GTACG: Um arcabouço computacional focado em genômica comparativa de bactérias de um mesmo ramo evolutivo

Caio Rafael do Nascimento Santiago[1], **Luciano Antonio Digiampietri**[2], **Leandro Marcio Moreira**[3]

[1]Programa Interunidades de Pós-graduação em Bioinformática –
Universidade de São Paulo (USP)
CEP 05508-090 – São Paulo – SP – Brasil

[2]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP – Brasil

[3]Departamento de Ciências Biológicas
Universidade Federal de Ouro Preto (UFOP) – Ouro Preto, MG – Brasil

{caio.santiago,digiampietri}@usp.br, lmmorei@gmail.com

Abstract. *This work presents GTACG: a framework for comparative genomics with a focus on facilitating usability and aimed to identify unique genetic characteristics in subgroups of bacterial genomes that share a certain phenotype. The results were validated with two case studies involving a set of 161 genomes of the* Xanthomonadaceae *family, in which 19 orthologous protein families unique to the plant-associated genomes were found, and 45 genomes of* Streptococcus pyogenes*, in which 15 families of orthologous proteins were found that serve as phylogenetic markers for the* emm *protein.*

Resumo. *Neste trabalho é apresentado o GTACG: um arcabouço computacional para genômica comparativa com foco em uma usabilidade facilitada e destinada a pesquisas para identificação de características genéticas únicas em subgrupos de genomas de bactérias que possuem um determinado fenótipo em comum. Os resultados foram validados com dois estudos de caso envolvendo um conjunto com 161 genomas da família* Xanthomonadaceae*, no qual foram encontradas 19 famílias proteicas ortólogas exclusivas aos genomas associados a plantas e 45 genomas de* Streptococcus pyogenes*, no qual foram encontrados 15 famílias de proteínas ortólogas que servem como marcadores filogenéticos para a proteína* emm*.*

1. Introdução

Os estudos genéticos datam das pesquisas realizadas no início do século XX [Fietto and Lamëgo 2015], e com o passar do tempo os métodos para se extrair e sequenciar DNA foram aperfeiçoados e popularizados dentro da comunidade científica, causando assim uma proliferação de genomas ou partes de genomas sequenciados [Bell et al. 2009]. A análise manual de um volume tão grande de dados é inviável, porém a massificação do sequenciamento de genomas abre novas possibilidades para análises comparativas, mais especificamente de análises genômicas sobre populações [Joyce et al. 2002].

A massiva quantidade de genomas sequenciados disponível nas bases de dados fornece maior grau de confiança para estudos populacionais, como a filogenia [Felsenstein 1988] e o estudo sobre fenótipos. Organismos de mesmo ramo evolutivo tendem a compartilhar determinados fenótipos, da mesma forma que compartilham determinados genes homólogos que são direta ou indiretamente responsáveis pela expressão desses fenótipos, pois a função desses genes é preservada através das gerações [Hardison 2003, Xia 2013]. Em geral, muitos estudos utilizam a homologia para inferir a função de proteínas desconhecidas, ou para estudar o comportamento dos organismos [Chervitz et al. 2011]. Porém, não são abundantes os estudos que estudam populações correlatas a fim de entender os mecanismos genéticos por trás de determinados fenótipos [Ilina et al. 2013, Obolski et al. 2018]. Mesmo existindo alguns estudos sobre o estudo populacional de fenótipos, ainda há uma carência de metodologias e ferramentas para automatizar a análise.

As bactérias são organismo interessantes para o estudo de populações devido a suas características genômicas. Bactérias possuem genomas relativamente pequenos (poucos milhões de pares de bases) formados majoritariamente de sequências codificantes (CDS). Isto permite análises detalhadas sobre pequenas diferenças fenotípicas que podem estar diretamente relacionadas aos genes.

