Identificação *in silico* de genes associados ao sistema CRISPR-Cas a partir de dados de metagenômica em organismos adversos

Palavras-Chave: Bioinformática, Ecologia microbiana, Sequenciamento de próxima geração

Barbara da Paixão Perez Rodrigues, Ilum Escola de Ciência, CNPEM

Prof. Dr. Leandro Nascimento Lemos, Ilum Escola de Ciência, CNPEM

Introdução:

Desde o lançamento do mapeamento do genoma humano a partir de esforços de pesquisadores de todo o mundo a bioinformática tem se tornado cada vez mais influente na ciência em pesquisas que estão na fronteira do conhecimento [1]. Este fato deve-se primordialmente, ao avanço das tecnologias de sequenciamento gênico, como as tecnologias *Illumina* baseadas em leituras curtas, que possibilitam a geração de dados genéticos em alta velocidade e confiabilidade [2]. Com isso, há a geração maior a cada dia de dados genéticos que abastecem grandes bancos de dados. A partir disso, uma ampla gama de experimentos e estudos é criada, por exemplo, para estudos de metagenômica das mais diferentes amostras, o que demanda a cada dia de mais profissionais de bioinformática para a sua análise.

A metagenômica consiste no conjunto de informações genéticas de microrganismos existentes em uma comunidade presente em um ambiente [3]. Através de ferramentas de bioinformática, então, é possível se identificar toda a comunidade microbiana de uma determinada localidade sem a exigência de seu cultivo em laboratórios, por meio da montagem de seu genoma. Tendo em vista que o corpo humano é composto por cerca de 90% de células estranhas a ele, o que é chamado de microbioma [4]. Estudos que tentam revelar esses indivíduos e entender as suas funções nos diferentes tecidos que o compõem têm alta relevância visando avanços na medicina ao verificar como essa composição se dá diante de doenças, diferentes costumes alimentares, locais frequentados e outros hábitos cotidianos [5]. Disbioses, ou seja, desequilíbrio desfavorável na diversidade da microbiota, por exemplo a intestinal é um fator que pode levar a muitas doenças, como obesidade, asma, artrite e doenças hepáticas [4].

Repetições Palindrômicas Curtas Agrupadas e Regularmente Interespaçadas (do inglês CRISPR: Clustered regularly interspaced short palindromic repeats) são elementos genéticos presentes nos genomas da maioria das bactérias e arquéias, e que estão envolvidas na resistência microbiana contra bacteriófagos (vírus que infectam bactérias e arquéias) [6]. A descoberta desses elementos genéticos foi realizada a partir da análise da sequência do genoma da bactéria *Streptococcus thermophilus*, onde foi possível encontrar locis que codificam um conjunto de repetições palindrômicas curtas e regularmente espaçadas [6]. O CRISPR combina sequências exclusivas de seu genoma intercalados com fragmentos curtos chamados de 'espaçadores' (do inglês: *spacers*), que demonstram alta similaridade com bacteriófagos ou plasmídeos [7].Associado a uma proteína, denomida de proteína *cas* (Crispr-associated), é formando o sistema CRISPR-Cas, que é responsável pelo reconhecimento de genomas exógenos e a sua clivagem após infecção [8].

A descoberta do sistema CRISPR-cas [9] permitiu uma revolução no campo da engenharia genética, onde foram desenvolvidas novas ferramentas de interesse biotecnológico para edição de genomas de animais, plantas e humanos, visando seu uso para a solução de problemas genéticos atuais

Embora o mecanismo deste sistema tenha sido sistematicamente explorado, os elementos genéticos móveis armazenados neles e a sua ampla gama de genes cas associados em seus diferentes sistemas ainda é pouco explorada [10]. Existem mais de 65 proteínas diferentes compreendidas nos genes associados ao sistema CRISPR, os quais podem ser divididos em muitos subtipos e classes [11], evidenciando o grande potencial de diferenciação desta maquinaria e seu aparente potencial de adaptação as condições ecológicas em que está inserido e pode se associar, como por exemplo em diferentes microbiomas do corpo humano [12].

