantamento de novas relações e inferências sobre a maquinaria vital das células. Este trabalho apresentou uma abordagem de análise de clusterização dos promotores reconhecidos pelo fator σ^{28} e a relação entre as proteínas traduzidas pelos genes associados a cada promotor. Assim, pode-se perceber que os promotores possuem sinal biológico evidente, no entanto este é divergente ao consenso biológico canônico apresentado na literatura. Além disso, percebeu-se que similaridades na composição da sequência não implicam em uma mesma função biológica. Por exemplo, as proteínas FliC e FliF apresentam funções diferentes: FliC refere-se à uma subunidade proteica enquanto FliF à um componente do corpo basal do flagelo (4). Considerando os resultados obtidos neste trabalho, pretende-se estender a análise de clusterização para os demais promotores, reconhecidos por outros fatores, ampliando a análise de

nucleotídeos para uma análise de valores estruturais, como estabilidade e curvatura.

AGRADECIMENTOS

Agradecemos a Universidade de Caxias do Sul pelo apoio ao projeto e a disponibilidade da bolsa de iniciação científica.

REFERÊNCIAS

(1) ATTWOOD, T. K. et al. Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective. In: MAHDAVI, M. A. (Ed.) **Trends and Methodologies**. Croácia: Intech, 2011. p. 1-2.

(2) KREBS, J.; GOLDSTEIN, S.; KILPATRICK, S. T. **Genes XI.** 11. ed. Sudbury, Massachusetts: Jones and Bartlett, 2014. 930 p.

(3) AVILA e SILVA, S.; ECHEVERRIGARAY, S. Bacterial Promoter Features Description and Their Application on *E. coliin silico* Prediction and Recognition Approaches. In: PÉREZ-SÁNCHEZ, H. (Ed.). **Bioinformatics**, Croacia: Intech, 2012. Cap. 10. p. 241-260.

(4) KAZMIERCZAK, M. J. et al. Alternative Sigma Factors and Their Roles in Bacterial Virulence. **Microbiology and Molecular Biology Reviews**. v. 69, n. 4, p. 527 - 543, 2005.

(5) MAES, A. et al. Role of polyadenylation in regulation of the flagella cascade and motility in *Escherichia coli*. **Biochimie**, v. 95, n. 2, p.410-418, fev. 2013.

(6) FONTANA, E. A. **Algoritmos de Clusterização Aplicados na Análise Genômica da Bactéria *Escherichia coli***. 2013. 75f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação). Universidade de Caxias do Sul, Caxias do Sul, 2013.

(7) CROOKS, G. E. et al. WebLogo: A Sequence Logo Generator. **Genome Research,** v. 14, n.6, p.1188-1190, 12 maio 2004.

(8) FRANCESCHINI, A. et al. STRING v 9.1: protein-protein interaction networks, with increased coverage and integration. **Nucleic Acids Research,** v. 41, n. 1, p.D808-D815, 2012.

(9) SHBURNER, M.; et al. Gene Ontology: tool for the unification of biology. **Nature Genetics,** v. 25, n. 1, p.25-29, 2000.

(10) SALGADO, H. et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. **Nucleic Acids Research,** v. 41, n. 1, p.D203-D213, 2012.

(11) HE, J. Improving feature representation of natural language gene functional annotations using automatic term expansion. In: IEEE SYMPOSIUM ON COMPUTATIONAL INTELLIGENCE IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, Sun Valley: IEEE, 2008, p.173-179.

(12) BANDYOPADHYAY, S.; WANG, J. T.; MAULIK, U. **Computational Intelligence and Pattern Analysis in Biology Informatics.** Nova Jersey: John Wiley & Sons, Inc, 2010. 372 p.

(13) HUERTA, A. M.; COLLADO-VIDES, J. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. **Journal of Molecular Biology,** v. 333, n. 2, p.261-278, out. 2003.

(14) GOÑI, J. R. et al. Determining promoter location based on DNA structure first-principles calculations. **Genome Biology,** v. 8, n. 12, p. R263, 2007.

(15) TAMAYO, R.; PRATT, J. T.; CAMILLI, A. Roles of Cyclic Diguanylate in the Regulation of Bacterial Pathogenesis. **Annual Review of Microbiology,** v. 61, n. 1, p.131-148, out. 2007.

(16) BROWN, P. N. et al. Mutational Analysis of the Flagellar Protein FliG: Sites of Interaction with FliM and Implications for Organization of the Switch Complex. **Journal of Bacteriology,** v. 189, n. 2, p.305-312, 3 nov. 2006.

(17) LI, C. et al. Differential Regulation of the Multiple Flagellins in Spirochetes. **Journal of bacteriology,** v. 192, n. 10, p.2596-2603, 19 mar. 2010.

Enviado: 13/08/2015
Revisado: 21/10/2015
Aceito: 06/11/2015