Binning de Sequências Anterior à Montagem em Metagenomas: um estudo de caso

Paulo Oliveira[1], **Kleber Padovani**[1], **Raíssa L. da Silva**[1], **Ronnie Alves**[1,2]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Federal do Pará (UFPA)
CEP – 66.075-110 – Belém – PA – Brasil

[2]Instituto Tecnológico Vale (ITV)
CEP – 66.055-090 – Belém – PA – Brasil

{p.paulo.f.oliveira,kleber.padovani,r.lorenna}@gmail.com, ronnie.alves@itv.org

Abstract. *This work, through an empirical study, aimed to answer the following question: Does binning over reads contribute to the production of better assemblies? We evaluated whether quantitative (genome binning) and qualitative (taxonomic binning) approaches bring benefits to the assembly of genomes from metagenome data through statistics which evaluate assemblies considering their sizes and qualities.*

Resumo. *Este trabalho, por meio de um estudo empírico, procurou responder a seguinte questão: O binning sobre reads colabora com a produção de melhores montagens? Buscou-se verificar se o uso das abordagens quantitativa (binning genômico) e qualitativa (binning taxonômico) traz benefícios para a montagem de genomas em metagenomas utilizando estatísticas de avaliação que consideram tamanho e conteúdo das montagens.*

1. Introdução

Na metagenômica, o agrupamento de fragmentos de DNA a um grupo taxonômico correspondente é chamado de *binning*, processo em que cada uma das sequências é alocada em um grupo que representa, de forma ideal, somente os fragmentos pertencentes a determinado táxon [Sedlar 2017]. Nesse sentido, o *binning* é significativo para a reconstrução ou recuperação de genomas de micro-organismos e pode permitir o conhecimento sobre o material genético do metagenoma como um todo.

A maioria dos algoritmos mostram que a precisão do *binning* é aprimorada conforme se aumenta o comprimento das sequências e, por isso, eles normalmente são aplicados após a montagem, sobre os *contigs* [Vollmers 2017]. No entanto, como o processo de montagem é suscetível a erros, as sequências produzidas pelos montadores podem não corresponder a sequências inteiramente corretas, como é o caso dos *contigs* quiméricos [Sedlar 2017].

Como alternativa para contornar este problema, o *binning* pode, também, ser aplicado antes da montagem, e, com isso, pode reduzir a complexidade computacional das montagens subsequentes. Dessa forma, o resultado do processo de *binning* pode ser usado não somente para a avaliação da diversidade taxonômica, mas, também, para facilitar a montagem do genoma [Sedlar 2017].

Nesse sentido, um aspecto importante é que o *binning* pode ser realizado diretamente nas *reads* obtidas do sequenciamento. Ou seja, o *binning* pode ser aplicado antes da montagem particionando as reads em *bins* (grupos) taxonômicos, que podem reduzir de forma significativa a complexidade da montagem de metagenomas.

O *binning* antes da montagem é uma estratégia análoga ao processo de facilitar a montagem de um quebra-cabeça, em que as peças seriam separadas em grupos e cada grupo é montado individualmente, o que pode ser considerado mais fácil do que montar todas peças misturadas. Esta estratégia é conhecida como divisão e conquista. Na investigação de [Constantinescu 2015], o *binning* genômico em conjunto com técnicas de aprendizado de máquina mostrou potencial em reduzir a complexidade e aumento de desempenho na montagem de sequências de DNA.

Este trabalho, por meio de um estudo empírico, procurou responder a seguinte questão: O binning sobre reads colabora com a produção de melhores montagens? Ou seja, buscou-se verificar se a redução de complexidade decorrente do uso das abordagens quantitativa e qualitativa, na análise de *binning* em amostras metagenômicas, traz benefícios ao desempenho da montagem.

2. Materiais e Métodos

2.1. Experimentos

Foram realizados três experimentos para a avaliação do *binning* de *reads* sobre a montagem utilizando subconjuntos de dados de *benchmarking* da iniciativa *Critical Assessment of Metagenomics Interpretation* (CAMI) [Sczyrba et al. 2017]. No primeiro experimento, conforme descrito a seguir, foi utilizada uma abordagem de *binning* taxonômico para a separação das *reads*, cujos *bins* foram posteriormente montados individualmente; no segundo experimento, foi utilizada uma abordagem de *binning* genômico; e, no terceiro, uma abordagem controle, sem a utilização de estratégias de *binning* de *reads*.