O desenvolvimento de pesquisas que desejam implementar ferramentas de bioinformática para exploração de dados metagenômicos, no entanto, não é tão simples ao passo que requer o uso de muitas etapas em que diferentes softwares são necessários. Dessa forma, o uso de pipelines computacionais, sequências de etapas que devem ser aplicadas aos dados visando sua análise e processamento de maneira organizada e eficiente, é de suma importância para formar a metodologia que abrange esses trabalhos.

Dado o contexto, este trabalho tem como objetivo propor um pipeline computacional automatizado que visa a identificação de sistemas CRISPR-Cas e seus protoespaçadores associados, bem como o histórico de infecção da comunidade microbiana do microbioma humano intestinal sob diferentes condições, como em disbiose. O processo tem como abrangência a reconstrução dos genomas bacterianos em metagenomas (MAGs) usando dados de sequenciadores Ilumina NovaSeq. Os genomas são reconstruídos, filtrados e identificados utilizando ferramentas que tem como princípio aprendizado de máquina e algoritmos complexos. A ferramenta automatizada será disponibilizada na linguagem de programação python em repositório público visando contribuir com a comunidade científica, bem como a formação de novos recursos no campo de Biologia Computacional e Metagenômica na Ilum – Escola de Ciência, faculdade do Centro Nacional de Pesquisa em Energia e Materiais (CNPEM).

Metodologia:

Para o desenvolvimento da plataforO processamento dos dados é totalmente feito utilizando o super-computador de alto desempenho (HPC) da Ilum – Escola de Ciência.

Para o desenvolvimento desde projeto foi utilizado um conjunto de dados de sequenciamento metagenômico da microbiota intestinal humana disponibilizados para uso público no repositório do Centro Nacional de Informações sobre Biotecnologia em sua partição de Arquivo de Leitura de Sequência. As sequências foram geradas por meio de sequenciadores da plataforma Illumina NovaSeq6000, gerando em média 23 milhões de sequências de leituras curtas (do inglês short reads).

Todo o trabalho é dividido em duas etapas principais, sendo que a primeira etapa (I), tem como foco a reconstrução e curadoria de genomas microbianos a partir de dados metagenômicos de pacientes com o objetivo de selecionar os genomas mais completos e com melhor qualidade capazes de gerar resultados mais precisos nas etapas seguintes. Já a segunda etapa (II) concentra-se na identificação dos loci CRISPR-Cas destes genomas e sua classificação [13], por meio de ferramentas de bioinformática que funcionam a partir de aprendizado de máquina [14].

