

Reconstrução da rede metabólica em escala genômica da bactéria
Burkholderia sacchari

Paulo Moises Raduan Alexandrino

DISSERTAÇÃO APRESENTADA
AO
PROGRAMA INTERUNIDADES EM BIOINFORMÁTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Orientador: Prof. Dr. André Fujita

Coorientador: Prof. Dr. José Gregório Cabrera Gomez

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq

São Paulo, março de 2016

Reconstrução da rede metabólica em escala genômica da bactéria
Burkholderia sacchari

Esta é a versão original da dissertação elaborada pelo
candidato Paulo Moises Raduan Alexandrino, tal como
submetida à Comissão Julgadora.

Comissão Julgadora:

- Prof. Dr. André Fujita (orientador) - IME/USP
- Prof. Dr. Gabriel Padilla - ICB/USP
- Dr^a. Cecília Coimbra Klein - CRG/Barcelona

Agradecimentos

Agradeço ao meu orientador, André Fujita, a liberdade que me foi dada para a escolha e desenvolvimento do tema da presente dissertação. Além disso, agradeço-lhe o auxílio financeiro que recebi para participar do curso de Engenharia de Biosistemas que aconteceu na cidade de Braunwald, Suíça, em setembro de 2015. Agradeço ainda aos meus amigos do laboratório de e-Science a amizade e o companheirismo. Dentre eles, sou grato aos peruanos Edu, Leandro, Waldir, Jorge, Leici e Urpy e aos colombianos Juan e Sabrina, que me alfabetizaram em cultura andina. E com relação aos brasileiros, destaco os amigos Fernando, Gustavo, Suzana, Gabriela, Thaianne, Caio, Gerson, Eric e David. Um agradecimento especial fica para Fernando e Juan, com quem pude, sempre ao lado de um cafezinho, filosofar e ter conversas engrandecedoras.

Agradeço ao meu co-orientador, José Gregório Cabrera Gomez, e à professora Luiziana Ferreira da Silva, ambos do Laboratório de Bioprodutos do Instituto de Biociências da USP, tanto a orientação como a oportunidade de ter o primeiro contato com a pesquisa na área de biotecnologia industrial. Agradeço os ensinamentos teóricos e práticos, a disposição e a paciência destes professores. Sou muito grato também ao auxílio que recebi para participar do 14^o Simpósio Internacional em Biopolímeros que aconteceu em Santos, em outubro de 2014. Além disso, agradeço aos amigos do Laboratório de Bioprodutos, em especial à Juliana Cardinali, o apoio nas atividades laboratoriais e os ensinamentos em biologia molecular, e à Thatiane Mendonça, o aprendizado na condução de experimentos em biorreator.

Agradeço ao professor Andreas Karoly e também à FAPESP a oportunidade e o incentivo financeiro que me foram dados para participar da Escola Avançada FAPESP em Bioenergia, na cidade de Campinas, em outubro de 2014.

Agradeço ao Mateus Lopez e ao Paulo Coutinho, ambos em nome da Braskem, empresa em que passei o último ano do mestrado desenvolvendo um projeto sobre a prospecção computacional de novas rotas metabólicas para a produção de químicos a partir de açúcares de segunda geração. Este projeto foi desenvolvido no âmbito do programa Inova Talentos, do CNPq, a quem também sou agradecido, e ajudou tanto a complementar minha formação acadêmica como a amparar minha saúde financeira quando a bolsa que recebia já estava chegando ao fim.

Agradeço à secretária do programa de pós-graduação Interunidades em Bioinformática, Patrícia Martorelli, o serviço prestado e a paciência para transmitir informações, dicas e normas do programa.

Agradeço finalmente à minha família, Paulo, Rose, Rosiane, Paula e Nina, pelo apoio e por terem estado sempre lá. Sempre estiveram.

Resumo

Burkholderia sacchari é uma bactéria que naturalmente produz polihidroxialcanoatos (PHAs) como reserva de carbono e de energia. PHAs são uma classe de polímeros biodegradáveis que apresentam uma série de aplicações industriais e *B. sacchari* se mostrou um organismo promissor para a produção em larga escala do polímero. Recentemente, o genoma da bactéria foi sequenciado, abrindo a possibilidade para a reconstrução de sua rede metabólica em escala genômica, o que poderia contribuir para o entendimento da fisiologia da bactéria, com implicações para sua melhoria genética visando a produção mais eficiente de PHAs. Desse modo, o primeiro objetivo deste trabalho é reconstruir a rede metabólica da bactéria *Burkholderia sacchari* a partir de seu conteúdo genômico. Para isso, o genoma sequenciado foi montado e anotado. A montagem foi feita usando 28 tratamentos diferentes que foram avaliados de acordo com três critérios: métricas de tamanho da montagem, critério de qualidade da montagem, e completude da anotação. Os três critérios foram usados como filtros para eleger a melhor montagem. O genoma eleito foi então usado para a reconstrução automática da rede metabólica usando a ferramenta ModelSEED, resultando em uma rede composta por 1512 reações e 1890 metabólitos. Em seguida, a reconstrução obtida foi sujeita à adequação de vocabulário por meio da plataforma MetaNetX, onde também foram realizadas dois tipos de análise da rede. A análise de balanço de fluxo mostrou que o modelo suporta crescimento. A análise de nocautes de reações mostrou uma inconsistência no metabolismo de trealose. Assim, curou-se manualmente a reconstrução e foi constatada uma falha na anotação da enzima fosfoglutamase, que foi considerada como candidata à reanotação. O processo de curadoria desta enzima indicou duas lacunas: a dificuldade em se percorrer a via de uma maneira interativa e a dificuldade em se modificar a reconstrução. Em decorrência das dificuldades encontradas, o segundo objetivo deste trabalho foi propor uma forma de representar a reconstrução de *B. sacchari* visando superar as lacunas encontradas. A forma proposta foi através de um banco de dados orientado a grafo e os resultados indicaram que esta representação consegue facilitar a tarefa de curadoria manual que precisará ser feita em trabalhos futuros de refinamento da reconstrução.

Palavras-chave: polihidroxialcanoatos, engenharia metabólica de sistemas, montagem *de novo*, anotação multidimensional, reconstrução metabólica, biocuradoria, modelo metabólico, banco de dados em grafo

Abstract

Burkholderia sacchari is a bacterium that naturally produces polyhydroxyalkanoates (PHAs) as a carbon and energy storage material. PHAs are a class of biodegradable polymers which have many industrial applications and the bacteria has emerged as a potential host for large-scale production of the polymer. *B. sacchari*'s genome has been recently sequenced, opening up the possibility for its metabolic network reconstruction. This could in turn be helpful to understand *B. sacchari*'s physiology, with implications to the genetic improvement of the bacterium, leading to a more efficient production of PHAs. Therefore, the first goal of this work is to reconstruct the metabolic network of *B. sacchari* from its genomic content. Hence, the sequenced genome was assembled and annotated. Assembly was performed with 28 different treatments, and each of them was evaluated according to three criteria: size metrics, assembly quality criteria and annotation completeness. All three criteria together were used as a filter to select the better assembly. The selected genome was used in the automatic reconstruction with ModelSEED, resulting in a network composed of 1512 reactions and 1890 metabolites. Then, the reconstruction was subject to vocabulary standardization with Meta-NetX, where two types of analysis were also carried out. Flux Balance Analysis showed that the model can support growth and Reaction Knockout Analysis revealed an inconsistency in trehalose metabolism. Therefore, the network was manually inspected and the shortcoming was found to be related to the incorrect annotation of the enzyme phosphoglucomutase, which was considered as a candidate for reannotation. The manual curation of this enzyme has indicated two gaps: the difficulty in walking through the pathway and the difficulty in altering the reconstruction. Facing these difficulties, the second goal of this work was to propose a representation of *B. sacchari*'s metabolic network as a graph database. Results suggest that this form of representation can facilitate the task of manual inspection that should be made in future work aiming to refine the reconstruction.

Keywords: polyhydroxyalkanoates, systems metabolic engineering, *de novo* assembly, multidimensional annotation, metabolic network reconstruction, biocuration, metabolic model, graph database

Sumário

Lista de Abreviaturas	vi
Lista de Símbolos	vii
Lista de Figuras	viii
Lista de Tabelas	ix
1 Introdução	1
1.1 Contextualização	1
1.2 Motivação	2
1.3 Objetivos	4
1.4 Organização do Trabalho	4
2 Conceitos Fundamentais e Definições	6
2.1 O genoma: o que é e como conhecer seu conteúdo	6
2.2 O conceito de anotação multidimensional do genoma	8
3 Montagem e anotação do genoma de <i>B. sacchari</i>	11
3.1 Considerações Preliminares	11
3.2 Revisão de Literatura	11
3.3 Estratégia adotada para <i>B. sacchari</i>	13
3.4 Materiais e Métodos	14
3.5 Resultados e Discussão	20
3.6 Conclusões	28
4 Reconstrução da rede metabólica <i>B. sacchari</i>	29
4.1 Considerações Preliminares	29
4.2 Revisão de Literatura	29
4.3 Proposta adotada para <i>B. sacchari</i>	31
4.4 Métodos	32
4.5 Resultados e Discussão	34
4.6 Conclusões	37
5 Representação da reconstrução	40
5.1 Considerações Preliminares	40
5.2 Lacunas encontradas no processo de curadoria manual	40

5.3	Proposta de solução das lacunas identificadas	44
5.4	Materiais e Métodos	45
5.5	Resultados e Discussão	46
5.6	Conclusões	50
6	Considerações Finais	51
6.1	Contribuições	51
6.2	Sugestões para Pesquisas Futuras	53
A	Artigo publicado	54
B	Créditos das imagens	57
	Referências Bibliográficas	58

Lista de Abreviaturas

BLAST	<i>Basic Local Alignment Search Tool</i>
EC	<i>Enzyme Commission</i>
ETL	<i>Extract-Transforms-Load</i>
HV	Hidroxivalerato
JSON	<i>JavaScript Object Notation</i>
OLC	<i>Overlap-Layout-Consensus</i>
P3HB	Polihidroxibutirato
pb	par de base
SBML	<i>Systems Biology Markup Language</i>
SCS	<i>Shortest Common Superstring</i>
SFF	<i>Standard Flowgram Format</i>
TSV	<i>Tab-Separated Values</i>

Lista de Símbolos

- [c] Compartimento do citoplasma
- [e] Região extracelular
- [x] Compartimento externo ao modelo
- [p] Espaço periplasmático

Lista de Figuras

1.1	Ciclo de engenharia metabólica tradicional e de sistemas	5
2.1	Ilustração do genoma de uma célula bacteriana	6
2.2	Visão geral das etapas do sequenciamento e da montagem de um genoma	8
2.3	O conceito de anotação multidimensional	9
3.1	Distribuição de tamanho dos <i>reads</i> provindos dos dados brutos.	15
3.2	Super-representação da sequência GACT	15
3.3	Tratamento dado ao conjunto de <i>reads</i>	18
3.4	Aumento cumulativo do tamanho das montagens conforme o aumento incremental do número de <i>contigs</i>	22
3.5	Avaliação da montagem MIRA com relação ao gene 16S	23
3.6	Avaliação da montagem Newbler com relação ao gene 16S	24
3.7	Avaliação da montagem CABOG com relação ao gene 16S	25
3.8	Comparação do conteúdo de gênico entre pares de anotações	27
3.9	Comparação do conteúdo de gênico restrito entre pares de anotações	28
4.1	Protocolo de Thiele e Palsson (2010)	31
4.2	Avaliação de ferramentas de reconstrução automática	32
4.3	A equação de formação de biomassa de <i>B. sacchari</i>	34
4.4	Classificação das reações de <i>B. sacchari</i> pelo MetaNetX	35
4.5	A enzima fosfoglucomutase em diferentes organismo.	38
4.6	Alinhamento entre duas enzimas fosfomanomutases	39
5.1	Percorrendo vias metabólicas com o MetaNetX	41
5.2	Exemplo de procedimento par a adição de reação à reconstrução com o MetaNetX	42
5.3	Exemplo de procedimento par a remoção de reação da reconstrução com o MetaNetX	43
5.4	Modelagem conceitual do banco de dados	46
5.5	Percorrendo vias metabólicas com o banco de dados	47
5.6	Acionando reações no banco de dados	48
5.7	Removendo reações do banco de dados	49
6.1	Proposta de divisão do gênero <i>Burkholderia</i>	52

Lista de Tabelas

3.1	Sequências de nucleotídeos disponíveis no NCBI para <i>Burkholderia sacchari</i>	19
3.2	Métricas de tamanho das montagens	20
3.3	Estatísticas das anotações.	26
4.1	Subconjunto dos 293 nocautes letais que envolvem o transporte de metabólitos entre o compartimento externo e a região extracelular	36
4.2	Reações envolvendo G1P que estão presentes em outros modelos metabólicos	37
4.3	Lista mostrando a disjunção entre os 293 nocautes letais obtidos antes da reanotação e os 288 nocautes letais obtidos na análise posterior	38

Capítulo 1

Introdução

Este capítulo descreve o contexto em que este trabalho se insere como também apresenta a motivação e seus objetivos.

1.1 Contextualização

Os polihidroxicanoatos (PHAs) são uma classe de polímeros produzidos naturalmente por bactérias e arqueas sob a forma de grânulos intracelulares (Anderson e Dawes, 1990). Apesar de serem atualmente concebidos como classe, o polihidroxibutirato (P3HB) foi durante quase meio século a única instância de PHA conhecida: sua descoberta se deu em 1926 por Lemoigne, ao passo que a primeira descrição de outros PHAs foi feita somente em 1974 por Wallen e Rohwedder (Lemoigne, 1926; Wallen e Rohwedder, 1974). Entre estes dois marcos, uma série de descobertas foram feitas sobre as propriedades e as funções do P3HB. Foi descoberto, por exemplo, que os micro-organismos produtores de P3HB acumulam esse polímero como material de reserva de carbono e de energia em decorrência de situações onde há excesso de fonte de carbono ao mesmo tempo em que há limitação de algum nutriente essencial, como por exemplo, nitrogênio, fósforo ou oxigênio (Steinbüchel e Füchtenbusch, 1998). Outras descobertas de grande relevância foram em torno da biodegradabilidade do P3HB. Micro-organismos conseguem degradar o polímero P3HB até seus monômeros e usá-los como fonte de carbono (Lepidi, 1972; Merrick e Doudoroff, 1964).

Apesar de sua biodegradabilidade e de poder ser produzido a partir de fontes renováveis, o fato de P3HB ser um polímero rígido e quebradiço fez com que suas aplicações industriais fossem reduzidas. Foi somente a partir de 1974, que um amplo interesse foi voltado aos PHAs pois a descoberta de novos monômeros permitiu que novos polímeros, com diferentes propriedades termo-mecânicas, pudessem ser produzidos (Sudesh *et al.*, 2000). Assim, vários PHAs foram caracterizados e uma abundância de aplicações foi proposta. Hoje, a escassez de propriedades termo-mecânicas não é mais uma lacuna, contudo o preço o é. Assim, diversos esforços ainda são feitos para otimizar e baratear a produção de PHAs.