2.2. Subamostragem

Em cada experimento, foram utilizadas três subamostragens de dados do CAMI, uma proveniente da única amostra completa de complexidade baixa fornecida pela iniciativa, outra obtida de uma dentre as 4 amostras de complexidade média (RM2, S001) e a terceira provém de uma dentre as 5 amostras de complexidade alta (S001), todas com cerca de 10% do número original de *reads* e o mesmo tamanho de *insert* de 270bp.

Para a avaliação do impacto da subamostragem nos experimentos, foi considerada a estimativa de cobertura de diversidade proposta em [Rodriguez and Konstantinidis 2014] para cada uma das subamostras, bem como para a amostra original, utilizando a ferramenta Prinseq (v0.20.4) [Schmieder and Edwards 2011] para preparação dos dados e a ferramenta Nonpareil (v3 com algoritmo *kmer*) [Rodriguez and Konstantinidis 2014] para o cálculo da estimativa.

Para a geração das subamostras, foi utilizada uma ferramenta para processamento de sequências em formato FASTA/Q disponibilizada gratuitamente na internet, denominada SeqTK (https://github.com/lh3/seqtk), na versão 1.2-r101-dirty. Foram utilizadas, respectivamente e obtidas aleatoriamente, as seeds 15492, 16416 e 30938 como

parâmetros para a ferramenta para a geração de conjuntos aleatórios para construção das subamostras de baixa, média e alta complexidade. Os *scripts* de geração das subamostras e cálculos de cobertura de diversidade, bem como os demais *scripts* utilizados nos métodos descritos a seguir, foram desenvolvidos com a linguagem *python* e estão disponíveis em https://sourceforge.net/p/binning-reads/files/.

2.3. Binning

Para cada subamostra foi realizado o *binning*, com ferramentas de *binning* (*binners*) taxonômico e genômico. No *binning* taxonômico, foi utilizado o *software* Kaiju (v1.4.4) [Menzel et al. 2016], em que foi possível obter os *bins* que correspondem a cada táxon identificado. Com isso, cada *bin* é composto pelas sequências de determinado táxon. Dos diversos *bins* gerados, e para cada um dos três tipos de subamostra, foram escolhidos os seis *bins* que apresentaram maior abundância e relação com os táxons utilizados pelo CAMI. Então, o resultado dessa etapa foi a obtenção de seis *bins* para subamostra de baixa, seis para média e seis para alta complexidade, totalizando dezoito *bins*.

De forma semelhante, foi realizado o *binning* genômico. Inicialmente, foi feita a conversão das amostras FASTQ para FASTA com o *software* SeqTK (adicionalmente, houve a necessidade de desenvolvimento de um *script* para o ajuste dos arquivos FASTA gerados para a correta identificação das *reads*). Em seguida, foi realizado o *binning* genômico, utilizando o *software* MetaProb (v2.0) [Girotto et al. 2016].

Com o MetaProb, as amostras de baixa, média e alta foram analisadas nos tempos de 3h:37m, 3h:33m e 27h, respectivamente, gerando, nesta ordem, 28, 38 e 78 *clusters* para as amostras de complexidades baixa, média e alta. Foi realizada uma análise taxonômica (Kaiju) em cada *cluster* a fim de identificar os táxons mais abundantes. A partir de *script* desenvolvido, os *clusters* com táxons mais abundantes e semelhantes foram mesclados gerando *bins* dos quais foram selecionados os seis mais abundantes e relacionados ao CAMI, de cada tipo de amostra.

Após as análises de *binning* taxonômico e genômico, os *bins* resultantes foram utilizados posteriormente na montagem para as avaliações seguintes.

2.4. Montagem e Obtenção das Estatísticas

Cada *bin* gerado pelos *binners* taxonômico e genômico foram montados com o montador MegaHit (v1.1.2) [Li et al. 2015] em suas configurações padrão. Da mesma forma, foram montadas as subamostras, de baixa, média e alta complexidade, diretamente, sem o uso de *binning*.

Posteriormente, as montagens (*contigs*) foram submetidas a três ferramentas e um *script* para então ser realizada a análise e comparação. Com o intuito de avaliar a qualidade da recuperação do genoma, foram utilizadas as métricas de [Parks 2015] da ferramenta CheckM (v1.0.11, seguindo o *workflow* de identificação automática de táxons *lineage_wf*). Dessas métricas, foram selecionadas as estimativas de completude, qualidade e contaminação do genoma montado.