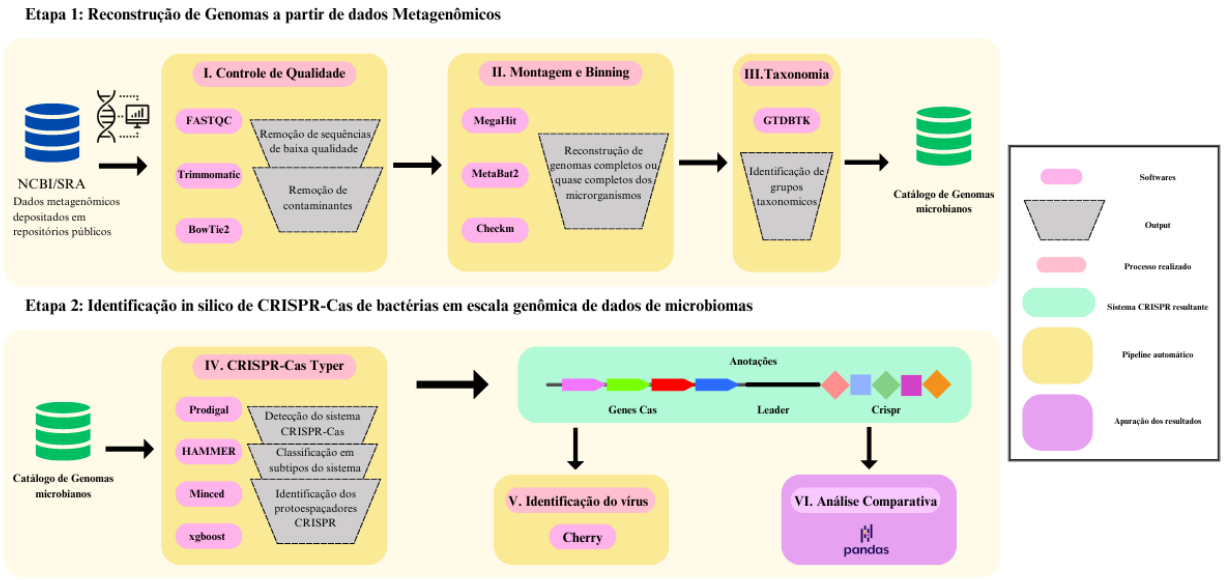

Figura 1. Pipeline proposto para a exploração do sistema CRISPR-Cas existente na microbiota intestinal humana. A etapa 1 ilustra suas três partições, Controle de qualidade (I), Montagem e Binning dos genomas (II) e Taxonomia dos microrganismos (III). Para a etapa 2 são precisos o passo IV de identificação dos sistemas CRISPR-Cas a partir do conjunto de genomas, V de identificação dos vírus presentes nos protoespaçadores do sistema e VI análise comparativa dos resultados obtidos a partir dos dados anteriores.

Etapa I – Reconstrução de Genomas a partir de dados Metagenômicos

Na primeira etapa do pipeline é realizada a montagem e reconstrução de genomas a partir de dados metagenômicos. O primeiro passo (I) Figura 1 consiste na análise do controle de qualidade utilizando três softwares de bioinformática. O processo inicial é feito pelo software FastQC [15], o qual realiza um conjunto de verificações dos arquivos de sequenciamento e produz um relatório que resume todos os resultados, como a pontuação de qualidade por base. Na próxima etapa a ferramenta Trimmomatic [16] realiza cortes em regiões específicas dos dados, visando eliminar qualquer possível contaminante. Por fim, é necessário o uso do software BowTie2 [17] para executar o alinhamento das sequências com o genoma correspondente ao hospedeiro da microbiota para eliminar qualquer resquício que não seja de DNA bacteriano.
Em seguida, a montagem dos genomas é feita com o software Megahit [18] (passo II, Figura 1). Após a montagem, a partir do software Metabat2 [40], foi realizado o processo de binning das contigs. O controle de qualidade dos genomas microbianos é feito pelo software CheckM2 [19]. Por fim, é feita a identificação taxonômica dos microrganismos utilizando o software GTDBTk [20].

Etapa II – Identificação in sílico de CRISPR-Cas de bactérias em escala genômica de dados de microbiomas

A segunda etapa do sistema consiste na busca por loci CRISPR e seus genes Cas associados presentes nos genomas do catálogo de genomas obtidos na etapa anterior. A identificação do sistema e suas classificações em subtipos foi feita utilizando o software CrisprCasTyper [21], que tem como princípio o uso de aprendizado de máquina (Deep Learning). Possibilitando também a visualização graficamente da estrutura dos sistemas CRISPRs. Esse modelo é implementado utilizando a biblioteca xgboost e tem a capacidade de identificar e classificar os diferentes subtipos de repetições do sistema com uma precisão média de 89% [22].

Com o loci CRISPR identificado e seus genes cas associados, o próximo passo consiste em analisar o que foi encontrado. A partir dos espaçadores (do inglês spacers) presentes no sistema, sabendo que eles consistem no material genético de móvel de um invasor que se integrou ao das bactérias, é feita a identificação de sequências correspondentes a vírus que já as infectam. Para isso será utilizada a ferramenta de bioinformática Cherry, a qual tem alta precisão no reconhecimento de

interações vírus-procarionte, com alta precisão e estabilidade até mesmo para contigs curtos [23]. Assim, com todos os passos anteriores finalizados, é construída uma planilha utilizando a biblioteca Pandas para reunir os resultados obtidos para as diferentes amostras. A fim de realizar todas as tarefas complexas que abrangem esta etapa, será feito um código em Python de modo a automatizar todo o processamento utilizando o software Nextflow, o qual será disponibilizado posteriormente em repositório público no GitHub.