É nesse contexto de busca por melhorias na produção de PHAs, em particular na busca por melhores linhagens, em que se insere a descoberta da bactéria *Burkholderia sacchari* LMG19450, que foi isolada no Brasil a partir do solo de uma plantação de cana-de-açúcar no trabalho de Gomez *et al.* (1996). Este estudo visava rastrear bactérias produtoras de PHAs a partir de carboidratos e também de ácido propiônico. *B. sacchari*, que até então era conhecida por este nome, se mostrou um micro-organismo promissor para ser usado na produção industrial de PHAs, pois apre-

sentou diversas características de interesse, como o acúmulo de altas taxas de polímero em relação à massa celular (até 68% de massa seca), versatilidade no uso de diferentes açúcares como substratos para a produção do polímero (incluindo substratos onde o Brasil se destaca mundialmente pela produção, como é o caso da sacarose), ausência de patogenicidade, acúmulo poli-3-hidroxi-butirato (P3HB) a partir de sacarose com um rendimento acima de 80% do máximo teórico, entre outras (Gomez *et al.*, 1996). Por outro lado, apesar dessas características, o estudo também identificou que a bactéria não converte ácido propiônico até hidroxivalerato (HV) com uma boa eficiência, e isso tem implicações negativas caso se queira produzir o copolímero P(3HB-co-HV) com bom rendimento. No mais, a bactéria é promissora, mas precisa de melhorias genéticas a depender de sua aplicação. Assim, desde que *B. sacchari* foi isolada, o Laboratório de Bioprodutos da Universidade de São Paulo empreendeu uma série de iniciativas tanto para melhor caracterizar a fisiologia da bactéria quanto para identificar potenciais de melhoria.

1.2 Motivação

O modo como os gargalos identificados em *B. sacchari* foram abordados pelo Laboratório de Bioprodutos seguiu um panorama comum: primeiro, com base na literatura disponível, levantava-se uma hipótese acerca do alvo a ser atacado. A seguir, o alvo selecionado era atacado por meio de modificações genéticas e, por fim, contrastava-se os resultados obtidos experimentalmente com os previstos. Essa forma geral de abordar o problema está esquematizada na figura 1.1 e ela implicou em dois tipos de resultados.

De um lado, o problema identificado foi atacado com sucesso ao se percorrer o ciclo uma única vez. Esse é o caso do trabalho de Silva (2000), onde o problema identificado foi o baixo rendimento que a bactéria apresentou na conversão de propionato a HV mencionado. A hipótese levantada foi a de que haveriam duas ou mais vias de catabolismo de propionato que o estariam desviando do seu fim preterido, o HV. Diante dessa hipótese, os autores planejaram modificações genéticas por meio da técnica de mutagênese aleatória com raios ultravioleta e selecionaram mutantes afetados no metabolismo de propionato. Como última etapa do ciclo, os mutantes foram testados em biorreatores e as análises indicaram que o rendimento havia aumentado significativamente. Embora este trabalho de Silva e colaboradores tenha culminado nas análises de caracterização posteriores feitas por Bramer *et al.* (2002) e mais tarde por Pereira *et al.* (2009), estas análises visaram a comprovação da hipótese levantada, sendo que o problema inicialmente identificado foi praticamente solucionado em um único ciclo de engenharia metabólica.

Por outro lado, o trabalho de Lopes *et al.* (2009) é um dos casos onde o problema identificado não foi solucionado com um único ciclo de engenharia metabólica. Neste exemplo, o gargalo identificado foi a baixa produtividade que *B. sacchari* apresentou na síntese de P3HB a partir de xilose em comparação com a síntese a partir de glicose. Então, o gene *xylA* foi selecionado como alvo com base na hipótese de que sua superexpressão poderia contribuir para o aumento da taxa de crescimento celular. Assim, o gene *xylA* foi clonado e superexpressado em *B. sacchari* como estratégia de modificação. Os experimentos foram conduzidos em biorreator, mas não indicaram aumento da produtividade, que era o problema inicial. Embora o estudo não tenha continuado, para que o problema inicialmente levantado fosse resolvido, seria necessário que uma nova hipótese fosse levantada, dando início a um novo ciclo.

Os dois exemplos apresentados tanto atestam que o ciclo tradicional de engenharia metabólica fornece bons resultados, quanto atestam suas limitações: a possibilidade de o ciclo ter que ser percorrido inúmeras vezes traz consigo o risco deste processo se revelar extenso e custoso. Isso vai ao encontro de outros trabalhos que confirmam tais limitações, como são o caso de Tee *et al.* (2014), Knuf e Nielsen (2012), Lee *et al.* (2011) e Blazeck e Alper (2010). Para contornar as limitações, estes trabalhos sugerem que o ciclo tradicional seja complementado com abordagens oriundas do campo da Biologia de Sistemas. Nesta nova abordagem, nomeada por Lee *et al.* (2011) de Engenharia Metabólica de Sistemas, ocupam papel de destaque os modelos computacionais que simulam o funcionamento do metabolismo microbiano em uma escala sistêmica.

A importância destes modelos está baseada no fato de que seu uso traria mais dinamismo ao processo de melhoria de linhagens. De acordo com Knuf e Nielsen (2012), isso se justifica pois os resultados das simulações usando os modelos poderiam guiar a escolha de alvos a serem atacados, complementando assim a geração de hipóteses baseadas somente na literatura. Além disso, ainda de acordo com esses autores, os modelos poderiam ser usados como arcabouço para a análise e interpretação de dados ômicos (transcriptômica, metabolômica, fluxômica, e assim por diante), o que serviria para refinar o modelo, melhorar sua capacidade preditiva e, em consequência, ajudaria na elaboração de hipóteses mais acuradas.

Exemplos de benefícios no uso de modelos computacionais sistêmicos podem ser encontrados em (Blazeck e Alper, 2010). Em um dos exemplos elencados, Blazeck e Alper conseguem aumentar o rendimento de treonina a partir de um nocaute triplo que havia sido simulado *in silico*. Sem a simulação, atestam os autores, essa estratégia seria inviável.

Assim, para tomar proveito dos benefícios atestados na literatura com respeito a esse tipo de abordagem, seria de grande valia construir um modelo computacional sistêmico para *Burkholderia sacchari*. Para isso, de acordo com Thiele e Palsson (2010a), é preciso antes fazer um processo conhecido por reconstrução da rede metabólica, um processo que envolve o catálogo e organização do conhecimento a respeito do genoma, da fisiologia e do metabolismo do organismo em questão. Em resumo, o modelo computacional é uma representação de uma reconstrução metabólica em um formato factível para simulações. Enquanto que a reconstrução é mais ampla, um catálogo. Entretanto, reconstrução e modelo são intimamente relacionados, pois a acurácia do modelo em prever o fenótipo está intimamente relacionada com quão bem anotada está a reconstrução. Nesse sentido, reconstruir uma rede metabólica em escala genômica é um empreendimento extenso e iterativo. Por exemplo, para *Escherichia coli*, o processo já dura mais de 10 anos e ainda restam muitas lacunas em sua reconstrução, muito embora tenham havido uma série de resultados positivos com o modelo dessa enterobactéria (Oberhardt *et al.*, 2009).

Apesar de extensa, as reconstruções podem ser facilitadas com o uso de ferramentas automatizadas que partem de um genoma anotado, ou até mesmo de um genoma montado, e reconstróem automaticamente a rede metabólica. Somado a isso está o fato de que *Burkholderia sacchari* teve seu genoma recentemente sequenciado no trabalho de Alexandrino *et al.* (2015). Assim, seria viável reconstruir a rede metabólica de *B. sacchari*, mesmo que de uma maneira incipiente e sujeita a esforços de curadoria manual em trabalhos posteriores.

1.3 Objetivos

Atestada a importância de se obter uma reconstrução para *B. sacchari* e a viabilidade deste empreendimento, o primeiro objetivo deste trabalho é reconstruir a rede metabólica em escala genômica a partir do sequenciamento da bactéria. Além disso, o segundo objetivo deste trabalho é propor uma forma de representar a reconstrução de modo a facilitar os extensos trabalhos de curadoria manual que serão feitos no futuro.

Com relação ao que foi dito, os objetivos específicos são:

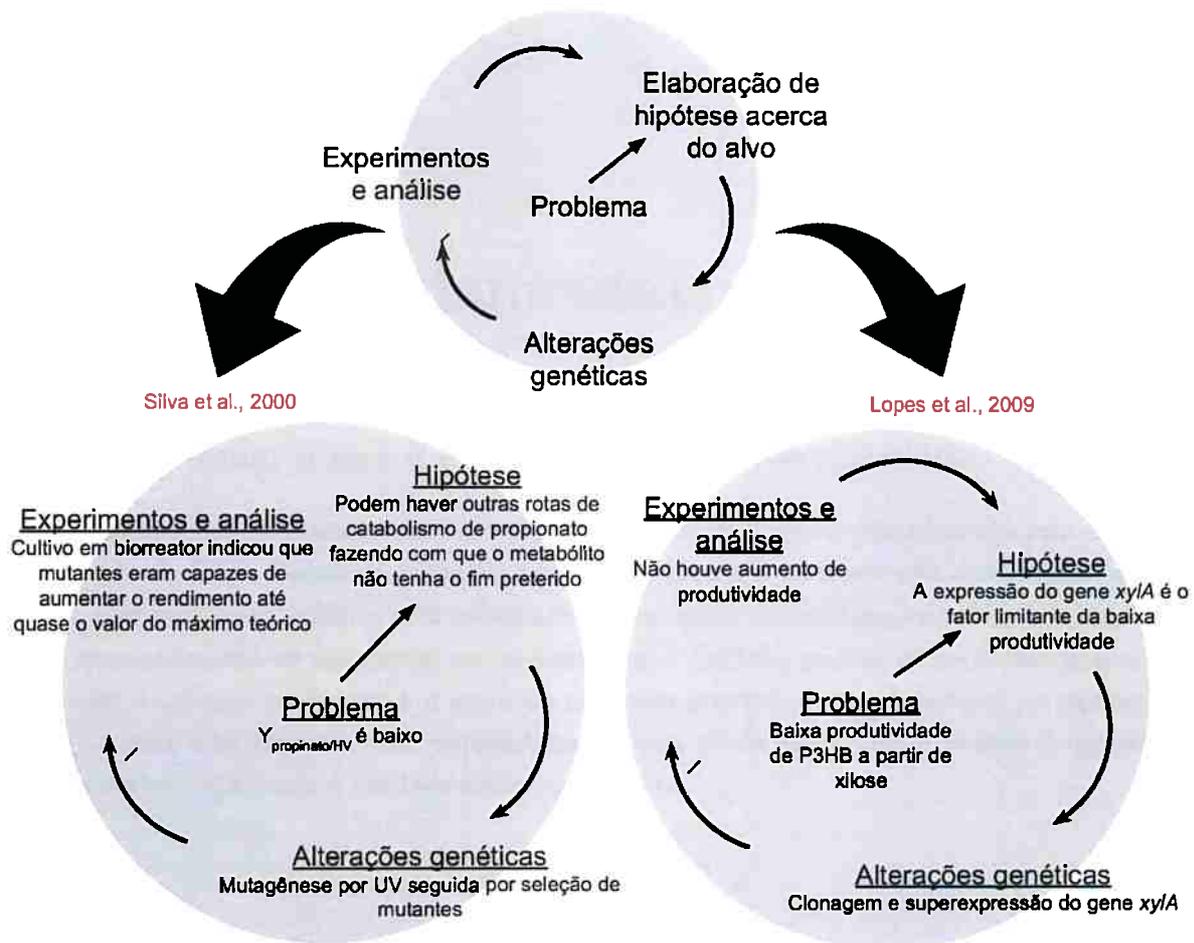
- I. Objetivo 1: reconstruir a rede metabólica em escala genômica a partir do sequenciamento da bactéria
 - I.i. Montar o conjunto de *reads* gerados pelo sequenciamento;
 - I.ii. Anotar o genoma montado;
 - I.iii. Reconstruir a rede metabólica a partir do genoma anotado;
- II. Objetivo 2: Propor uma forma de representar a reconstrução de modo a facilitar as etapas de curadoria manual
 - II.i. Representar a reconstrução de *B. sacchari* como um banco de dados orientado a grafo

A pormenorização dos objetivos específicos e os métodos empregados para tanto são apresentados nos capítulos apropriados, conforme a organização a seguir:

1.4 Organização do Trabalho

O trabalho está dividido em seis capítulos, incluindo esta Introdução. No Capítulo 2, são apresentados os conceitos fundamentais para a compreensão dos capítulos subsequentes. O Capítulo 3 mostra como o genoma de *B. sacchari* foi montado, anotado e validado. O Capítulo 4 trata do processo de reconstrução da rede metabólica, feita a partir do genoma anotado. Além disso, este capítulo trata das análises que foram feitas na rede reconstruída e mostra uma etapa de curadoria manual. Depois, no Capítulo 5, duas lacunas identificadas durante as etapas de curadoria manual são mostradas, além de uma proposta de solução para resolvê-las. Finalmente, no Capítulo 6 é apresentada uma visão geral do trabalho, suas contribuições e sugestões para pesquisas futuras.

a)



b)

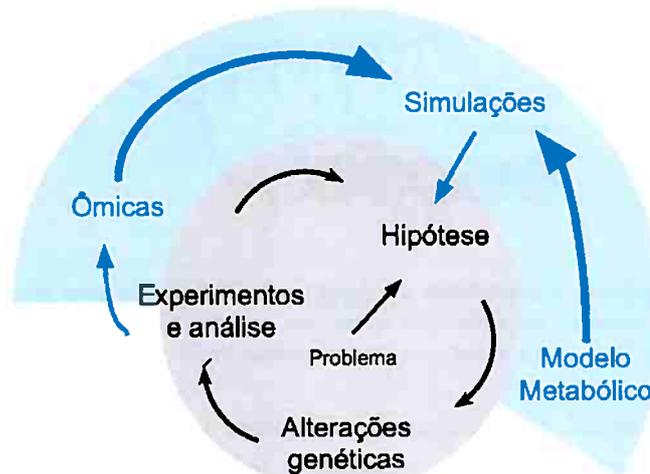


Figura 1.1: Ciclo de engenharia metabólica tradicional e de sistemas. Em (a), é mostrado o panorama geral adotado pelo Laboratório de Bioprodutos da USP para a melhoria genética de *Burkholderia sacchari*, no contexto da engenharia metabólica tradicional. A partir desse panorama, são exemplificadas duas instâncias: o trabalho de Silva (2000) e o trabalho de Lopes et al. (2009). Em (b), é mostrada a sugestão Knuf e Nielsen (2012) para a complementação do ciclo tradicional com abordagens de Biologia de Sistemas, dando resultado a um ciclo de engenharia metabólica expandido.

Capítulo 2

Conceitos Fundamentais e Definições

2.1 O genoma: o que é e como conhecer seu conteúdo

O genoma é todo o conteúdo genético presente em um organismo e este conteúdo está codificado na molécula de DNA. Bactérias, como é o caso de *Burkholderia sacchari*, não apresentam envelope nuclear e seu conteúdo genético está localizado em um único compartimento celular, o citoplasma. Neste compartimento se encontram os cromossomos e também podem se encontrar plasmídeos, sendo que o genoma neste caso é a soma do conteúdo genético destes dois tipos de replicons. A figura 2.1 traz uma representação esquemática de uma célula bacteriana e mostra o que seria o genoma nestes organismos e sua base molecular, o DNA.