Para avaliar a montagem dos genomas, foram adotadas as medidas de comparação de desempenho de montagens proposta em [Mikheenko et al. 2015]. Utilizou-se a ferramenta MetaQuast (v4.2) sobre os genomas montados para alinhar com os genomas de

referências, identificando as medidas de fração de recuperação do genoma, fragmentação da montagem e a estatística N50. As medidas de fração de recuperação de genoma e N50 também foram avaliadas nas montagens sem *binning* para as sequências de baixa, média e alta complexidade.

Com o objetivo de auxiliar na análise taxonômica, foi executado a ferramenta de predição de genes FragGeneScan (v1.30), sobre os genomas montados para medir a quantidade de sequências codificadoras [Rho et al. 2010] encontradas em cada montagem do táxon correspondente. Posteriormente foram removidas, via *script*, as sequências redundantes para contagem de sequências codificadoras distintas.

Para a comparação de desempenhos das abordagens com e sem aplicação de *binning*, foram considerados somente táxons comuns nos grupos taxonômicos e genômicos, ou seja, que estavam presentes no dois grupos. Essa comparação pode ser vista na seção seguinte.

3. Resultados

Os resultados a seguir são referentes às métricas adotadas pelas ferramentas CheckM, MetaQuast e FragGeneScan. Além disso, somente foram usados os táxons comuns em todas as abordagens e que estão presentes no estudo disponibilizado pelo CAMI, totalizando oito espécies (genomas)[1].

Foram encontrados os táxons *Albidovulum xiamenense*, *Paracoccus denitrificans* e *Pseudomonas aeruginosa* na amostra de complexidade baixa, identificados na Tabela 1 pelos números 1, 2 e 3, respectivamente; os taxons *Azospirillum brasilense*, *Moorella thermoacetica*, *Phaeospirillum fulvum* e *Sinorhizobium meliloti* na amostra de complexidade média, identificados pelos números 4, 5, 6 e 7; e o táxon *Salegentibacter salarius* na amostra de complexidade alta, identificado pelo número 8.

A Tabela 1 apresenta os valores para as métricas obtidas com as ferramentas CheckM, MetaQuast e FragGeneScan. Na maioria dos *bins*, tanto para as métricas do CheckM quanto para a estatística proveniente da análise do FragGeneScan, o *binning* taxonômico obteve maior qualidade em relação ao genômico.

A partir dos resultados do MetaQuast, também são apresentados na Tabela 1, a fração alcançada do genoma correspondente ao *bin*, o índice de fragmentação das montagens e as taxas de N50 para cada táxon. Considerando, isoladamente, as métricas de fração de genoma e N50, as montagens que não fizeram uso de abordagens de *binning* apresentaram melhores resultados. Contudo, é válido observar que, para alguns táxons, a diferença entre os valores foi muito pequena.

A análise comparativa entre as abordagens com e sem *binning* considerou apenas essas duas métricas (N50 e fração do genoma) pois não foi possível calcular as demais métricas para a montagem sem *binning* sem que isso implicasse em um viés do *binning* pós-montagem, uma vez que o agrupamento por táxon é necessário para a obtenção dessas métricas.

Levando-se em consideração as cinco métricas (Qualidade, Fragmentação, Fração

[1]Informações complementares a respeito das avaliações realizadas estão disponíveis em https://sourceforge.net/projects/binning-reads/files/docs/details.xlsx

Tabela 1. Comparativo dos resultados das ferramentas de análise.

	CheckM		MetaQuast								FragGeneScan	
	Qualidade		Fração do genoma (%)			Fragmentação(%)		N50			CDS	
ID	Taxonômico	Genômico	Taxonômico	Genômico	Sem binning	Taxonômico	Genômico	Taxonômico	Genômico	Sem binning	Taxonômico	Genômico
1	**13,67**	4,17	6,753	2,531	**6,775**	74,66	**71,5**	607	585	**646**	**1477**	1391
2	**48,14**	0	75,557	82,264	**96,224**	23,71	**18,31**	2146	4877	**6864**	5792	**9857**
3	31,87	**33,64**	29,471	67,019	**86,557**	42,67	**12,26**	1016	13218	**116689**	**5319**	4301
4	**12,89**	3,78	**10,048**	4,997	9,466	**57,29**	65,79	**707**	698	**707**	**9870**	6357
5	10,53	**27,99**	35,275	12,143	**37,578**	**53,91**	61,78	721	724	**758**	**3898**	1444
6	**14,96**	8,33	1,02	0,482	**2,371**	**65,21**	70,62	649	670	**681**	**4012**	3207
7	**14,64**	0	49,916	72,934	**73,457**	**55,47**	59,3	812	869	**870**	16442	**54785**
8	0	**14,27**	31,497	8,715	**71,04**	**49,82**	76,83	853	718	**1123**	3084	**3600**