Nesse artigo é apresentado o GTACG (*Gene Tags Assessment by Comparative Genomics*), um arcabouço computacional dedicado ao estudo comparativo de genomas de bactérias, que contém um *pipeline* completo para estudos de genômica comparativa para análise de características fenotípicas.

2. O Arcabouço

O ambiente como um todo do GTACG pode ser dividido entre *back-end* e *front-end*. A divisão se faz necessária, pois no back-end estão as ferramentas e algoritmos destinados à preparação dos dados genômicos fornecidos pelo usuário e é necessário um conhecimento básico em computação para executar as etapas de seu *pipeline*. Por outro lado, o front-end é destinado à visualização dos resultados e não exige conhecimentos específicos em computação. Os resultados são exportados na forma de um *website*, não sendo necessária a configuração de um servidor e facilitando a publicação e compartilhamento dos resultados por parte do pesquisador [Santiago et al. 2019].

A partir de informações genômicas (a sequência de DNA do genoma e a identificação das sequências codificantes) e fenotípicas o arcabouço agrupa as CDS em diferentes níveis de homologia utilizando um algoritmo próprio [Santiago et al. 2018] e estes agrupamentos são utilizados como base para as demais análises.

Os resultados abrangem uma ampla gama de possibilidade, para permitir ao usuário estruturar uma pesquisa de forma completa. Entre os resultados para comparação de genomas dois se sobressaem. O primeiro é apresentado na forma de filogenias, usando métodos bem estabelecidos como *neighbor-join* combinado com *k-mers*, mas se destacam os métodos mais robustos como a *supertree*, construída utilizando como base as famílias de genes compartilhados. O segundo é a geração de métricas para identificar genes correlatos com características fenotípicas separadas em três diferentes categorias. A primeira dessas categorias é a conformação das famílias, definida por famílias (individualmente ou em combinação) únicas ou majoritárias a um grupo específico de genomas. Nesta

categoria são apresentadas métricas para indicar quantas CDS ou genomas presentes na família pertencem aos genomas dos grupos de interesse. A informação é disponibilizada na forma absoluta e percentual, indicando deste modo o quão representativo este grupo é para a família em questão. A segunda categoria apresentada é referente aos alinhamentos das famílias, identificando de forma relativa quantas bases são mais correlatas a determinado grupo, e para expressar isso de forma numérica foi criada uma métrica de dissimilaridade. A última categoria é sobre as filogenias inferidas a partir dos alinhamentos múltiplos das sequências de cada família, com o objetivo de determinar o quão bem estão separados os genomas de determinado grupo em relação aos outros. Para isso foi criada a métrica *Most Isolated SubTree* (MIST) que mostra de forma numérica qual o tamanho da maior sub-árvore formada apenas por sequências do grupo em estudo.

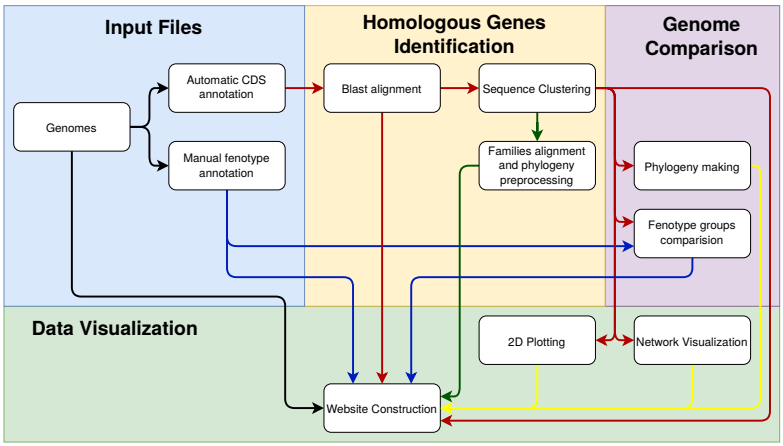

Figura 1. *Pipeline* de execução do GTACG organizadas em três pilares de processamento: identificação de genes homólogos, comparação de genomas e visualização de resultados [Santiago et al. 2019].