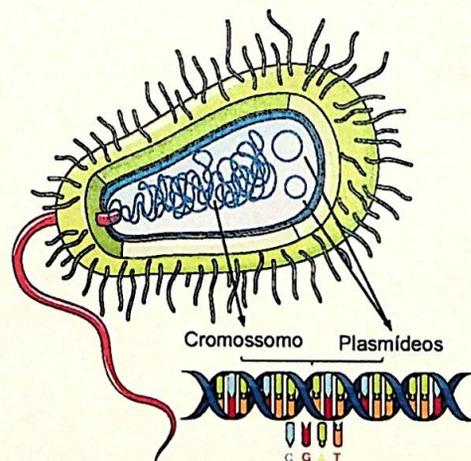


Figura 2.1: Ilustração de uma célula bacteriana genérica mostrando a presença de um cromossomo e de dois plasmídeos em seu citoplasma. O genoma, por ser a totalidade do material genético de um organismo, seria, neste caso, o conjunto dos três elementos genéticos presentes e sua base molecular é o DNA.

O genoma é composto por genes e elementos não codificantes. Saber a ordem dos nucleotídeos que o compõe se tornou fundamental em diversos ramos das ciências da vida, por exemplo, no presente trabalho o conhecimento do genoma irá permitir o conhecimento do conjunto de genes da bactéria *B. sacchari*, que por sua vez será importante para se inferir o conjunto de reações químicas que acontecem por meio das enzimas codificadas pelo conjunto de genes. Assim, é possível dividir o conhecimento do conteúdo genômico em dois aspectos: o primeiro é o conhecimento da ordem dos nucleotídeos e isso é possível através de técnicas de sequenciamento de DNA e de montagem. Já o segundo aspecto se refere ao conhecimento do teor informacional de uma sequência de DNA e isso

é obtido através de técnicas de anotação. Abaixo são detalhadas as etapas desses dois aspectos.

2.1.1 Inferência da ordem dos nucleotídeos no genoma

Desde os primeiros métodos de sequenciamento de DNA que surgiram na década de 1970, cujos exemplos incluem os trabalhos pioneiros de Sanger e Coulson (1975) e de Maxam e Gilbert (1977), até as inovações mais recentes, foi desenvolvida uma diversidade de tecnologias de sequenciamento: microeletroforese, sequenciamento por hibridação, observação em tempo real de moléculas isoladas, entre várias outras (Shendure e Ji, 2008). Entretanto, apesar desse amplo leque de opções desenvolvidas, todas elas compartilham da mesma limitação técnica: a incapacidade de determinar a ordem dos nucleotídeos de trechos muito longos de DNA como, por exemplo, de um cromossomo inteiro. De fato, todos os métodos de sequenciamento são restritos à determinação de fragmentos menores, cujos comprimentos dependem da tecnologia empregada (Loman e Pallen, 2015). Somente à caráter de ilustração, se fosse imaginado um espectro de variação de tamanho dos fragmentos, em um extremo estariam pequenos fragmentos de 35 pares de bases, caso da tecnologia *Genome Analyzer*, de 2008, da empresa *Illumina* (Hernandez *et al.*, 2008). No outro extremo estariam fragmentos relativamente bem maiores, da ordem de 15.000 pares de base, caso da tecnologia *PacBio RS*, de 2013, da empresa *Pacific Biosciences* (Heiner *et al.*, 2013).

Nesse contexto, para sequenciar o genoma completo de um organismo, é preciso contornar a limitação tecnológica apresentada. Para tanto, foi desenvolvida uma série de estratégias, incluindo o *primer walking*, o sequenciamento direcionado a sítios de restrição de enzimas e o sequenciamento *shotgun* (Simpson e Pop, 2015). Dentre elas, se destaca a estratégia *shotgun*, que se mostrou a maneira mais bem sucedidas de contornar a limitação técnica do sequenciamento (Kaiser *et al.*, 2003). Proposta no final dos anos 1970 por Staden (Staden, 1979), esta estratégia se baseia na fragmentação aleatória do genoma, resultando em inúmeros trechos que são, em seguida, sequenciados. Esses fragmentos sequenciados, chamados de *reads*, precisam ser então organizados mediante o uso de algoritmos computacionais para que possam revelar sequências contíguas a eles subjacentes, conhecidas por *contigs* (Miller *et al.*, 2010). A estratégia *shotgun* é mostrada nos itens (a) e (b) da figura 2.2.

O processo de organização dos *reads* até os *contigs* é referido na literatura por montagem dos *reads* e é classificado em duas categorias. O primeiro caso é quando já se dispõe de informações sobre a sequência do genoma de interesse. Nesse caso, a montagem dos *reads* se dá pelo seu alinhamento contra o genoma de referência. O segundo caso é a montagem *de novo* do genoma, ou seja, quando os *reads* são organizados sem o auxílio de informações prévias sobre o genoma de interesse (Kisand e Lettieri, 2013). Neste caso, a ordenação dos *reads* se dá por meio da mescla de sobreposições identificadas entre pares de *reads*, podendo ser vista com mais clareza nos itens (c) e (d) da figura 2.2. Apesar da classificação delineada, a montagem usando um genoma de referência e a montagem *de novo* não são abordagens mutuamente excludentes e ambas podem ser usadas no mesmo processo de montagem de *reads* (Pop, 2009).

2.1.2 Anotação do genoma

O processo de anotação visa atribuir informação biológica às sequências obtidas através da montagem do genoma. Este processo acontece em duas etapas sucessivas, ambas envolvendo o uso de métodos computacionais (Stein, 2001).

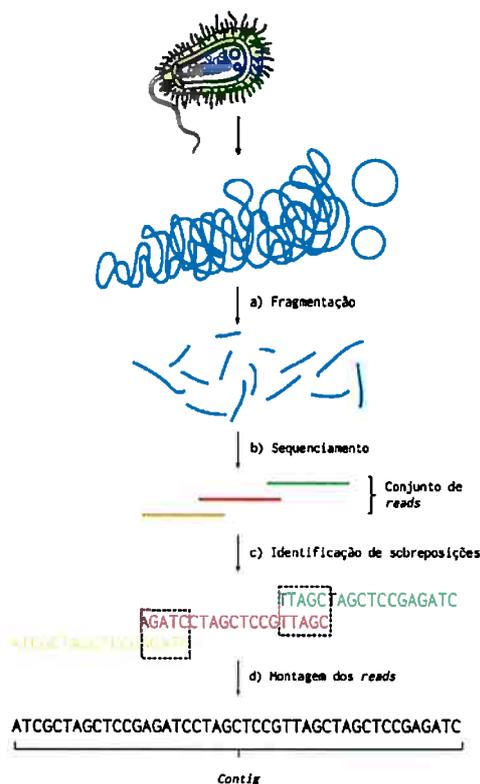


Figura 2.2: Visão geral das etapas envolvidas no sequenciamento shotgun e na montagem de novo de um genoma. Após a fragmentação do genoma (a), os fragmentos resultantes são sequenciados (b) para a inferência de suas sequências de nucleotídeos, sendo que essas sequências são chamadas na literatura de reads. Em seguida, algoritmos computacionais identificam sobreposições entre pares de reads (c) e os mesclam, fazendo com que sejam obtidas sequências maiores, chamadas de contigs (d). Adaptado de Shendure e Ji (2008) e Baker (2012)

A primeira etapa se refere à predição de genes no genoma, mas pode também envolver a predição de RNAs não codificantes como, por exemplo, o RNA ribossomal e o RNA transportador. Já a segunda etapa se refere à atribuição de função propriamente dita aos genes preditos na etapa anterior. Nesta etapa são atribuídos metadados às características identificados com base em métodos de homologia. Esses metadados podem ser informações a respeito da localização deste *feature* no genoma, a função que exerce ou até mesmo o número EC (*Enzyme Commission Number*) relacionado à enzima que o gene codifica.

Normalmente as duas etapas mencionadas estão disponíveis em uma única ferramenta que pode ser usada através de uma interface *web* ou por meio de uma aplicação local. Exemplos dessas ferramentas são: BaSYS (Van Domselaar *et al.*, 2005), IGS (Galens *et al.*, 2011), Prokka (Seemann, 2014) e RAST (Overbeek *et al.*, 2014).

2.2 O conceito de anotação multidimensional do genoma

A descrição da anotação de genomas no item anterior conceitualiza este processo como a catalogação de componentes celulares e a atribuição de metadados aos mesmos. O trabalho de Reed *et al.* (2006) vem complementar essa concepção da anotação ao propor o conceito de anotação multidimensional de genomas. Sua unidimensionalidade, segundo os autores, se deve ao fato de que esta anotação é uma lista de componentes celulares. No próximo nível da hierarquia, as anotações bi-

dimensionais levam em consideração as descrições sobre interações entre os componentes celulares. Assim, a anotação bidimensional não é somente uma lista de componentes, como a anotação unidimensional o é, mas sim uma rede de interação. Em razão dos diferentes tipos de interação que existem na célula, as anotações bidimensionais podem se referir a redes metabólicas, redes de sinalização e/ou redes de interação entre proteínas. Já anotações tridimensionais adicionam descrições espaciais à rede de interação entre os componentes da rede e, por último, anotações quadridimensionais levariam em conta também as mudanças que acontecem no genoma por meio do processo evolutivo. Entretanto, estes dois últimos níveis de anotação são atualmente apenas teóricos de acordo com Reed e colaboradores, pois atualmente só existem métodos e informações para se fazer as duas primeiras dimensões de anotação.

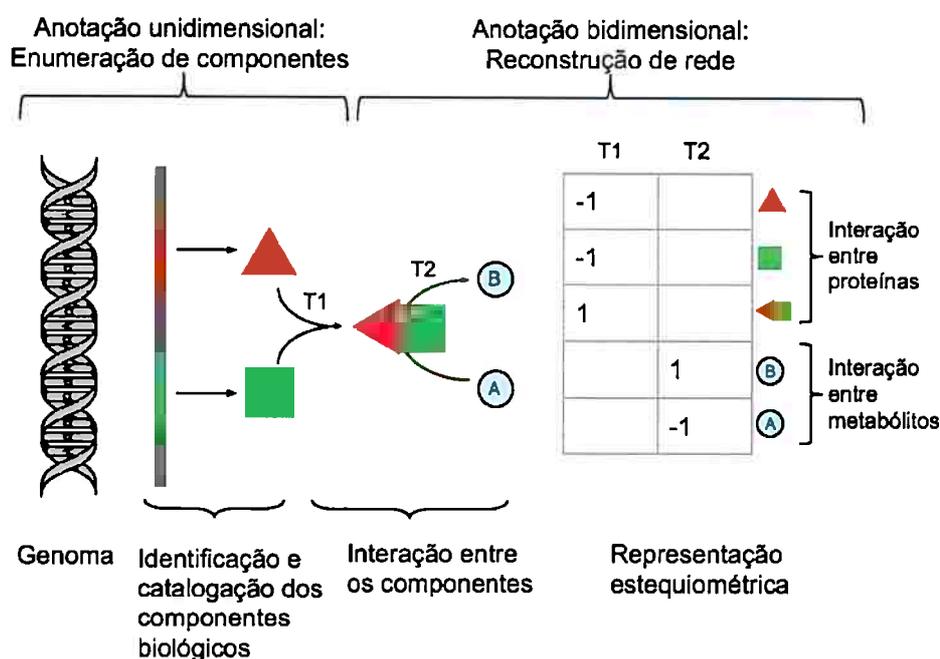


Figura 2.3: O conceito de anotação multidimensional. Adaptado de Reed et al. (2006)

2.2.1 Reconstrução metabólica

A reconstrução metabólica é uma instância de anotação bidimensional na qual a rede de interação que está sendo reconstruída é a rede metabólica. Assim, é necessário anteriormente conceituar o que são redes metabólicas: elas são uma representação de um conjunto de reações bioquímicas que acontecem em um organismo. As reconstruções metabólicas, por sua vez, são a representação de uma rede metabólica de uma forma estruturada e contendo metadados. Esses metadados podem ser informações genéticas e bioquímicas, por exemplo, para uma determinada reação, os metadados podem informar qual é a enzima que catalisa aquela reação. O formato geralmente usado para representar reconstruções metabólicas é a planilha eletrônica. A reconstrução é vista como uma base de conhecimento específica para um organismo. Dentro dessa base, há uma lista de reações bioquímicas e, por conseguinte, é possível representar essas reações num formato computacional e fazer simulações. A acurácia dessas simulações depende da qualidade das informações presentes no banco de dados e isso faz com que o processo de reconstruir a rede metabólica seja iterativo com o processo de se chegar a um modelo computacional preditivo para o funcionamento do metabolismo.

Modelos metabólicos em escala genômica

Entre os metadados presentes em uma reconstrução metabólica, estão informações sobre a estequiometria das reações. Desse modo, pode ser inferida a matriz estequiométrica desta reconstrução. As dimensões dessa matriz são m por n , onde m se refere ao número de metabólitos no sistema e n se refere ao número de reações. Já o conteúdo dessa matriz diz respeito à estequiometria de cada metabólito para cada reação da rede. Por exemplo, o elemento (i, j) desta matriz indica o coeficiente estequiométrico que o metabólito i apresenta na reação j .

A matriz estequiométrica é uma peça central nas análises do metabolismo, onde a maioria dos métodos de análise gravita em torno de uma única descrição matemática (Llaneras e Picó, 2008): a conservação de massa dos metabólitos internos de um sistema. Seja $x = (x_1, x_2, \dots, x_m)$ o vetor que se refere às concentrações dos metabólitos internos do sistema, na unidade (mol/L), $v = (v_1, v_2, \dots, v_n)$ o vetor que se refere aos valores dos fluxos nas reações do sistema, na unidade ($\text{mol} \cdot \text{L}^{-1} \text{h}^{-1}$) e S a matriz estequiométrica, a referida conservação de massa é representada pela equação 2.1:

$$\frac{dx}{dt} = S \cdot v \quad (2.1)$$

Com essa descrição matemática, é possível fazer simulações computacionais sobre metabolismo de modo a inferir as taxas de fluxo das reações. Essas simulações são muito importantes em engenharia metabólica pois o conhecimento dos fluxos permite saber, por exemplo, quanto a célula está consumindo de determinado substrato e quanto está produzido de determinado produto. Os métodos desenvolvidos para resolver a equação 2.1 se classificam, de acordo com Gornbert e Nielsen (2000) em modelos cinéticos e modelos estequiométricos. Os primeiros trabalham com a equação de conservação de massa, além de parâmetros cinéticos. O objetivo deste tipo de análise é determinar a relação entre os fluxos das reações e as concentrações internas dos metabólitos. A limitação desta abordagem é que a determinação experimental da concentração dos metabólitos é muito desafiadora. Já os modelos estequiométricos são mais difundidos pois usam algumas premissas que visam facilitar na modelagem. Nesses modelos, se considera estado estacionário, que é a situação onde não há variação na concentração dos metabólitos internos ao sistema. A suposição de estado estacionário também pode ser interpretada como a ausência de acúmulo de metabólitos internos. Matematicamente, a condição de estado estacionário é descrita pela equação 2.2:

$$\forall i \in \{1, \dots, m\}, \frac{dx_i}{dt} = 0 \Rightarrow S \cdot v = 0 \quad (2.2)$$

Os modelos estequiométricos partem da equação 2.2 e a abordam de diversas maneiras. Dentre elas, ganharam destaque a Análise de Balanço de Fluxo. Nesse contexto, a maioria dos autores considera modelo metabólico como sendo o formato computacional simulável que contém a matriz estequiométrica, além de algumas restrições que servem para a Análise de Balanço de Fluxo. Dentre estes autores, alguns defendem ainda que o modelo não basta ser simulável, mas também deve ser preditivo, como é o caso de Rolfsson e Palsson (2015).