Táxons nas amostras de complexidade: **Baixa** **Média** **Alta**

do Genoma, N50 e número de CDS), o *binning* taxonômico mostrou-se melhor. Esse resultado se confirma sob três perspectivas. Primeiramente, por contagem total de melhor resultado dentre as cinco métricas, onde em 57,5% dos melhores resultados foram obtidos pelo *binning* taxonômico. Analisando por táxon, os melhores resultados foram obtidos pelo *binner* taxonômico em 62,5% dos táxons. Na análise de cada métrica isoladamente, em 62,5% dos casos o *binning* taxonômico obteve melhores resultados em 4 das métricas, e somente com a métrica N50 o *binning* genômico foi superior.

4. Discussão e Conclusões

Nos experimentos realizados, o *binning* pré-montagem não foi benéfico para a montagem subsequente, no entanto, a investigação mais aprofundada de alguns aspectos pode ser válida. Dentre eles, podemos citar a exploração de mais *binners* taxonômicos e genômicos (como os citados em [Sczyrba et al. 2017] e [Mande 2012]), com o intuito de se verificar seus desempenhos na montagem de genomas. Além disso, novos métodos de agrupamento de sequências - incluindo técnicas de aprendizado de máquina, como a ferramenta apresentada por [Vervier et al. 2018] - são outras possibilidades que podem ser investigadas para avaliação de melhorias na montagem decorrentes da aplicação de binning sobre reads.

A investigação do impacto do *binning* pré-montagem nas amostras completas do CAMI, sem a utilização de subamostragem, pode produzir resultados distintos dos alcançados, dado o possível viés contido na escolha aleatória de *reads*. Ainda considerando o conjunto de dados, ressalta-se que os experimentos foram realizados em ambiente controlado, em que se conhecem os táxons contidos nas amostras, porém, o cenário metagenômico é distinto deste e, considerando organismos ainda não sequenciados, o *binning* genômico pode apresentar melhor adequação ao problema. Dessa forma, torna-se importante também a avaliação de *binners* em cenários com escassez de referências.

Embora os resultados com *binning* tenham se apresentado inferiores aos resultados sem *binning*, é válido citar o desempenho do *binning* taxômico contra o desempenho do *binning* genômico, que se mostra mais adequado ao *binning* sobre *reads*, ao se tratar de benefícios à montagem. Contudo, conforme mencionado, acredita-se que essa comparação pode produzir resultados distintos quando a diversidade contida nas amostras

correspondem a organismos ainda não sequenciados, podendo, nesse caso, alcançarmos resultados mais favoráveis aos *binners* genômicos, que não dependem de referência.

Na métrica N50 o melhor resultado foi com o *binning* genômico, isso pode ter ocorrido devido às montagens maiores obtidas, porém, com inconsistências resultantes, por exemplo, de quimeras. Contudo, ainda em direção ao que fora mencionado, também pode ser um indicativo de falta de informação a respeito dos táxons considerados, já que três das cinco métricas utilizadas na avaliação dependem de referência e isso, também, pode ter influenciado negativamente a avaliação do *binning* pré-montagem.

Referências

Constantinescu, R.-I. (2015). A machine learning approach to dna shotgun sequence assembly. Master's thesis, University of the Witwatersrand.

Girotto, S., Pizzi, C., and Comin, M. (2016). Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*.

Mande, S. S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, 13:669–681.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2015). Metaquast: evaluation of metagenome assemblies. *Bioinformatics*, 32:1088–1090.

Parks, D. H. (2015). Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25:1043–1055.

Rho, M., Tang, H., and Ye, Y. (2010). Fraggenescan: predicting genes in short and error-prone reads. *Nucleic acids research*, 38(20):191–191.

Rodriguez, L. M. and Konstantinidis, K. T. (2014). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics*.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27:863–864.

Sczyrba, A., Hofmann, P., and Belmann, P. (2017). Critical assessment of metagenome interpretation – a benchmark of computational metagenomics software. *Nature methods*, 14(11):1063–1071.

Sedlar, K. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*, 15:48–55.

Vervier, K., Mahé, P., and Vert, J.-P. (2018). *MetaVW: Large-Scale Machine Learning for Metagenomics Sequence Classification*, pages 9–20. Springer New York, New York, NY.

Vollmers, J. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS ONE 12.1*.