Resultados e Discussão:

O projeto encontra-se em andamento e até o momento tem todos os scripts da etapa I do pipeline já definidos e em execução. Assim, a atenção agora está na etapa II e na averiguação dos seus resultados e códigos de execução.

Bibliografia

[1] BAYAT, A. **Science, medicine, and the future: Bioinformatics**. BMJ, v. 324, n. 7344, p. 1018–1022, 2002. Disponível em: <https://www.bmj.com/lookup/doi/10.1136/bmj.324.7344.1018>. Acesso em: 12 jan. 2024.

[2] ALTVEŞ, Safaa; YILDIZ, Hatice Kübra; VURAL, Hasibe Cingilli. Interaction of the microbiota with the human body in health and diseases. **Bioscience of Microbiota, Food and Health**, v. 39, n. 2, p. 23–32, 2020. Disponível em: <https://www.jstage.jst.go.jp/article/bmfh/39/2/39_19-023/_article>. Acesso em: 7 ago. 2024.

[3] CONG, Le; RAN, F. Ann; COX, David; *et al*. Multiplex Genome Engineering Using CRISPR/Cas Systems. **Science**, v. 339, n. 6121, p. 819–823, 2013. Disponível em: <https://www.science.org/doi/10.1126/science.1231143>. Acesso em: 7 ago. 2024.

[4] CRAWLEY, Alexandra B.; HENRIKSEN, Emily D.; STOUT, Emily; *et al*. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. **Scientific Reports**, v. 8, n. 1, p. 11544, 2018. Disponível em: <https://www.nature.com/articles/s41598-018-29746-3>. Acesso em: 7 ago. 2024.

[5] JINEK, Martin; CHYLINSKI, Krzysztof; FONFARA, Ines; *et al*. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. **Science**, v. 337, n. 6096, p. 816–821, 2012. Disponível em: <https://www.science.org/doi/10.1126/science.1225829>. Acesso em: 7 ago. 2024.

[6] MAKAROVA, Kira S.; HAFT, Daniel H.; BARRANGOU, Rodolphe; *et al*. Evolution and classification of the CRISPR–Cas systems. **Nature Reviews Microbiology**, v. 9, n. 6, p. 467–477, 2011. Disponível em: <https://www.nature.com/articles/nrmicro2577>. Acesso em: 7 ago. 2024.

[7] MAKAROVA, Kira S.; WOLF, Yuri I.; IRANZO, Jaime; *et al*. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. **Nature Reviews Microbiology**, v. 18, n. 2, p. 67–83, 2020. Disponível em: <https://www.nature.com/articles/s41579-019-0299-x>. Acesso em: 7 ago. 2024.

[8] MORAES-ALMEIDA, Mariele Santos. **Edição Gênica por CRISPRCas9: da teoria à prática**. São Paulo, SP: Open Access, 2022.

[9] MÜNCH, Philipp C.; FRANZOSA, Eric A.; STECHER, Bärbel; *et al*. Identification of Natural CRISPR Systems and Targets in the Human Microbiome. **Cell Host & Microbe**, v. 29, n. 1, p. 94-106.e4, 2021. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1931312820305734>. Acesso em: 7 ago. 2024.

[10] NAVGIRE, Gauri S.; GOEL, Neha; SAWHNEY, Gifty; *et al*. Analysis and Interpretation of metagenomics data: an approach. **Biological Procedures Online**, v. 24, n. 1, p. 18, 2022. Disponível em: <https://biologicalproceduresonline.biomedcentral.com/articles/10.1186/s12575-022-00179-7>. Acesso em: 7 ago. 2024.