Capítulo 3

Montagem e anotação do genoma de *B. sacchari*

3.1 Considerações Preliminares

O genoma de *Burkholderia sacchari* foi sequenciado com a estratégia *shotgun* pelo Laboratório de Bioprodutos da USP, que recorreu à empresa Macrogen para a prestação do serviço. A tecnologia de sequenciamento usada foi 454 GS FLX Titanium e o serviço como um todo compreendeu, além do sequenciamento propriamente dito, a montagem do conjunto de *reads* gerados. Pelo fato da montagem que foi entregue ter tido um caráter preliminar, onde foi usado um único *software* e com os parâmetros padrão, o presente capítulo visa olhar com mais cautela para este processo, tanto por meio de montagens mais robusta, quanto pela validação dessas montagens. Como a bactéria nunca havia sido previamente sequenciada, a montagem dos *reads* gerados tem que necessariamente seguir a abordagem *de novo*. Desse modo, abaixo é apresentada uma breve descrição sobre o estado da arte com respeito a esse tipo de montagem. Depois, será mostrada qual foi a estratégia adotada para *B. sacchari*, os métodos usados, e, finalmente, os resultados obtidos.

3.2 Revisão de Literatura

A montagem *de novo* de genomas sequenciados com a abordagem *shotgun* envolve, como dito no capítulo anterior, a junção de *reads* por meio da identificação de sobreposições em suas extremidades. Como são centenas de milhares ou milhões de *reads*, se faz necessário o uso de métodos computacionais para resolver a tarefa de montá-los para a obtenção do genoma - e essa tarefa se mostrou ser computacionalmente bastante desafiadora, com diversas formalizações apontando que o problema seja NP-completo Nagarajan e Pop (2009).

Uma das primeiras formalizações tratava a montagem de *reads* como uma instância do problema de se achar o *Shortest Common Superstring* (SCS). Essa abordagem visava encontrar a menor *string*¹ que contém todos os *reads* como *substrings*. Além do inconveniente de que o problema SCS ser NP-completo (Gallant *et al.*, 1980; Maier, 1978; Råihä e Ukkonen, 1981), uma limitação desta abordagem é sua premissa. Nela, é assumido que o genoma é o resultado do menor caminho a ser percorrido no grafo, entretanto, sabe-se que há muitos elementos repetitivos no genoma e a

¹*string*, em Ciências da Computação, se refere a uma sequência de caracteres

abordagem SCS, por otimizar o menor caminho, acaba por colapsar as repetições, enviando o genoma obtido.

Outras formalizações mais realistas, que levam em conta as repetições, foram então propostas. Elas se baseavam no uso de grafos, mais especificamente na construção da montagem como a busca por caminhos em um grafo formado a partir dos *reads* e serviram de base para a maioria dos *software* de montagem atuais. Entretanto, todas essas abordagens também foram provadas ser NP-completas (Medvedev *et al.*, 2007).

Apesar de tanto as formalizações SCS quanto as baseadas em grafos tenham se mostrado computacionalmente intratáveis, Nagarajan e Pop (2009) comenta que essas provas se basearam em casos de piores cenários possíveis, onde o tamanho do *read* é menor que o padrão da repetição, e argumentam que em cenários favoráveis, o problema da montagem pode ser devidamente tratável. Desse modo, a complexidade decorre da própria composição do genoma e dos *reads*. Em um trabalho posterior, os mesmos autores argumentam que dependendo da estrutura dos *reads*, a montagem pode variar de trivial, passando por computacionalmente intratável até impossível (Nagarajan e Pop, 2013). Entretanto, alegam os autores que casos triviais são raros e que na maioria da vezes a complexidade da montagem se faz presente.

A forma de resolver o impasse da complexidade da montagem para a maioria dos genomas (ou complexidade prática para Simpson e Pop (2015)), foi através do desenvolvimento de algoritmos baseados em heurísticas. Para o caso da formalização SCS, o método usado foi o guloso (*greedy*). Exemplos de montadores que usam esse método são, de acordo com Narzisi e Mishra (2011), TIGR, PHRAP, CAP3, entre outros. Com relação às formalizações baseadas em grafos, três métodos foram desenvolvidos, de acordo com Simpson e Pop (2015): Overlap-Layout-Consensus (OLC) cujas implementações são encontradas em CELERA, CABOG, ARACHNE; grafos de *de Bruijn*, cujos exemplos são Velvet e Euler e, por fim, *string graphs*, onde um representante é o Edena.

Esses métodos são úteis pois permitem chegar a uma solução aproximada quando o problema da montagem se revelar ambíguo, onde vários genomas são possíveis de serem reconstruídos a partir do mesmo conjunto de dados. Por outro lado, por serem aproximações, o resultado da montagem usando essas heurísticas pode ser fragmentado e/ou sujeito a erros Ghodsi *et al.* (2013). Além disso, de acordo com Howison *et al.* (2013), por ter que escolher deliberadamente uma solução entre uma diversidade de possibilidades, as heurísticas acabam por afunilar o espaço de soluções possíveis até “uma estimativa pontual da verdadeira sequência do genoma”, deixando fora desse espaço soluções que poderiam ser igualmente viáveis.

Se Howison *et al.* (2013) alerta para a multiplicidade de hipóteses à jusante do processo de montagem com algoritmos, Koren *et al.* (2014) alerta para a multiplicidade de hipóteses que podem ser geradas à montante: diante do fato de que há uma grande quantidade de montadores disponíveis para escolha, cada qual funcionando de acordo com um algoritmo, diante do fato de que dentro dos próprios montadores há uma série de parâmetros que podem ser modificados e, por último, diante do fato de que o conjunto de *reads* pode ser tratado de diversas maneiras, então é possível permutar os fatores supracitados de modo a obter uma quantidade quase ilimitada de genomas montados.

Diante dessa multiplicidade de hipóteses a respeito do genoma tanto à montante quanto à jusante da etapa de montagem, duas perguntas tem sido levantadas. Essas hipóteses seriam ao menos consistentes entre si? E se não, haveria alguma estratégia que levaria a uma melhor reconstrução do genoma? Dois estudos que repercutiram bastante na comunidade bioinformática ajudam a jogar luz

nessas questões: Assemblathon Earl *et al.* (2011) e GAGE Salzberg *et al.* (2012). Neles, a resposta para a primeira questão é que não há consistência nem entre montadores diferentes e até mesmo, em alguns casos, nem entre os resultados para o mesmo montador. A resposta à segunda questão é que não há uma estratégia única que seja a melhor. Cada montador seria bom em alguns aspectos, mas perderia em outros. A maior conclusão destes estudos é que não há uma receita para a montagem *de novo* de genomas. Foi a partir dessa conclusão que Koren *et al.* (2014) sugere que a montagem *de novo* de genomas deveria ser feita usando uma ampla variedade de estratégias e, então, elas deveriam ser avaliadas de acordo com algum critério de interesse para que seja eleita a montagem mais útil em relação à pergunta biológica que está sendo levantada.

Com relação à avaliação de montagens, os métodos e técnicas se dividem num sentido geral em dois grupos: métricas internas, que usam somente informações tomadas da própria montagem e métricas externas, que requisitam informações adicionais (Ekblom e Wolf, 2014).

As estatísticas padrão foram as primeiras métricas internas desenvolvidas para a avaliação das montagens e elas se baseiam em critérios de tamanho. São exemplos: N50, tamanho do maior *contig*, tamanho total da montagem, entre outras. Apesar de essas métricas ainda serem umas das mais utilizadas, elas não conseguem capturar a qualidade da montagem Ekblom e Wolf (2014); Nagarajan e Pop (2013); Narzisi e Mishra (2011); Vezzi *et al.* (2012a,b) e um exemplo disso é a conclusão do estudo de Vezzi *et al.* (2012a), que mostrou que a métrica N50 teve uma baixa correlação com qualidade dos *contigs montados*. Assim, as limitações dessas estatísticas padrão fez com que fossem desenvolvidas métricas internas mais robustas, como o uso de características para a detecção de inconsistências internas, que foi proposto por Phillippy *et al.* (2008).

Apesar dos avanços nas métricas internas, são as métricas externas que permitem uma avaliação mais acurada com respeito à qualidade da montagem. Nesse grupo se incluem o mapeamento contra um genoma de referência, o sequenciamento de transcriptoma, entre outros. Entretanto, Vezzi *et al.* (2012a) alertam que é preciso ter cautela com as métricas externas e argumentam, como exemplo, sobre as limitações da métrica de alinhar os *contigs montados* contra um genoma de referência e contar o número de *misassemblies*.

3.3 Estratégia adotada para *B. sacchari*

Embora Koren *et al.* (2014) está se baseando em estudos feitos usando dados da tecnologia Illumina para propor sua sugestão (de montar o genoma usando uma série de estratégias e depois selecionar a melhor baseada em critérios de interesse), ela ainda continua válida para dados 454 ao se levar em conta estudos que comparam montadores específicos para dados dessa tecnologia. Um desses exemplos é o trabalho de Finotello *et al.* (2012), que compara cinco montadores tanto com relação à métricas internas quanto com relação à métricas externas e conclui que não houve nenhum *software* cujo desempenho foi muito melhor do que os outros. Ao contrário, eles apresentaram “diferentes características em termos de acurácia e completude da montagem”.

Assim, será usada a sugestão de Koren e colaboradores neste trabalho, onde o genoma será montado usando vários *software* e diferentes tratamentos do conjunto de *reads* bruto. Então, os genomas montados serão sujeitos a avaliações com vistas a escolher a melhor montagem.

A escolha dos montaremos foi pautada em revisões já feitas acerca de montadores úteis para pirosequenciamento e a escolha do método de tratamento dos *reads* foi pautada na estrutura e

qualidade do conjunto de *reads* e na qualidade do sequenciamento usado. Por fim, a escolha das métricas foi pautada em métricas internas, externas e também em métricas sobre o da anotação dos genomas montados uma vez que isso tem íntima relação com a qualidade da reconstrução e justifica o motivo de montagem e anotação estarem juntas neste capítulo.

3.4 Materiais e Métodos

3.4.1 Obtenção e descrição do conjunto de *reads*

O conjunto bruto de *reads* foi obtido através do sequenciamento *de novo* do genoma de *B. sacchari* usando o sequenciador 454 GS FLX Titanium (Alexandrino *et al.*, 2015). Os dados brutos do sequenciamento estão disponíveis em um arquivo no formato SFF (*Standard Flowgram Format*) e foram depositados na base de dados *Sequence Read Archive*, do *National Center for Biotechnological Information*. Nesta base de dados, os *reads* podem ser acessados através do identificador SRX736502.

Após o sequenciamento de um genoma é bastante recomendado a análise do conjunto de *reads* brutos por *software* de controle de qualidade (Edwards e Holt, 2013). Estes *software* descrevem o conjunto de *reads* gerados por meio de gráficos e estatísticas como, por exemplo, a distribuição do valor de qualidade para cada posição do *read*, a presença de adaptadores indesejados, o conteúdo GC, entre outras funcionalidades. Com os resultados dessa descrição é possível inferir a qualidade do sequenciamento feito e isso permitiria o posterior desenho experimental da estratégia de pré-processamento dos conjunto de *reads* antes da montagem. Assim, com relação ao controle de qualidade, foram usados dois *software*: o FASTQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>) e o Prinseq (Schmieder e Edwards, 2011). Este foi usado através de seu servidor *web* no endereço <http://prinseq.sourceforge.net/> e aquele foi usado por intermédio da plataforma Galaxy (Blankenberg *et al.*, 2010; Giardine *et al.*, 2005; Goecks *et al.*, 2010), que é um conjunto de ferramentas e algoritmos de bioinformática que integra, entre outros, o FASTQC.

Uma vez que tanto o FASTQC quanto o Prinseq exigem que os dados estejam em formato FASTQ e como os dados brutos gerados pelo sequenciador 454 estavam em formato SFF (*Standard Flowgram Format*), foi necessário convertê-los ao formato adequado. Para isso foi usada a ferramenta *SFF converter*, disponível dentro do próprio Galaxy, com parâmetro para não modificar os dados. O arquivo FASTQ gerado pelo Galaxy foi usado internamente pelo FASTQC, enquanto que para o Prinseq, o arquivo precisou ser exportado para seu servidor.

Com os *softwares* de controle de qualidade foi possível constatar que os dados brutos continham 785.669 *reads*, que somavam um total de 395.269.906 pares de base. A média de tamanho do conjunto de *reads* é de 503,10 pb, com desvio padrão de 39,83 pb. O menor e o maior *read* continham, respectivamente, 53 e 1.200 pb. Essas informações a respeito da distribuição de tamanho dos *reads* podem ser vistas na figura 3.1.

Além disso, observa-se que o conjunto de *reads* apresenta uma super-representação dos nucleotídeos GACT em sua extremidade 5', mostrada na figura 3.2. Entretanto, foi verificado que esses nucleotídeos são inerentes à própria tecnologia 454 e servem somente para calibrar o sinal de luz que a reação de polimerização de DNA emite. Por ser uma sequência exógena, a maioria dos montadores específicos para o pirosequenciamento consegue identificar e eliminar estes quatro nucleotídeos devidamente.

- Newbler
- CAP3
- MIRA
- SeqMan
- CLC Assembly Cell

Dentre estes sete montadores levantados, o CLC e o SeqMan, por serem acessíveis mediante pagamento, não foram selecionados para uso neste trabalho. Além disso, PCAP454 não foi encontrado e a referência que o citava, no trabalho de Finotello *et al.* (2012), estava desatualizada ². Os montadores restantes CABOG, MIRA, Newbler e CAP3 se mostraram viáveis e, por isso, foram os escolhidos para a montagem do genoma de *B. sacchari*.

3.4.3 Seleção da estratégia de pré-processamento dos *reads*

As estratégias usadas para o pré-processamento do conjunto de *reads* são divididas, de modo geral, em remoção de trechos de baixa qualidade nas extremidades destes fragmentos e em eliminação de *reads* (Schmieder e Edwards, 2011). Esta última estratégia pode ainda ser dividida em eliminação de *reads* seguindo algum critério de interesse (por exemplo eliminar *reads* com relação ao tamanho, número de N's, complexidade, entre outros), onde o objetivo é filtrar sequências de baixa qualidade, e em eliminação de *reads* aleatórios, onde o objetivo é realizar o *downsample* do conjunto de *reads*. Nesse contexto, levando em consideração a descrição que foi feita do conjunto de *reads* de *B. sacchari*, foram desenhadas as estratégias de pré-processamento a seguir. Para cada uma das três estratégias apresentadas são apresentadas justificativas acerca das escolhas que foram tomadas.