3. Resultados e Discussões

As ferramentas desenvolvidas estão disponíveis para acesso no *github* (em `https://github.com/caiorns/GTACG-backend`) e podem ser executadas em computadores com sistema operacional Linux. Os resultados visuais foram estruturados no formato de um *website* com diferentes níveis de detalhamento. Na tela inicial do *website* estão contidas as informações mais macroscópicas. Este primeiro nível está relacionado a visualização e interação com dados genômicos, como filogenias, visualização de genomas e dados estatísticos. Esta tela também há a possibilidade de estabelecer filtros que levam o usuário a acessar os níveis seguintes de detalhamento dos dados, em que são listadas as famílias de genes de acordo com métricas estatísticas.

O nível seguinte é destinado ao estudo estatístico das famílias. A famílias de genes podem ser pesquisadas por meio de métricas como: o número de genomas que as compartilham, número de sequências, distribuição do comprimento das sequências, função anotada, métricas baseadas nos grafos, métricas baseadas nos alinhamentos, métricas baseadas nas filogenias e métricas baseadas nas anotações dos grupos de genomas.

O último nível é dedicado ao detalhamento de famílias. Neste nível, cada família tem uma página com suas respectivas informações. Neste nível é possível ver as informações referentes a cada um dos genes, combinada com as informações de seus

respectivos genomas (identificação do genoma e a anotação dos grupos). Para cada sequência também está presente um *link* de acesso a uma busca do BLAST contra a base de aminoácidos do NCBI. Também é possível visualizar o alinhamento, filogenia da famílias e informações de alinhamentos locais entre cada uma das sequências.

4. Estudo de caso

Com o objetivo de validar o arcabouço e sua potencialidade de auxiliar nos estudos fenótipos de genomas dois estudos de caso foram realizados, porém, por limitação de espaço, apenas um será apresentado. 161 genomas provenientes da família *Xanthomonadaceae* foram analisados com o objetivo principal de identificar possíveis genes relacionados à associação adaptativa com plantas, quer como fitopatógenos ou não. Dos 169 genomas, 139 genomas apresentam essa característica (os pertencentes aos gêneros *Xanthomonas* e *Xylella*), por outro lado, os 22 genomas dos gêneros *Pseudoxanthomonas* e *Stenotrophomonas* não apresentam essa característica.

Nenhuma família ortóloga encontrada consegue, individualmente, separar os dois grupos, isto é, ser compartilhada por todos os genomas associados a plantas, e ao mesmo tempo não estar presente nos demais. Porém foram encontrados resultados interessantes e que são consistentes com a filogenia encontrada pela *supertree*. Foram encontradas 19 famílias de genes compartilhados por ao menos 90% dos genomas associados a plantas e ausentes a todos os demais. Destaca-se que esses genomas ausentes são os mesmos identificados como um grupo separado na filogenia.

Entre as 19 famílias de proteínas identificadas em pelo menos 134 dos 139 genomas dos microrganismos associados a plantas, oito famílias estão envolvidas na degradação de N-glicanos, todas na mesma região genômica, e são responsáveis pela clivagem dos N-glicanos em diferentes ligações glicosídicas (Quadro 2 e Figura 22 encontrados na tese). A interação de patógenos de plantas é propiciada pela evolução das proteínas ligadas à virulência bacteriana para induzir a virulência e modular a resposta imune das plantas, isso concomitante com a evolução das proteínas vegetais para reconhecer os efeitos da infecção bacteriana e induzir resposta imunológica especializada levando à resistência. Os receptores de reconhecimento de padrões são responsáveis por reconhecer padrões moleculares associados a patógenos e pela ativação de gatilhos imunológicos precisam de N-glicosilação para mediar a imunidade da planta [Häweker et al. 2010], e sua degradação impacta negativamente na resposta imune da planta.