[11] RUSSEL, Jakob; PINILLA-REDONDO, Rafael; MAYO-MUÑOZ, David; *et al*. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. **The CRISPR Journal**, v. 3, n. 6, p. 462–469, 2020. Disponível em: <https://www.liebertpub.com/doi/10.1089/crispr.2020.0059>. Acesso em: 7 ago. 2024.

[12] TREMBLAY, Julien; SCHREIBER, Lars; GREER, Charles W. High-resolution shotgun metagenomics: the more data, the better? **Briefings in Bioinformatics**, v. 23, n. 6, p. bbac443, 2022. Disponível em: <https://academic.oup.com/bib/article/doi/10.1093/bib/bbac443/6780270>. Acesso em: 7 ago. 2024.

[13] WANG, Joy Y.; DOUDNA, Jennifer A. CRISPR technology: A decade of genome editing is only the beginning. **Science**, v. 379, n. 6629, p. eadd8643, 2023. Disponível em: <https://www.science.org/doi/10.1126/science.add8643>. Acesso em: 7 ago. 2024.

[14] WINTER, Sebastian E.; BÄUMLER, Andreas J. Gut dysbiosis: Ecological causes and causative effects on human disease. **Proceedings of the National Academy of Sciences**, v. 120, n. 50, p. e2316579120, 2023. Disponível em: <https://pnas.org/doi/10.1073/pnas.2316579120>. Acesso em: 7 ago. 2024.

[15] BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014. Disponível em: <https://academic.oup.com/bioinformatics/article/30/15/2114/2390096>. Acesso em: 7 ago. 2024.

[16] CHAUMEIL, Pierre-Alain; MUSSIG, Aaron J; HUGENHOLTZ, Philip; *et al*. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. **Bioinformatics**, v. 36, n. 6, p. 1925–1927, 2020. Disponível em: <https://academic.oup.com/bioinformatics/article/36/6/1925/5626182>. Acesso em: 7 ago. 2024.

[17] CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. *In*: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. San Francisco California USA: ACM, 2016, p. 785–794. Disponível em: <https://dl.acm.org/doi/10.1145/2939672.2939785>. Acesso em: 7 ago. 2024.

[18] KANG, Dongwan D.; LI, Feng; KIRTON, Edward; *et al*. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. **PeerJ**, v. 7, p. e7359, 2019. Disponível em: <https://peerj.com/articles/7359>. Acesso em: 7 ago. 2024.

[19] LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 2012. Disponível em: <https://www.nature.com/articles/nmeth.1923>. Acesso em: 7 ago. 2024.

[20] LI, Dinghua; LUO, Ruibang; LIU, Chi-Man; *et al*. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. **Methods**, v. 102, p. 3–11, 2016. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1046202315301183>. Acesso em: 7 ago. 2024.

[21] PARKS, Donovan H.; IMELFORT, Michael; SKENNERTON, Connor T.; *et al*. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. **Genome Research**, v. 25, n. 7, p. 1043–1055, 2015. Disponível em: <http://genome.cshlp.org/lookup/doi/10.1101/gr.186072.114>. Acesso em: 7 ago. 2024.

[22] RUSSEL, Jakob; PINILLA-REDONDO, Rafael; MAYO-MUÑOZ, David; *et al*. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. **The CRISPR Journal**, v. 3, n. 6, p. 462–469, 2020. Disponível em: <https://www.liebertpub.com/doi/10.1089/crispr.2020.0059>. Acesso em: 7 ago. 2024.

[23] SHANG, Jiayu; SUN, Yanni. CHERRY: a Computational metHod for accuratE pRediction of virus–pRokarYotic interactions using a graph encoder–decoder model. **Briefings in Bioinformatics**, v. 23, n. 5, p. bbac182, 2022. Disponível em: <https://academic.oup.com/bib/article/doi/10.1093/bib/bbac182/6589865>. Acesso em: 7 ago. 2024.