Remoção de trechos nas extremidades dos *reads*

Foi constatado que todos os quatro algoritmos de montagem selecionados realizam automaticamente a remoção de extremidades de baixa qualidade seguindo critérios próprios. Assim, optou-se por deixar que cada algoritmo tomasse sua decisão com respeito à estratégia de remoção de extremidades. No entanto, também foi constatado que o CAP3 foi o único montador que não consegue reconhecer GACT como um adaptador 454 e, por isso, não remove esta sequência da extremidade 5' dos *reads*. Diante disso, foi preciso remover a sequência GACT dos conjunto de *reads* bruto antes da aplicação do montador CAP3 e, para tanto, foi usado o programa *Trim sequences*, disponível através da plataforma *Galaxy*. Já para os outros montadores, CABOG, Newbler e MIRA, foram usados brutos sem modificação.

Eliminação de *reads* de baixa qualidade

Há dois artigos na literatura que estudam os fatores que influenciam a qualidade dos *reads* sequenciados com tecnologia 454 e esses artigos se mostraram bastante informativos para a escolha da estratégia de eliminação de *reads* baseadas em critérios. De acordo com Huse *et al.* (2007), é uma pequena quantidade de *reads* de baixa qualidade que contém a maioria dos erros encontrados

² A referência é a página *web* <http://www.medcomp.medicina.unipd.it/pcap454.html>

no sequenciamento 454 GS++. Além disso, a maioria desses erros encontrados foi associada a *reads* cujo tamanho se afastava consideravelmente da média e *reads* que continham a presença de até mesmo uma única base N. Assim, o autor argumenta que é necessário eliminar *reads* com tais características para a obtenção de um conjunto de *reads* com maior qualidade.

Entretanto, o estudo mais recente de Gilles *et al.* (2011) argumenta que o padrão de erros da tecnologia 454 GS FLX Titanium é diferente da tecnologia 454 GS++ avaliada por Huse e colaboradores. Ao avaliar somente erros exclusivos para 454 GS FLX Titanium, Gilles e colaboradores asseveram que para esta tecnologia a presença de erros está distribuída mais uniformemente e na média há poucos erros por *read*. Além disso, os autores atestam que no contexto da tecnologia 454 GS FLX Titanium, a remoção de *reads* com erros pode não ser efetivo na melhora da qualidade global da montagem.

Com relação ao exposto, optou-se por não realizar a eliminação de *reads* de baixa qualidade.

Eliminação de *reads* aleatórios

No estudo comparativo entre montadores feito por Finotello *et al.* (2012), os autores fazem o *downsample* do conjunto de *reads* da seguinte maneira: o dado bruto tinha 72× de cobertura e foi reduzido até 4× por meio de intervalo de 4×, isto é, por meio de reduções sucessivas que subtraíam 4× de cobertura a cada passo. No total, os autores obtiveram 18 subconjuntos a partir do original.

No caso do presente trabalho com *B. sacchari* optou-se por uma distância menor entre o conjunto de *reads* de menor cobertura e o maior. Essa opção foi baseada no manual de instruções do montador MIRA, onde é sugerido que a cobertura seja mantida entre 18 e 60× de cobertura para dados 454. Este manual se mostrou a única fonte de informações, entre a literatura consultada, a respeito de uma faixa de valores de cobertura ideais.

O dado bruto de *B. sacchari* tinha 49× de cobertura e foi filtrado para seis conjuntos de *reads*:

- 90% do total de *reads*, equivalente a 44,57× de cobertura;
- 80% do total de *reads*, equivalente a 39,62× de cobertura;
- 70% do total de *reads*, equivalente a 34,67× de cobertura;
- 60% do total de *reads*, equivalente a 29,71× de cobertura;
- 50% do total de *reads*, equivalente a 24,76× de cobertura;
- 40% do total de *reads*, equivalente a 19,80× de cobertura;

Como o Newbler e o CABOG³ requerem o formato de entrada SFF e o MIRA e o CAP3 requerem o formato FASTQ, duas abordagens foram adotadas e são para se fazer cumprir o *downsample* proposto: para o Newbler e o CABOG, usou-se o programa *sff_extract* (disponível na própria suíte de aplicativos do Newbler) para filtrar *reads* do arquivo SFF. Já no caso de MIRA e CAP3, usou-se o programa *seqtk* (<https://github.com/lh3/seqtk>). Essas duas abordagens são mostradas na figura 3.3.

³O CABOG, na verdade, não exige diretamente o arquivo SFF, mas sim o arquivo no formato FRG(frag). Entretanto o CABOG possui o algoritmo *sffToCA* que converte o SFF para FRG.

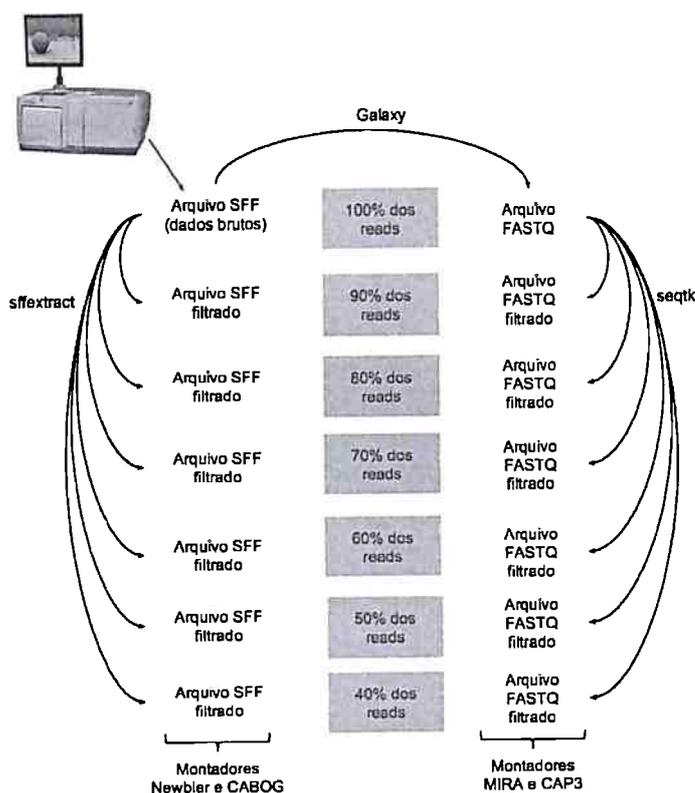


Figura 3.3: Tratamento dado ao conjunto de reads. Os dados brutos gerados pelo sequenciamento 454 estão formatados em um arquivo SFF. A filtragem desse conjunto de reads bruto foi feita de acordo com os valores da coluna do meio da figura e precisou ser feita de duas maneiras diferentes. Para os montadores Newbler e CABOG usou-se a estratégia mostrada na coluna da esquerda, onde o arquivo bruto SFF foi filtrado pelo programa *sffextract*. Para os montadores MIRA e CAP3 usou-se a estratégia da direita, onde o arquivo bruto SFF foi convertido ao formato FASTQ através da plataforma Galaxy e este arquivo FASTQ foi filtrado com o uso do programa *seqtk*.

3.4.4 Montagem do genoma usando os algoritmos escolhidos

A montagem com o algoritmo Newbler foi feita usando o programa *runAssembly* da suíte de aplicativos *GS Data Analysis Software*, disponível através do endereço <http://454.com/products/analysis-software/>. Os parâmetros usados foram os padrão do programa. Para o montador MIRA, a montagem foi baseada nas instruções presentes em seu manual, disponível em <http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html>. O arquivo manifesto exigido para este montador foi configurado com os seguintes valores para o parâmetro *job*: *genome*, *de novo* e *accurate*. Todos os outros parâmetros foram mantidos como padrão. Com relação ao montador CABOG, cada arquivo SFF obtido através do processo de *downsample* precisou ser convertido ao formato FRG através do programa *sffToCA*, disponível no endereço eletrônico <http://wgs-assembler.sourceforge.net/wiki/index.php/SffToCA>. Os parâmetros usados foram os padrões. Finalmente, o montador CAP3 foi obtido através do endereço <http://seq.cs.iastate.edu/cap3.html>. De acordo com a documentação presente no endereço mencionado, a montagem deve ser feita usando arquivos no formato FASTA separados dos arquivos de qualidade QUAL. Desse modo, recorreu-se à plataforma Galaxy para transformar cada um dos sete arquivos FASTQ obtidos nas etapas de tratamento em pares de arquivos FASTA e QUAL. Com relação à montagem, todos os parâmetros usados foram os padrões do algoritmo.

3.4.5 Anotação das montagens

A anotação das montagens foi feita com a ferramenta RAST (Aziz *et al.*, 2008; Brettin *et al.*, 2015; Overbeek *et al.*, 2014). O arquivo FASTA gerado pelos montadores foi submetido na plataforma web com todos os parâmetros padrão, com exceção do *Taxonomy ID*, que é específico para o organismo em questão e que para *Burkholderia sacchari* é 159450.

3.4.6 Avaliação dos genomas

Métricas padrão de montagem

As montagens foram avaliadas quanto à quanto às estatísticas padrão foi feita com a ferramenta QUAST (Gurevich *et al.*, 2013), através de sua aplicação pela linha de comando. Dentre as várias métricas disponibilizadas pelo QUAST, foram selecionadas as seis métricas a seguir: número de *contigs* gerados pelas montagens, número de *contigs* maiores de 1.000 pb, tamanho total da montagem, tamanho total da montagem considerando *contigs* maiores de 1000 pb, tamanho do maior contig e N50. Esta última se refere ao tamanho no qual a soma de todos os *contigs* que são maiores ou iguais a este tamanho representa pelo menos 50% do tamanho total da montagem. Como já foi mencionado na Revisão de Literatura, métrica de tamanho podem não ter boa correlação com a qualidade (Vezzi *et al.*, 2012a) e, por isto, a avaliação dessas métrica teve um caráter exploratório e eliminatório.

Alinhamento com sequências conhecidas

Para validar as montagens, foram alinhadas todas as sequências disponíveis para *B. sacchari* no NCBI contra cada uma das montagens. As sequências em questão são descritas na tabela 3.1.

Nome	Nº de acesso NCBI	Referência
Burkholderia sacchari strain IPT10 16S ribosomal RNA gene, partial,sequence	NR_025097	Brämer <i>et al.</i> (2001)
Burkholderia sacchari strain IPT101 xylose isomerase gene, complete cds	FJ374785	Lopes <i>et al.</i> (2009)
Burkholderia sacchari prp gene cluster, complete sequence	AY033092	Bramcr <i>et al.</i> (2002)
Burkholderia sacchari strain IPT101 ATP synthase beta chain (atpD) gene, partial cds	HQ398450	Estrada-de los Santos <i>et al.</i> (2013)
Burkholderia sacchari strain IPT101 glutamate synthase large subunit (gltB) gene, partial cds	HQ398498	Estrada-de los Santos <i>et al.</i> (2013)
Burkholderia sacchari strain IPT101 GTP binding protein-like (lepA) gene, partial sequence	HQ398545	Estrada-de los Santos <i>et al.</i> (2013)
Burkholderia sacchari strain IPT101 recombinase A-like (recA) gene, partial sequence	HQ398592	Estrada-de los Santos <i>et al.</i> (2013)
Burkholderia sacchari strain LMG 19450 DNA gyrase subunit B (gyrB) gene, partial cds	HQ849212	Lcmaire <i>et al.</i> (2012)

Tabela 3.1: Sequências de nucleotídeos disponíveis no NCBI para *Burkholderia sacchari*

O método de alinhamento usado foi o BLAST (*Basic Local Alignment Search Tool*), através de sua aplicação pela linha de comando. Para alinhar as sequências escolhidas contra as montagens,

foi preciso indexar todos os conjuntos de *contigs*.

Estatísticas da anotação e comparação do conteúdo gênico

A avaliação das montagens se deu por meio do total de *features* obtidos. Esses *features* se referem ao conjunto de genes que codificam tanto para proteínas quanto para RNAs. Essa informação é disponibilizada pelo próprio anotador RAST. Além disso, o anotador faz a inferência do provável número de genes faltantes a determinada anotação com base na comparação com as anotações em seu banco de dados. Essa informação também foi usada como critério de avaliação. Com relação ao conteúdo das anotações, foram feitas comparações par a par de todas as anotações para obter os genes que estavam presentes em uma, mas que não o estavam na outra. Para isso, foi usado o programa *Function Based Comparison* do RAST.

3.5 Resultados e Discussão

3.5.1 Estatísticas das montagens

A montagem resultou nas estatísticas mostradas na tabela 3.2.

Tabela 3.2: Métricas de tamanho, obtidas com a ferramenta *QUAST*, para cada um dos 28 tratamentos aplicados às montagens.

Montador	Tratamento	Nº de contigs (≥ 0 bp)	Nº de contigs (≥ 1000 bp)	Tamanho total (≥ 0 pb, Mpb)	Tamanho total (≥ 1000 pb, Mpb)	Tamanho do maior contig	N50
MIRA	40	286	159	7,42	7,34	601.926	142.362
	50	273	166	7,41	7,35	819.183	189.313
	60	272	152	7,41	7,35	470.080	185.597
	70	250	152	7,40	7,35	542.291	189.369
	80	245	124	7,40	7,33	685.421	189.991
	90	270	145	7,41	7,35	686.205	189.308
	100	292	150	7,42	7,35	686.322	189.355
CABOG	40	87	87	7,16	7,16	570.114	166.060
	50	80	80	7,16	7,16	569.773	191.037
	60	86	86	7,17	7,17	569.808	188.353
	70	81	81	7,16	7,16	569.803	170.493
	80	82	82	7,16	7,16	569.808	188.353
	90	84	84	7,16	7,16	569.809	170.765
	100	86	86	7,17	7,17	569.809	170.765
Newbler	40	118	92	7,25	7,24	569.723	195.614
	50	116	84	7,25	7,24	603.044	209.081
	60	114	78	7,25	7,24	603.061	214.207
	70	115	81	7,26	7,24	603.291	215.069
	80	121	80	7,26	7,24	603.370	208.943
	90	125	84	7,26	7,24	603.370	207.931
	100	121	79	7,26	7,24	603.119	208.943
CAP3	40	804	293	7,78	7,45	184.123	48.282
	50	894	261	7,86	7,46	343.185	71.142
	60	1.068	271	7,97	7,47	425.804	84.771
	70	1.191	264	8,06	7,47	216.046	68.857
	80	1.327	256	8,16	7,48	462.612	83.853
	90	1.457	282	8,27	7,51	352.436	85.116
	100	1.692	359	8,51	7,58	220.836	66.466

Com relação ao número de contigs, usualmente se usa a premissa de quanto menor o número de contigs, melhor seria a montagem. Assim, analisando todos os contigs, o CABOG se sai melhor, seguido de Newbler, MIRA e CAP3. Nota-se uma fragmentação muito grande da montagem feita por esse último montador. Entretanto, pelo fato de que o CABOG apresentou o mesmo número de *contigs* tanto para maiores de 0 pb quanto para maiores de que 1000 pb, nota-se que o montador filtrou automaticamente, em razão dos parâmetros padrão usados, os *contigs* abaixo de 1000 pb. Assim, é mais razoável comparar os *contigs* maiores de 1000 pb para todas as montagens e essa comparação mostra que o Newbler passa o CABOG, ficando em primeiro lugar. MIRA continua em terceiro e CAP3 fica em último, ainda fragmentando bastante, mas não tanto como antes.

