Análise de erros de anotação de genes exclusivos

Priscilla Koch Wagner[1], **Guilherme Duarte Mattos**[1], **Luciano Antonio Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)
Caixa Postal 02125 – São Paulo – SP – Brazil

Abstract. *This paper presents an approach to identify annotation errors in exclusive genes. This study shows computational tools that can be used by researchers and professionals in this field because of its generic purposes and provide more reliability in a phylogenetic analysis aiming exclusive genes and their origins. The developed tools were used in a study case with genes from Xanthomonas species.*

Resumo. *Este artigo apresenta uma abordagem para a identificação de erros de anotação em genes exclusivos. O estudo apresenta ferramentas computacionais que possuem caráter genérico e podem ser utilizadas por pesquisadores e profissionais da área, bem como possibilitam uma maior confiabilidade em uma análise filogenética visando a identificação de genes exclusivos e suas origens. É apresentado um estudo de caso com genes de espécies de bactérias do gênero Xanthomonas, em que as ferramentas foram aplicadas.*

1. Introdução

Para o estudo da relação evolutiva entre os organismos, denominado filogenia, existem diversas ferramentas computacionais que auxiliam nas etapas de um processo de análise filogenética [Miyamoto and Cracraft 1991]. Entre essas etapas estão: a anotação, comparação e alinhamento de sequências de nucleotídeos [Setubal and Meidanis 1997].

Uma das motivações de se realizar uma análise filogenética é identificar a origem de um determinado gene ou organismo [Matioli 2001]. As peculiaridades dos seres vivos podem ser dadas por genes exclusivos, que são identificados através da análise filogenética de grupos de espécies (ou sub-espécies) próximas geneticamente. Por meio do estudo filogenético é possível conhecer a origem de um gene exclusivo de dado genoma, trazendo avanço sobre os conhecimentos acerca do processo evolutivo da espécie em que esse gene exclusivo se apresenta. Esse estudo pode também trazer mais conhecimentos sobre comportamentos e características da espécie sob análise, visto que esse gene pode ser determinante para um fenótipo também exclusivo.

Durante uma análise filogenética, é necessário que haja uma verificação da anotação dos genes a fim de certificar-se se os genes sendo estudados tiveram o processo de anotação de sua função genética realizado de maneira correta e, caso o objetivo seja identificar genes exclusivos, esse processo é ainda mais delicado pois erros podem se propagar durante a análise e afetar diretamente os resultados.

O presente artigo visa a apresentar uma estratégia para identificar os erros de anotação identificados em genes considerados exclusivos a um genoma. O artigo também busca destacar uma parte do trabalho que ainda será apresentado a comunidade científica como uma nova abordagem para identificar a provável origem de genes exclusivos.

2. Solução Proposta

Com as motivações citadas anteriormente, a solução proposta neste artigo consiste em aplicar uma abordagem para identificação de prováveis erros de anotação em genes potencialmente exclusivos, criando-se ferramentas parametrizáveis ao longo do processo.

A solução proposta para a identificação dos genes realmente exclusivos tem suas principais atividades realizadas por duas ferramentas, sendo que ambas possuem resultados complementares e fazem parte de processos distintos dentro dessa abordagem, conforme apresentado na Figura 1. Tal abordagem pode contemplar estudos para identificação de genes exclusivos de maneira genérica bem como auxiliar na identificação de erros de anotação (exclusivos ou não).

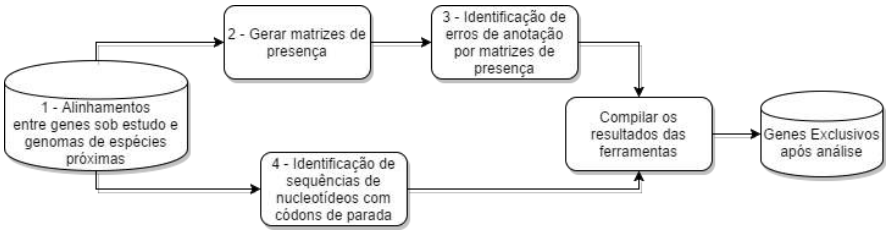

Figura 1. Abordagem adotada para a identificação de genes exclusivos

Para as etapas prévias da análise filogenética, foram utilizadas ferramentas já existentes desenvolvidas pelo grupo de pesquisa em que o trabalho se insere [Pereira et al. 2014, Pereira 2014] (número 2 na Figura 1) para geração de matrizes de presença; bem como o uso da ferramenta BLAST [Altschul et al. 1990] (número 1 na Figura 1) para realizar o alinhamentos dos genes escolhidos para análise com os genomas do *National Center for Biotechnology Information gene and genome database* (NCBI)[1]