Adicionalmente, outras proteínas encontradas estão envolvidas na adaptação, incluindo duas peptidases, três proteínas hipotéticas e um homólogo a XAC0501, este último que é codificado pelas *LesA/LipA* é um fator de virulência chave necessário para a patogenicidade de *Xylella fastidiosa* em videiras [Nascimento et al. 2016]. Outros quatro genes também podem estar relacionados com a adaptação e já foram relatados em outros estudos relacionados a virulência. O *hspA*, o *cyoD* o *leuB* e o *leuB* [Lin et al. 2010, Lunak and Noel 2015, Moreira et al. 2017, Laia et al. 2009].

Esta análise do repertório dos genes identificados pelo GTACG permite inferir que o arcabouço computacional desenvolvido se mostrou eficiente na busca de informações genéticas correlacionadas com informações fenotípicas, uma vez que os genes identificados automaticamente como exclusivos a genomas associados a plantas já foram descritos como capazes de modular a adaptação bacteriana à planta hospedeira.

5. Conclusão

No decorrer deste texto foi apresentado o GTACG (*Gene Tags Assessment by Comparative Genomics*) um arcabouço computacional que contempla todo o ciclo de vida, do ponto de vista computacional, de uma pesquisa sobre genômica comparativa de bactérias. O principal foco que norteou todo o processo de desenvolvimento deste arcabouço foi que, a partir de genomas de um mesmo ramo evolutivo, o pesquisador conseguisse encontrar características genéticas relacionadas com características fenotípicas. Para isso, foi definida uma ampla gama de métricas e ferramentas de buscas que permitem ao pesquisador extrair informações acerca de todo o pan-genoma.

Este arcabouço tem como objetivo a comparação de genomas, por meio de métricas, filogenias e visualizações. Os genomas são processados com base nas famílias de genes e diferentes resultados comparativos são produzidos, incluindo estatísticas básicas e métricas de correlação sobre os grupos fenótipos estabelecidos pelo usuário. São apresentadas várias opções de filogenia, desde abordagens clássicas a métodos mais robustos baseados na famílias de genes, como a *supertree*.

Os resultados foram construídos com base em um *website* estático e interativo. Isto traz diversos benefícios, entre eles a facilidade que um usuário sem conhecimentos profundos sobre computação tem de gerar e compartilhar resultados de uma pesquisa, pois os dados são facilmente publicáveis. A interação dos usuários com os resultados é bastante dinâmica, potencializada pela abordagem *top-down*, em que o usuário a partir de dados genômicos consegue se aprofundar até encontrar informações genéticas que poderiam justificar suas hipóteses iniciais (com base na anotação fenotípica).

Entre os principais resultados produzidos pelo presente trabalho estão: (i) o arcabouço em si, disponibilizado no GitHub (endereço apresentado na seção 3) e que já está sendo utilizado por diferentes grupos de pesquisa; (ii) um artigo publicado no periódico Journal of Computational Biology [Santiago et al. 2018] que apresenta o algoritmo proposto e desenvolvido para o agrupamento de genes em famílias e (iii) um artigo publicado no periódico Frontiers in Genetics [Santiago et al. 2019] que apresenta o arcabouço produzido. A tese pode ser encontrada a partir do seguinte link:
`https://www.teses.usp.br/teses/disponiveis/95/95131/tde-02032020-102628/`

Referências

Bell, G., Hey, T., and Szalay, A. (2009). Beyond the Data Deluge. *Science*.

Chervitz, S. A., Deutsch, E. W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S.-A., Stoeckert, C. J., Taylor, C. F., Taylor, R., and Ball, C. A. (2011). Data Standards for Omics Data: The Basis of Data Sharing and Reuse. pages 31–69.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual review of genetics*, 22(1):521–565.

Fietto, L. G. and Lamëgo, M. R. d. A. (2015). História e importância da genômica. In Moreira, L. M., editor, *Ciências genômicas: fundamentos e aplicações*, pages 21–26. Sociedade Brasileira de Computação.