O número de *contigs*, por sua vez, tem íntima relação com o tamanho total da montagem. Assim, também é viável comparar somente os tamanhos totais quando se consideram os *contigs* maiores ou iguais a 1000 pb. As maiores montagens são de CAP3 e MIRA, nesta ordem, e este fato se mostra um reflexo de que os dois montadores também foram aqueles que produziram mais *contigs*. Em se tratando dos outros montadores, CABOG produziu as menores montagens e Newbler ficou na terceira posição.

Com relação ao maior *contig*, o MIRA se mostrou o montador capaz de produzir os maiores contigs, seguido por Newbler, CABOG e CAP3. E, por último, sobre o N50, o Newbler se mostrou o montador cujas montagens tem o maior N50, seguido de CABOG, MIRA e CAP3. Entretanto, N50 tem relação com o tamanho total da montagem e este, como dito, tem relação com o número de contigs obtidos. Assim, a plotagem cumulativa é uma alternativa para tirar o viés do número de *contigs*. Por meio desta plotagem, mostrada na figura 3.4, é possível visualizar como o incremento do tamanho total da montagem quando se leva gradativamente mais *contigs* em consideração. Nela, ainda se mantém o distanciamento observado do CAP3 para com os demais montadores.

Somadas todas essas observações, o panorama que se tem é que Newbler, CABOG e MIRA O CAP3 apresentam diferentes pontos fortes. Newbler fica em primeiro em mais quesitos, CABOG na maioria deles e MIRA se destacou na métrica de maior *contig*. Por outro lado, CAP3 apresentou os piores resultados em todas as métricas analisadas. Desse modo, optou-se por excluí-lo das etapas subsequentes de avaliação.

3.5.2 Alinhamento das montagens com sequências conhecidas

Das oito sequências de nucleotídeos alinhadas, sete delas se alinharam por inteiro em todos os 21 conjuntos de *contigs* obtidos nas montagens, após a eliminação do montador CAP3. Naturalmente, o alinhamento teve alguns *mismatches*, mas não foi esse o fator mais relevante. Por outro lado, a sequência NR_025097, do gene 16S, foi a responsável por alinhamentos que diferiram consideravelmente entre os montadores. O resultado é mostrado nas figuras 3.5, 3.6 e 3.7. Nestas figuras, a linha horizontal superior de cada subfigura representa o gene 16S, que conta com 1.523 pares de base. E a linha inferior representa o *contig* cujo alinhamento ao gene 16S se mostrou o mais extenso dentre todos os alinhamentos encontrados com a ferramenta BLAST.

O MIRA se destacou pois todos os alinhamentos continham o gene 16S por inteiro. O Newbler apresentou três dos oito alinhamentos por inteiro e o CABOG não apresentou nenhum. Outro destaque para o MIRA foi que, comparando aos alinhamentos inteiros do Newbler, o do MIRA se encontrou sempre em um fragmento de 5000 pares de base, enquanto que o do Newbler estava em um fragmento de 1.200 pares de base.

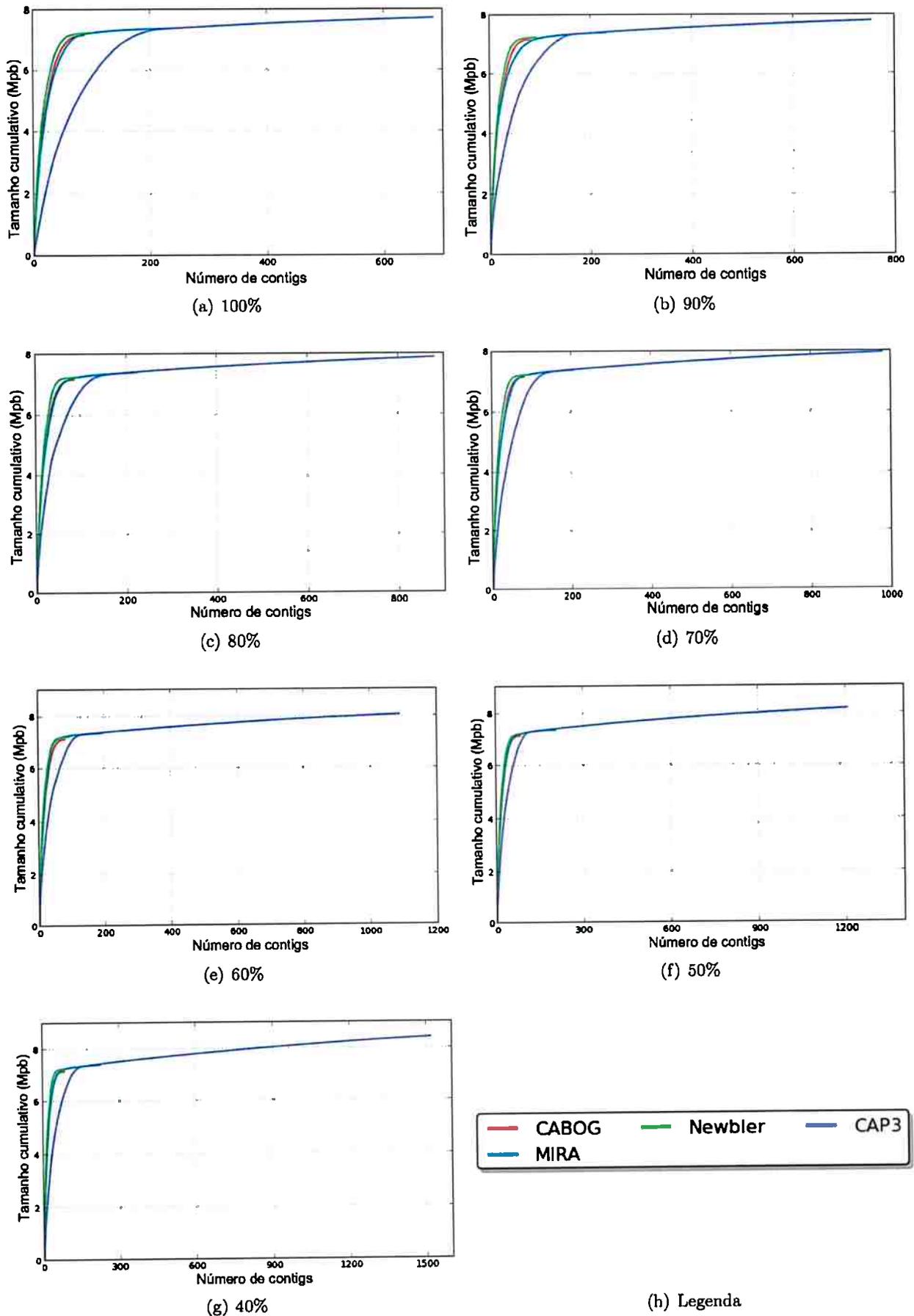


Figura 3.4: Aumento cumulativo do tamanho da montagem conforme o incremento do número de contigs para cada um dos tratamentos dados ao conjunto de reads.

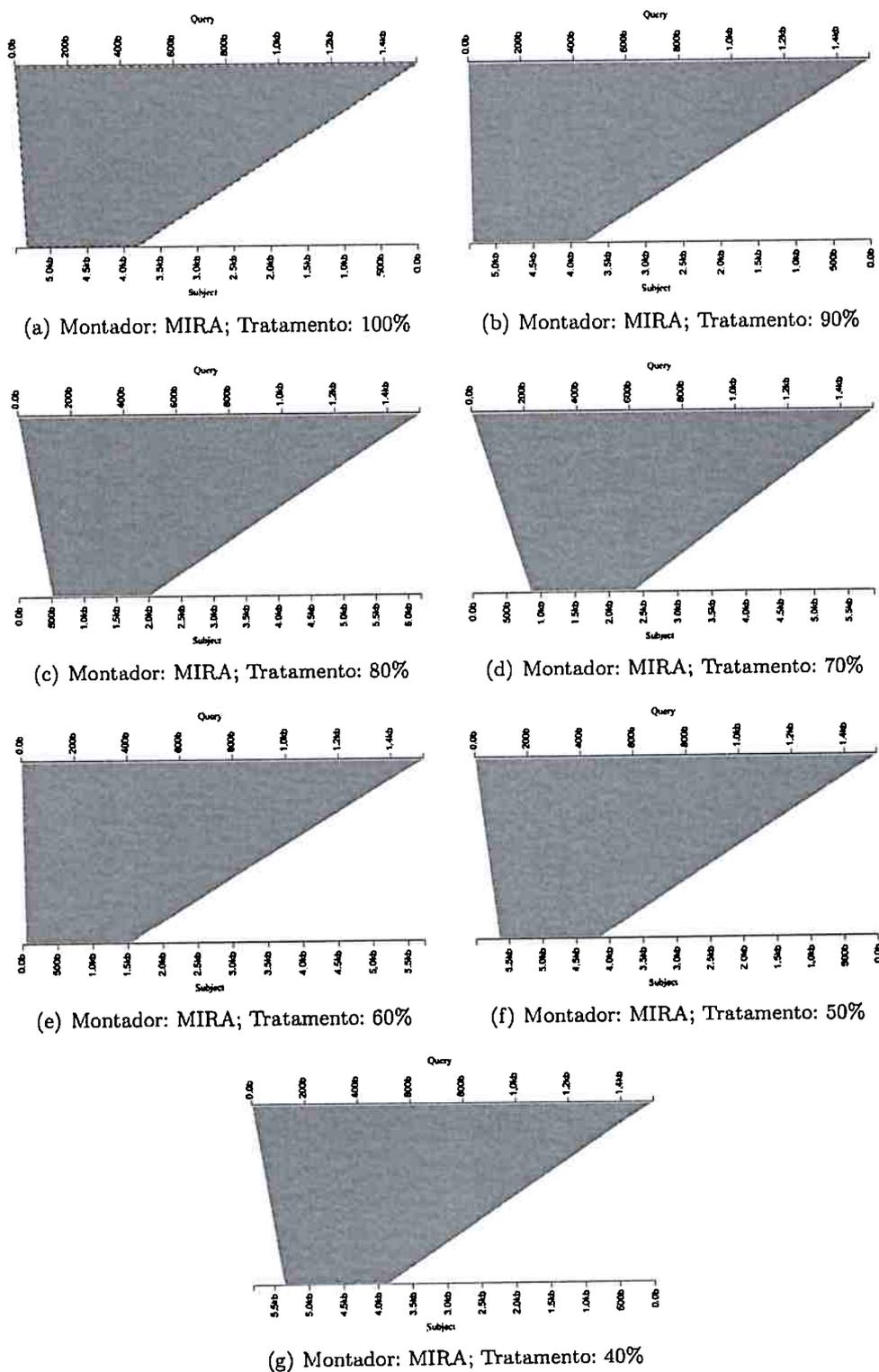


Figura 3.5: Resultado do alinhamento entre o gene 16S e o conjunto de contigs de cada montagem que usou o montador MIRA. Cada uma das figuras a-g representa um tratamento dado ao conjunto de reads e, nelas, a linha horizontal superior se refere ao gene 16S, enquanto que a linha inferior se trata do contig que apresentou o maior alinhamento dentre todos os obtidos. A visualização de todos os alinhamentos foi feita usando a ferramenta Kablammo.

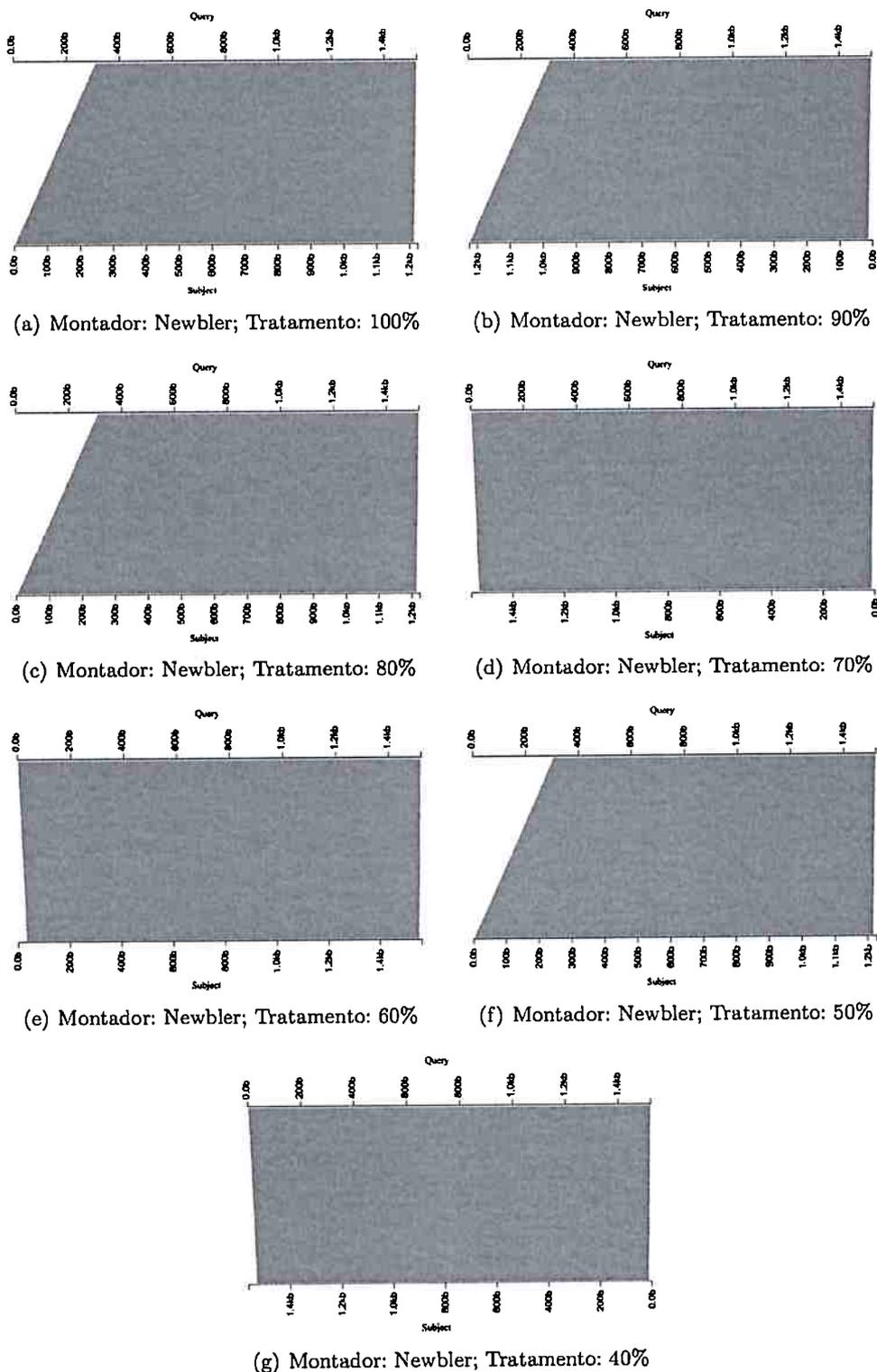


Figura 3.6: Resultado do alinhamento entre o gene 16S e o conjunto de contigs de cada montagem que usou o montador Newbler. Cada uma das figuras a-g representa um tratamento dado ao conjunto de reads e, nelas, a linha horizontal superior se refere ao gene 16S, enquanto que a linha inferior se trata do contig que apresentou o maior alinhamento dentre todos os obtidos. A visualização de todos os alinhamentos foi feita usando a ferramenta Kablammo.