As ferramentas desenvolvidas neste projeto referem-se aos números 3 e 4 da Figura 1. A primeira ferramenta (número 3 da Figura 1) possui o objetivo de analisar os potenciais genes exclusivos e identificar possíveis erros de anotação tendo como entrada matrizes geradas a partir de comparações entre genes anotados com os genomas filogeneticamente próximos. Além de identificar potenciais erros de anotação, a ferramenta também tem o objetivo de segregar os dados para identificar se o erro de anotação é relativo a outras sequências dentro do próprio genoma ou encontrados em outros genomas. A segunda ferramenta (número 4 da Figura 1) complementa os resultados da primeira para a análise adotada, submete os dados a uma dupla checagem, identificando quais das sequências de nucleotídeos poderiam sintetizar uma proteína, ou seja, se efetivamente a sequência encontrada pode corresponder a um gene.

3. Experimentos e Resultados

As ferramentas desenvolvidas possuem cunho genérico, portanto se aplicam para quaisquer genes. Entretanto, com intuito de testar sua aplicabilidade, essas ferramentas foram aplicadas num estudo de caso com genomas de bactérias do gênero *Xanthomonas*, que são bactérias patogênicas, visto que seus genes têm sido amplamente estudados e possuem grande importância econômica na agricultura [Zhang et al. 2015]. Foram selecionados 15 genomas de *Xanthomonas*, apresentados na Tabela 1.

[1]http://www.ncbi.nlm.nih.gov/

Tabela 1. Genomas do gênero *Xanthomonas* utilizados

Índice	Nome da espécie	Índice	Nome da espécie
1	*Xanthomonas oryzae pv. oryzae MAFF 311018*	9	*Xanthomonas axonopodis pv. citri str. 306*
2	*Xanthomonas campestris pv. campestris str. ATCC 33913*	10	*Xanthomonas campestris pv. raphani 756C*
3	*Xanthomonas axonopodis Xac29-1*	11	*Xanthomonas campestris pv. campestris str. B100*
4	*Xanthomonas campestris pv. campestris str. 8004*	12	*Xanthomonas oryzae pv. oryzae KACC 10331*
5	*Xanthomonas fuscans subsp. fuscans str. 4834-R*	13	*Xanthomonas oryzae pv. oryzicola BLS256*
6	*Xanthomonas campestris pv. vesicatoria str. 85-10*	14	*Xanthomonas oryzae pv. oryzae PXO99A*
7	*Xanthomonas axonopodis pv. citrumelo F1*	15	*Xanthomonas albilineans GPE PC73*
8	*Xanthomonas citri subsp. citri Aw12879*		

Os genes dos 15 genomas foram comparados uns com os outros por meio do alinhamento de sequências e com base em limiares de identidade e cobertura foram geradas matrizes de presença dos genes em cada um dos genomas. Com base neste resultado, foram criadas famílias de genes homólogos. Foram criadas duas matrizes, uma contendo os resultados dos genes contra os genes e outra contendo os resultados dos genes contra os genomas dos 15 organismos. A partir das matrizes, foi aplicada a primeira ferramenta, que identificou como resultado que 28.996 genes (entre os 65.120 genes desses genomas) estavam envolvidos em erros de anotação intergenoma. Ao analisar os erros dentro do próprio genoma, foram destacados 5.142 genes com essa inconsistência. Ao comparar os genes que eram considerados exclusivos antes da análise da anotação e os genes que permaneceram sendo considerados como exclusivos após a análise, temos a distribuição por genomas (Figura 2 (a)). Nessa distribuição, destacam-se as espécies *Xanthomonas abilineans GPE PC73* (índice 15 na Tabela 1) e *Xanthomonas axonopodis pv. citri str. 306* (índice 9 na Tabela 1) que preservaram quase todos os genes como exclusivos.

Com esses resultados, segregou-se os genes com potenciais erros de anotação para a continuidade da análise. Com o arquivo gerado pelo BLAST (contendo o alinhamento desses genes contra os 15 genomas), aplicou-se a segunda ferramenta a fim de se obter as sequências de nucleotídeos que foram utilizadas na segunda matriz de entrada da primeira ferramenta (neste caso, focando-se apenas nos erros que, a princípio, eram considerados genes exclusivos). O objetivo é verificar se essas sequências podem efetivamente corresponder a genes corroborando com o resultado anterior (da primeira ferramenta) que indica que há um potencial erro de anotação.