Hardison, R. C. (2003). Comparative Genomics. *PLoS Biol*, 1(2).

Häweker, H., Rips, S., Koiwa, H., Salomon, S., Saijo, Y., Chinchilla, D., Robatzek, S., and von Schaewen, A. (2010). Pattern Recognition Receptors Require N-Glycosylation to Mediate Plant Immunity. *Journal of Biological Chemistry*, 285(7):4629–4636.

Ilina, E. N., Shitikov, E. A., Ikryannikova, L. N., Alekseev, D. G., Kamashev, D. E., Malakhova, M. V., Parfenova, T. V., Afanas'ev, M. V., Ischenko, D. S., Bazaleev, N. A., Smirnova, T. G., Larionova, E. E., Chernousova, L. N., Beletsky, A. V., Mardanov, A. V., Ravin, N. V., Skryabin, K. G., and Govorun, V. M. (2013). Comparative Genomic Analysis of Mycobacterium tuberculosis Drug Resistant Strains from Russia. *PLoS ONE*, 8(2):e56577.

Joyce, E. A., Chan, K., Salama, N. R., and Falkow, S. (2002). Redefining bacterial populations: a post-genomic reformation. *Nature Reviews Genetics*, 3(6):462–473.

Laia, M. L., Moreira, L. M., Dezajacomo, J., Brigati, J. B., Ferreira, C. B., Ferro, M. I. T., Silva, A. C. R., Ferro, J. A., and Oliveira, J. C. F. (2009). New genes of Xanthomonas citri subsp. citri involved in pathogenesis and adaptation revealed by a transposon-based mutant library. *BMC microbiology*, 9(1):12.

Lin, C.-H., Lee, C.-N., Lin, J.-W., Tsai, W.-J., Wang, S.-W., Weng, S.-F., and Tseng, Y.-H. (2010). Characterization of Xanthomonas campestris pv. campestris heat shock protein A (HspA), which possesses an intrinsic ability to reactivate inactivated proteins. *Applied microbiology and biotechnology*, 88(3):699–709.

Lunak, Z. R. and Noel, K. D. (2015). A quinol oxidase, encoded by cyoABCD, is utilized to adapt to lower O2 concentrations in Rhizobium etli CFN42. *Microbiology*, 161(Pt 1):203.

Moreira, L. M., Soares, M. R., Facincani, A. P., Ferreira, C. B., Ferreira, R. M., Ferro, M. I. T., Gozzo, F. C., Felestrino, B., Assis, R. A. B., Garcia, C. C. M., and others (2017). Proteomics-based identification of differentially abundant proteins reveals adaptation mechanisms of Xanthomonas citri subsp. citri during Citrus sinensis infection. *BMC microbiology*, 17(1):155.

Nascimento, R., Gouran, H., Chakraborty, S., Gillespie, H. W., Almeida-Souza, H. O., Tu, A., Rao, B. J., Feldstein, P. A., Bruening, G., Goulart, L. R., and others (2016). The type II secreted lipase/esterase LesA is a key virulence factor required for Xylella fastidiosa pathogenesis in grapevines. *Scientific reports*, 6:18598.

Obolski, U., Gori, A., Lourenço, J., Thompson, C., Thompson, R., and French, N. (2018). Identifying Streptococcus pneumoniae genes associated with invasive disease using pangenome-based whole genome sequence typing.

Santiago, C., Assis, R. D., Moreira, L. M., and Digiampietri, L. A. (2019). Gene Tags Assessment by Comparative Genomics (GTACG): A user-friendly framework for bacterial comparative genomics. *Frontiers in Genetics*.

Santiago, C., Pereira, V., and Digiampietri, L. (2018). Homology Detection Using Multilayer Maximum Clustering Coefficient. *Journal of Computational Biology*, 25(12):1328–1338.

Xia, X. (2013). *Comparative Genomics*. SpringerBriefs in Genetics. Springer Berlin Heidelberg, Berlin, Heidelberg.