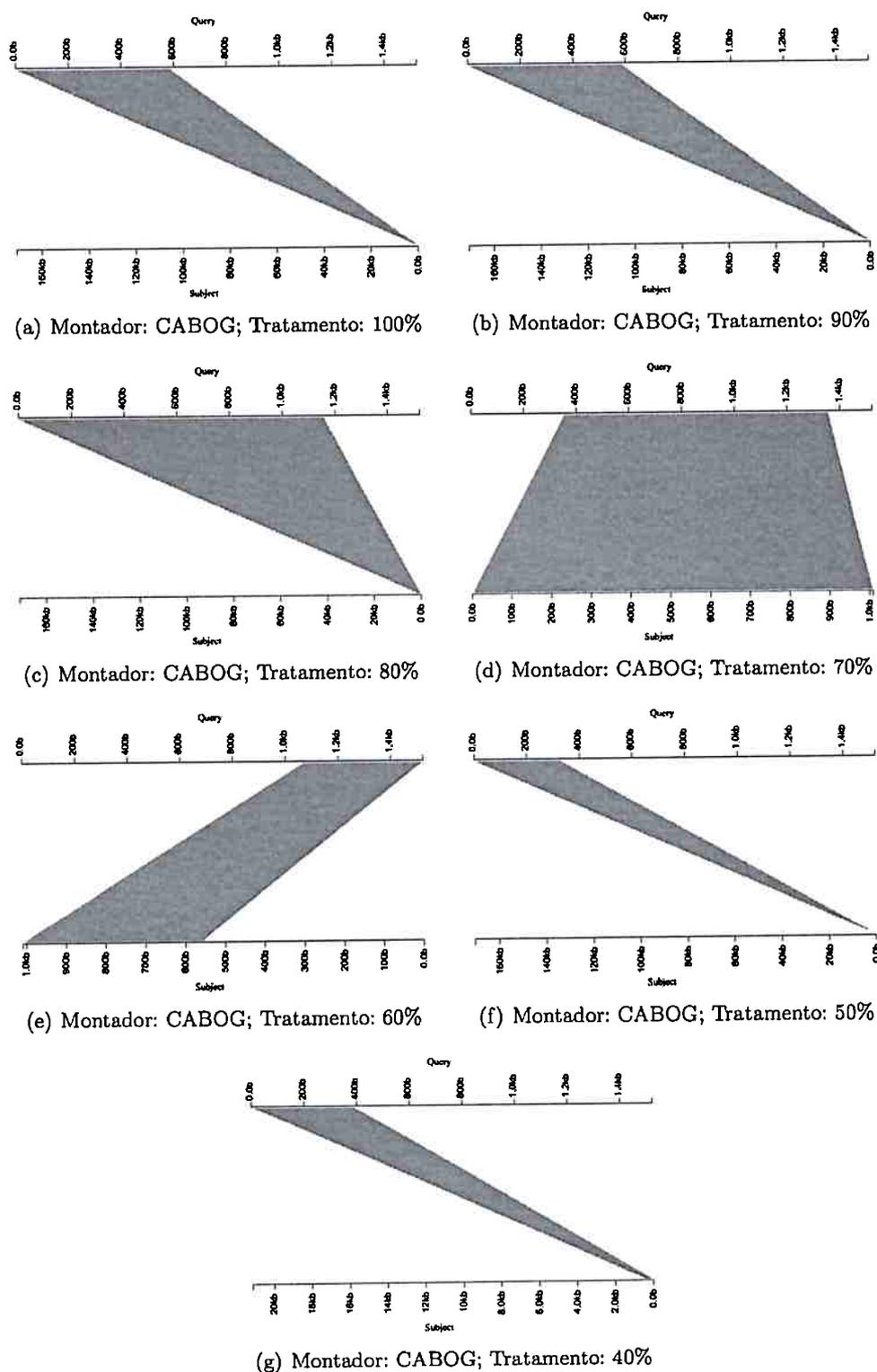


Figura 3.7: Resultado do alinhamento entre o gene 16S e o conjunto de contigs de cada montagem que usou o montador CABOG. Cada uma das figuras a-g representa um tratamento dado ao conjunto de reads e, nelas, a linha horizontal superior se refere ao gene 16S, enquanto que a linha inferior se trata do contig que apresentou o maior alinhamento dentre todos os obtidos. A visualização de todos os alinhamentos foi feita usando a ferramenta Kablammo.

3.5.3 Estatísticas das anotações

As anotações com o RAST resultaram nos dados mostrados na tabela 3.3. Nela é possível observar que a montagem com o MIRA resultou em anotações com mais *features*. Isso pode ter relação com a maior tamanho que da montagem doas MIRAs. Apesar de ter mais *features*, o MIRA teve mais prováveis genes ausentes.

Tabela 3.3: Estatísticas das anotações.

Montador	Tratamento	Número de features	Número de possíveis genes ausentes
MIRA	40%	6971	79
	50%	6.929	73
	60%	6.935	73
	70%	6.931	67
	80%	6.931	69
	90%	6.931	80
	100%	6.933	77
CABOG	40%	6.713	61
	50%	6.708	63
	60%	6.721	63
	70%	6.706	64
	80%	6.718	69
	90%	6.708	69
	100%	6.703	69
Newbler	40%	6.776	73
	50%	6.770	74
	60%	6.774	68
	70%	6.770	77
	80%	6.772	77
	90%	6.780	77
	100%	6.775	78

3.5.4 Comparação do conteúdo gênico

As comparações, com respeito ao conteúdo de *features*, entre pares de anotação resultou nas informações sumarizadas nas figuras 3.8 e 3.9. Ambas as figuras dizem respeito ao número de *features* que uma determinada anotação contém em relação à anotação com que está sendo comparada. Entretanto, a diferença entre as figuras é que para a primeira, são considerados todos os *features* anotados, e para a segunda, são considerados somente os *features* que dispunham de número EC e/ou que se relacionavam a mecanismos de transporte.

Antes do detalhamento dos resultados, é importante pontuar como as figuras 3.8 e 3.9 devem ser interpretadas. Para isso, o exemplo a seguir pode ser ilustrativo. Tomando a figura 3.8 como o exemplo, em sua primeira linha e segunda coluna há o número 1. Este valor deve ser interpretado como a quantidade de *features* que são exclusivos da anotação da montagem 100_MIRA (montagem

que usou o software MIRA com 100% do conjunto de reads) quando esta é comparada com a montagem 90_MIRA.

Desse modo, como é vantajoso para os próximos passos da reconstrução metabólica que o conteúdo da anotação do genoma seja o mais completo possível, seria interessante buscar, nas figuras mencionadas, por colunas em cujas células estão os menores valores possíveis.

O panorama que se observa é que, para a figura 3.8, as montagens MIRA 50 e 60 se destacaram pois suas colunas foram as que possuíam os menores resultados. Já com relação à figura 3.9, se destacaram as mesmas anotações usando o MIRA, além de quatro anotações das montagens usando o CABOG: 100_CABOG, 90_CABOG, 70_CABOG e 60_CABOG.

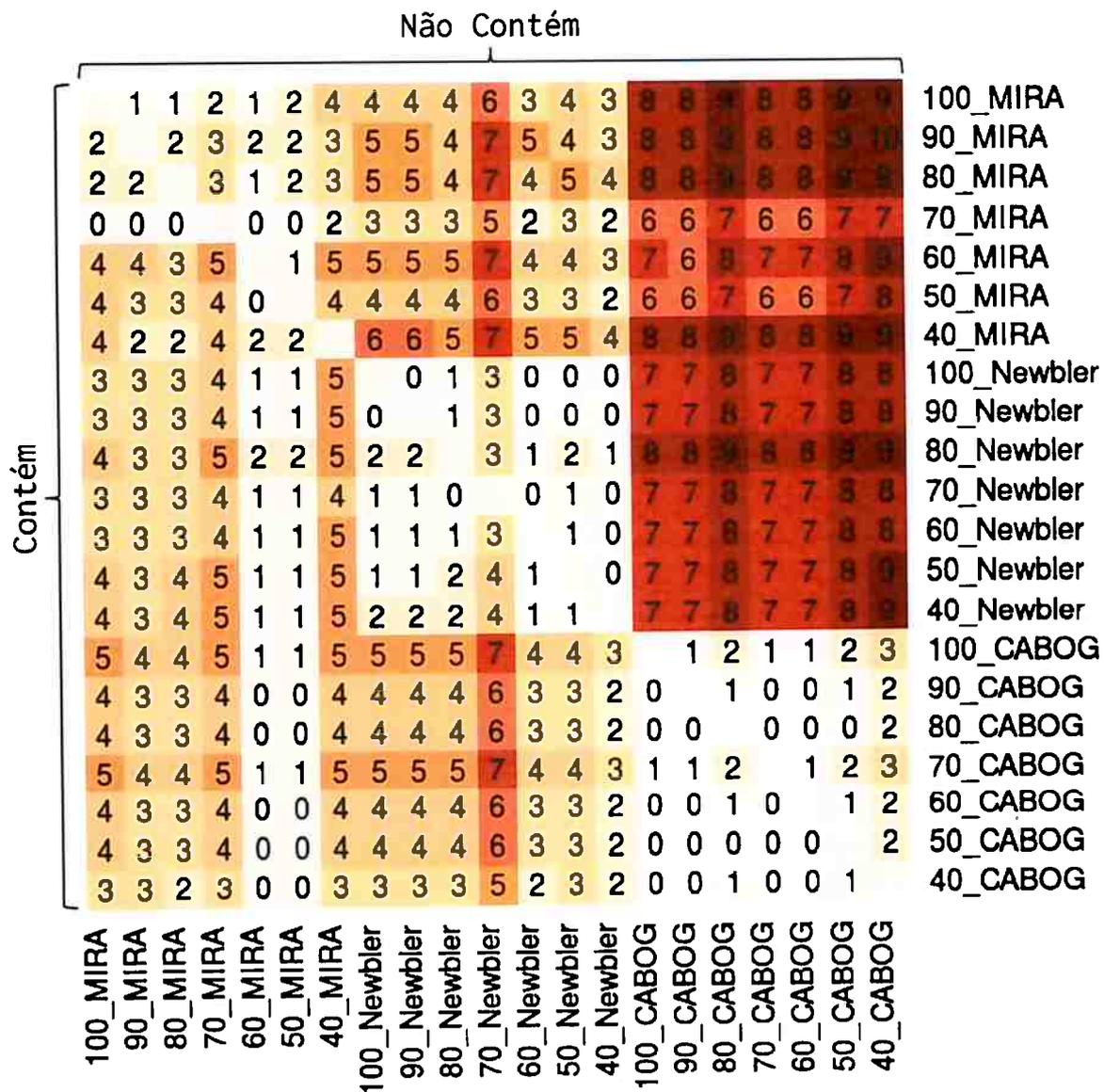


Figura 3.8: Comparação do conteúdo de gênico total entre pares de anotações. Para cada par, o valor mostrado representa quantos produtos gênicos estavam presentes em determinada reação (linhas), mas que não o estavam na outra (colunas)

		Não Contém																					
Contém		1	0	1	0	1	1	4	4	3	4	3	4	3	1	1	2	1	1	2	1	100_MIRA	
		1		1	1	0	0	0	3	4	4	4	3	2	3	0	0	1	0	0	1	1	90_MIRA
		0	1		1	0	1	1	3	4	4	4	3	3	4	1	1	2	1	1	2	1	80_MIRA
		0	0	0		0	0	0	2	3	3	3	2	2	3	0	0	1	0	0	1	0	70_MIRA
		3	3	3	4		1	3	3	4	4	4	4	4	3	1	0	2	1	1	2	2	60_MIRA
		3	2	3	3	0		2	4	4	3	4	3	3	2	0	0	1	0	0	1	1	50_MIRA
		1	0	1	1	0	0		4	4	3	4	3	3	2	0	0	1	0	0	1	1	40_MIRA
		2	2	2	2	0	0	2	0	0	1	0	0	0	0	0	0	1	0	0	1	0	100_Newbler
		2	2	2	2	0	0	2	0		0	1	0	0	0	0	0	1	0	0	1	0	90_Newbler
		2	2	2	2	0	0	2	1	1		1	0	1	0	0	0	1	0	0	1	0	80_Newbler
		2	2	2	2	0	0	2	1	1	0		0	1	0	0	0	1	0	0	1	0	70_Newbler
		2	2	2	2	0	0	2	1	1	0	1		1	0	0	0	1	0	0	1	0	60_Newbler
		3	2	3	3	0	0	2	1	1	1	2	1		0	0	0	1	0	0	1	1	50_Newbler
		3	2	3	3	0	0	2	2	2	1	2	1	1		0	0	1	0	0	1	1	40_Newbler
		4	3	4	4	1	1	3	4	4	4	4	4	4	3		1	2	1	1	2	2	100_CABOG
		3	2	3	3	0	0	2	4	4	3	4	3	3	2	0		1	0	0	1	1	90_CABOG
	3	2	3	3	0	0	2	4	4	3	4	3	3	2	0	0		0	0	0	1	80_CABOG	
	3	2	3	3	0	0	2	4	4	3	4	3	3	2	0	0	1		0	1	1	70_CABOG	
	3	2	3	3	0	0	2	4	4	3	4	3	3	2	0	0	1	0		1	1	60_CABOG	
	3	2	3	3	0	0	2	4	4	3	4	3	3	2	0	0	0	0	0		1	50_CABOG	
	2	2	2	2	0	0	2	3	3	2	3	2	3	2	0	0	1	0	0	1		40_CABOG	
	100_MIRA	90_MIRA	80_MIRA	70_MIRA	60_MIRA	50_MIRA	40_MIRA	100_Newbler	90_Newbler	80_Newbler	70_Newbler	60_Newbler	50_Newbler	40_Newbler	100_CABOG	90_CABOG	80_CABOG	70_CABOG	60_CABOG	50_CABOG	40_CABOG		

Figura 3.9: Comparação do conteúdo de gênico entre pares de anotações. Para cada par, o valor mostrado representa quantos produtos gênicos que continham o número EC ou que se relacionavam com mecanismos de transporte estavam presentes em determinada reação (linhas), mas que não o estavam na outra (colunas)

3.6 Conclusões

Foi escolhida a montagem MIRA 29 para dar continuidade no processo de reconstrução metabólica. A escolha se deveu ao fato dessa montagem ter tido resultados satisfatórios nos três critérios estabelecidos em comparação com as outras montagens.

Capítulo 4

Reconstrução da rede metabólica em escala genômica de *B. sacchari*

4.1 Considerações Preliminares

Neste capítulo, pretende-se reconstruir a rede metabólica de *B. sacchari* a partir do genoma anotado que foi escolhido no capítulo anterior. Primeiro, será feita uma breve revisão de literatura visando mostrar como o modo de reconstruir redes metabólicas se alterou ao longo dos últimos 30 anos e qual é seu estado da arte. Depois se afirmará a abordagem usada para o caso de *B. sacchari*, a seguir serão mostrados os métodos empregados e, por fim, os resultados obtidos.

4.2 Revisão de Literatura

Antes da era genômica, o processo de reconstrução de redes metabólicas residia basicamente no uso de informações a respeito da caracterização bioquímica das enzimas de determinado organismo e também de no uso de informações da literatura e essas reconstruções eram limitadas a poucos organismos, como *Escherichia coli*, *Clostridium acetobutylicum* e *Bacillus subtilis* (Feist *et al.*, 2009).