Como parametrização da ferramenta foi utilizada uma taxa de identidade de 96%, bem como uma cobertura mínima entre o alinhamento do gene do próprio genoma e o melhor alinhamento do gene com outro genoma também de 96%. Estas porcentagens foram utilizadas porque, para este conjunto de genomas, maximizam a transitividade das relações de homologia. Com a aplicação da segunda ferramenta sobre os dados, foram obtidos 243 sequências de nucleotídeos com a presença de um códon de parada dentro dos seis quadros de leitura (ou seja, elas não correspondem a genes).

Na Figura 2 (b) são identificadas as quantidades de sequências de nucleotídeos (que alinharam com os genes anotados) e que possuem um códon de parada nos seis quadros de leitura, agrupadas por genoma. A Tabela 1 mostra os índices dos genomas que fazem parte do eixo horizontal da Figura 2 (b). Com tais resultados, os genes, que foram alinhados com estas 243 sequências de nucleotídeos, listados como saída da segunda ferramenta podem continuar sendo considerados como genes exclusivos. Como resultado final, foi criada uma lista com os genes identificados como realmente exclusivos após as

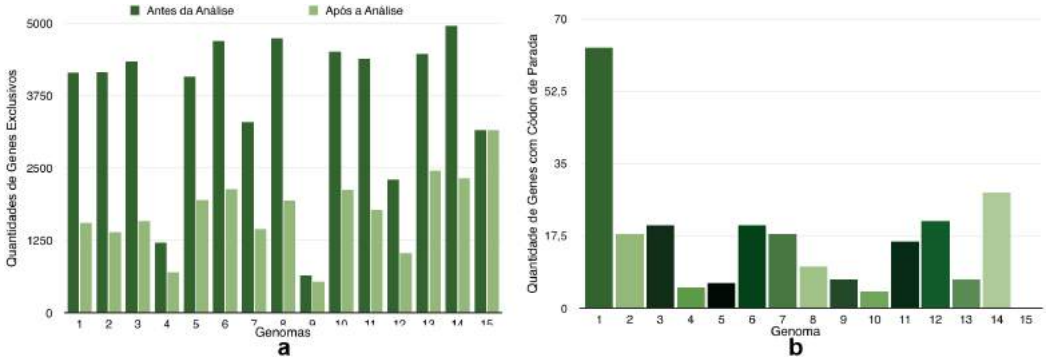

Figura 2. (a) Genes exclusivos antes e após análise de nucleotídeos (b) Distribuição das sequências com códons de parada por genomas

verificações das ferramentas aplicadas (desconsiderando as sequências que tinham características de "genes homólogos" mas que não podem codificar proteínas).

4. Conclusão

O presente artigo apresentou ferramentas computacionais para uma abordagem específica na identificação de genes exclusivos, a fim de trazer maior qualidade aos resultados de uma análise filogenética. As ferramentas possuem cunho genérico e podem ser utilizadas por pesquisadores e profissionais da área no processo de análise filogenética.

Como trabalhos futuros, planeja-se realizar: a investigação da intersecção dos genes com erros de anotação com os genes que se confirmaram exclusivos pela segunda ferramenta; a integração das ferramentas citadas; o desenvolvimento de uma interface amigável para o usuário; o desenvolvimento de ferramentas complementares a fim de se identificar a origem biológica do gene exclusivo filtrado pelas ferramentas apresentadas.

Referências

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Matioli, S. R. (2001). *Biologia Molecular e Evolução*. Editora Holos, Ribeirão Preto.

Miyamoto, M. and Cracraft, J. (1991). *Phylogenetic Analysis of DNA Sequences*. Oxford University Press.

Pereira, V. M. Y. (2014). Montagem e análise de genomas a partir de metagenomas. Master's thesis, Universidade de São Paulo - Escola de Artes, Ciências e Humanidades.

Pereira, V. M. Y., Costa, C. I., and Digiampietri, L. A. (2014). Uma ferramenta baseada em algoritmos genéticos para a ordenação de montagens parciais de genomas. In *VIII Brazilian e-Science Workshop (BRESCI2014)*.

Setubal, J. C. and Meidanis, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing Company, Boston, EUA.

Zhang, Y., Jalan, N., Zhou, X., Goss, E., Jones, J. B., Setubal, J. C., Deng, X., , and Wang, N. (2015). Positive selection is the main driving force for evolution of citrus canker-causing xanthomonas. In *ISME Journal*.