Com o sequenciamento do primeiro organismo não viral em 1995 a biologia entra na era pós-genômica. A partir de então, há uma disponibilidade cada vez maior de genomas sequenciados e anotados e esse fato acaba sendo um divisor de águas com respeito ao modo como as redes metabólicas vinham sendo reconstruídas. Ao contrário de antes, o processo de reconstrução usa os genomas anotados como esqueleto onde são adicionadas informações coletadas na literatura e a primeira reconstrução da rede metabólica usando essa abordagem foi feita para a bactéria *Haemophilus influenzae* em 1999 (Edwards e Palsson, 1999). Ela foi reconstruída baseada na anotação do genoma que havia sido feita por Fleischmann *et al.* (1995). Este trabalho também foi acompanhado de estudos computacionais da rede reconstruída.

No ano seguinte, houve a reconstrução de mais um organismo modelo: *Escherichia coli*, no trabalho de Edwards e Palsson (2000a). Da mesma forma como para *H. influenzae*, *E. coli* foi reconstruída a partir de bases de dados e também teve suas capacidades analisadas *in silico*. Entretanto, estudos posteriores fizeram a ligação entre modelo e experimentos. No trabalho de Edwards e Palsson (2000b) a reconstrução foi usada para prever o fenótipo de uma linhagem modificada e no trabalho de Edwards *et al.* (2001) o mesmo modelo foi usado para prever o fenótipo em diferentes

condições fisiológicas.

O sucesso dessas abordagens tanto *in silico* quanto *in vivo* incentivou a reconstrução de outros organismos e fez com que começassem esforços em direção à criação de ferramentas para automatizar partes do processo. Na primeira metade dos anos 2000, essas ferramentas ainda eram incipientes, mas acabou por modificar o modo como a reconstrução era feita. Um exemplo dessas primeiras ferramentas é o desenvolvimento do *software* Pathway Tools por Karp *et al.* (2002). O impacto dessas ferramentas no processo de reconstrução aconteceu em 2005, quando Borodina *et al.* (2005) publicam a primeira rede reconstruída de uma maneira semi-automatizada e fazem uma análise crítica das lacunas deste tipo de abordagem.

Em paralelo à criação das ferramentas, foi sendo desenvolvida uma metodologia para o processo de reconstrução. Reed *et al.* (2006) propõe a fundamentação conceitual da área, onde a reconstrução da rede metabólica passa a ser vista como uma instância do processo de anotação multidimensional de genomas. Além disso, neste trabalho os autores propõe um guia para o processo de reconstrução. Essas orientações se amadurecem no trabalho de Feist *et al.* (2009), onde os autores descrevem como processo de reconstrução que está sendo feito à época e como elas estão sendo curadas e validadas. Em 2010, o amadurecimento da área culmina no protocolo de instruções no trabalho de Thiele e Palsson (2010a), mostrado na figura 4.1. Ainda hoje, ele representa o padrão-de-ouro em termos de metodologia para a condução de reconstruções metabólicas. Se antes os organismos eram reconstruídos com base em esforços esparsos, a partir de 2010, diversos organismos são reconstruídos a partir desse guia.

Também por volta do fim dos anos 2000, florescem avanços na automatização das reconstrução no sentido de torná-las completamente autônomas. Esses *softwares* criam reconstruções automaticamente e alguns deles, como no caso do novo Pathway Tools (Karp *et al.*, 2010), geram até mesmo os modelos metabólicos e também permitem etapas de análise da rede. Em resumo, o protocolo de Thiele e Palsson é incorporado nas ferramentas de reconstrução metabólica que estão sendo desenvolvidas e muitas reconstruções partem do uso dessas ferramentas automatizadas como ponto de partida para a reconstrução. Desse modo, fica evidente a importância das plataformas no processo de reconstrução metabólica.

Com relação aos desenvolvimentos acima mencionados, Monk *et al.* (2014) se propõe a avaliar o estado da arte na área e faz uma análise crítica acerca do conhecimento metabólico de 117 reconstruções metabólicas. Os autores percebem um descompasso entre o ritmo de geração de reconstruções e o ritmo de geração de reações bioquímicas novas. Este aumentava a uma velocidade bem menor do que aquele, indicando que maioria das reconstruções são baseadas em outras já feitas e pouco contribuem na expansão do conhecimento metabólico como um todo. Além disso, os autores apontam para uma outra limitação que permeia a maioria das reconstruções: a falta de padronização no vocabulário a respeito de metabólitos e de reações. Os autores argumentam que a falta de padronização é um impeditivo para o estudo comparativo entre reconstruções metabólicas e afirmam que das 117 reconstruções analisadas, somente foi possível a comparação de 53 em razão dessa ausência de controle de vocabulário.

A falta de padronização é um tema recorrente em outros estudos críticos sobre as reconstruções. Ravikrishnan e Raman (2015) analisaram 99 reconstruções e, assim como Monk e colaboradores, constatam que há falta de padronização de metabólitos e reações para a maioria das reconstruções analisadas e que isto é um gargalo para a comparação entre reconstruções. Entretanto, Ravikrish-

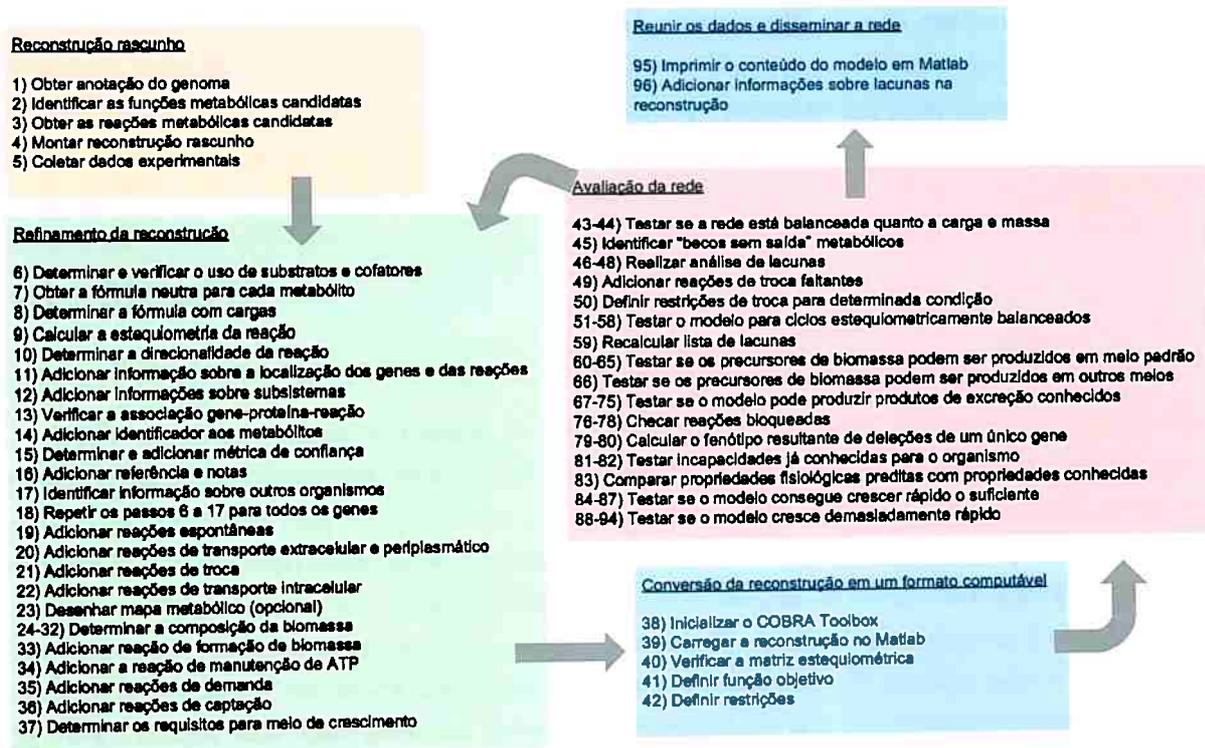


Figura 4.1: Protocolo de Thiele e Palsson (2010)

nan e Raman vão além da identificação do problema e sugerem que esse gargalo poderia ser contornado com o uso de ferramentas apropriadas como o MetRxn (Kumar *et al.*, 2012) e o MetaNeX (Ganter *et al.*, 2013), mas que ainda são largamente subutilizadas.

4.3 Proposta adotada para *B. sacchari*

Tomando como base o estado da arte descrito, onde as ferramentas de reconstrução automatizada desempenham um papel central, se essas ferramentas dispusessem de todo o conhecimento bioquímico disponível, a questão a respeito de como conduzir o processo de reconstrução poderia ser reduzida à questão de escolher a melhor plataforma dentre as disponíveis. Como isso não é válido, isto é, como estas ferramentas estão restritas a uma ou poucas bases de dados, a questão sobre a escolha do *software* passa a ser ofuscada por uma questão muito mais premente: é necessário que as reconstruções geradas pelas ferramentas automáticas sejam padronizadas, tal como foi alertado por Monk *et al.* (2014) e Ravikrishnan e Raman (2015). Padronizar reconstruções traria mais robustez às etapas de curadoria manual e faria com que diferentes reconstruções fossem comparáveis. Dessa maneira, no presente trabalho, o uso de ferramentas de padronização de vocabulário é uma questão central.

Com relação a essas ferramentas de padronização e unificação de bases de dados, Pfau *et al.* (2015) comenta que MetaNetX é atualmente a ferramenta mais completa tanto porque padroniza uma maior variedade de informações como porque permite, além da adequação de vocabulário, a análise da rede metabólica reconstruída.

Assim, a proposta de reconstrução para *B. sacchari* compreenderá a escolha de uma ferramenta de reconstrução automatizada e a padronização desta reconstrução com a plataforma MetaNetX. A

rede reconstruída e com vocabulário controlado será então sujeita a duas análises que serão feitas dentro da própria plataforma do MetaNetX: Análise de Balanço de Fluxo e Análise de Nocautes de Reações. A primeira visa testar se a rede metabólica é capaz de suportar crescimento, ou seja, se não haveria lacunas na formação de biomassa enquanto que a segunda visa testar o impacto da ausência de reações individuais no crescimento celular.

4.4 Métodos

4.4.1 Seleção de ferramenta para reconstrução automática

Na literatura consultada, o trabalho de Hamilton e Reed (2014) foi o único que comparou diferentes ferramentas automáticas de reconstrução metabólica. Por isso, ele foi usado como guia para a escolha da ferramenta para ser aplicada ao genoma anotado de *B. sacchari*. Nesse estudo, os autores compararam quatro ferramentas de reconstrução automática com respeito às etapas do protocolo de Thiele e Palsson. Cada ferramenta foi investigada com relação ao modo como executa cada uma das etapas (ou conjunto de etapas): se a etapa é executada automaticamente sem o intermédio do usuário, se a etapa é executada mediante assistência do usuário e se a etapa não é realizada. A figura 4.2 mostra uma visão geral das ferramentas analisadas e das etapas consideradas.

Legenda		Subliminal	ModelSEED	RAVEN	PathwayTools
	<ul style="list-style-type: none"> Automático Requer assistência Sem suporte 				
Reconstrução rascunho	1) Obter anotação do genoma 2) Identificar as funções metabólicas candidatas 3) Obter as reações metabólicas candidatas 4) Montar reconstrução rascunho				
Refinamento da reconstrução	6) Determinar e verificar o uso de substratos e cofatores 7, 8) Obter a fórmula neutra para cada metabólito 9, 43-44) Calcular a estequiometria da reação 10) Determinar a direcionalidade da reação 11) Adicionar informação sobre a localização dos genes e das reações 12) Adicionar informações sobre subsistemas 13) Verificar a associação gene-proteína-reação 14) Adicionar identificador aos metabólitos 15) Determinar e adicionar métrica de confiança 16) Adicionar referência e notas 17) Identificar informação sobre outros organismos 19) Adicionar reações espontâneas 20) Adicionar reações de transporte extracelular e periplasmático 22) Adicionar reações de transporte intracelular 23) Desenhar mapa metabólico (opcional) 24-33) Determinar a composição da biomassa 34) Adicionar a reação de manutenção de ATP 35,36) Adicionar reações de demanda 37) Determinar os requisitos para meio de crescimento				
Avaliação da rede	45) Identificar "becos sem saída" metabólicos 46-48) Realizar análise de lacunas 51-58) Testar o modelo para ciclos estequiometricamente balanceados 60-66) Testar se os precursores de biomassa podem ser produzidos em meio padrão 67-75) Testar se o modelo pode produzir produtos de excreção conhecidos 76-78) Checar reações bloqueadas 79-80) Calcular o fenótipo resultante de deleções de um único gene 81-83) Testar incapacidades já conhecidas para o organismo 83) Comparar propriedades fisiológicas preditas com propriedades conhecidas 84-94) Testar se o modelo consegue crescer rápido o suficiente				
Etapas omitidas	5, 18, 21, 38-42, 49-50, 59, 95-96				

Figura 4.2: As ferramentas de reconstrução automática incorporaram o protocolo de Thiele e Palsson. Extraído de Hamilton e Reed (2014).

Dentre as ferramentas avaliadas, Model SEED se mostrou a mais adequada para reconstruir a rede metabólica de *B. sacchari*. Primeiro pois é a ferramenta que mais executa etapas automaticamente. Segundo porque consegue inferir a composição de biomassa da bactéria e determinar o requerimento quanto à ATP de manutenção (Subliminal também consegue inferir a composição de biomassa, entretanto não consegue inferir ATP de manutenção e também executa menos passos automaticamente). Além disso, embora o estudo de Hamilton e Reed não mostre esse fato, Model SEED tem integração com o anotador RAST usado na anotação do genoma montado. Desse modo, Model SEED foi a ferramenta de reconstrução escolhida.

4.4.2 Reconstrução automática com a ferramenta selecionada

O genoma anotado que foi selecionado no capítulo anterior foi submetido à reconstrução automática usando a ferramenta ModelSEED, versão 1.0, disponível no endereço eletrônico <http://seed-viewer.theseed.org/seedviewer.cgi?page=ModelView>, com parâmetros *default*.

4.4.3 Controle de vocabulário

A ferramenta MetaNetX foi usada através de sua aplicação *web*, disponível em <http://metanetx.org/>. Através dessa aplicação o modelo metabólico construído pelo Model SEED e em formato SBML foi submetido. Os parâmetros usados são os seguintes: para o quesito *Mapping to MNXref*, o parâmetro utilizado foi *All*. E para o quesito SBML *type* o parâmetro utilizado foi *SEED*.

4.4.4 Análise da rede reconstruída

Análise de Balanço de Fluxo

Apesar de toda a complexidade envolvida no análise de balanço de fluxos, a plataforma MetaNetX só disponibiliza um único botão para que essa análise seja executada. Por trás desse botão, o MetaNetX assume que a função a ser otimizada é a formação de biomassa e que os limites superiores e inferiores para o fluxo pelas reações são aqueles configurados originalmente no modelo, não havendo possibilidade de alterações. Além disso, o MetaNetX assume que os metabólitos externos são aqueles que são exteriores ao compartimento do citosol. A ABF foi executada no modelo de *B. sacchari* por meio do MetaNetX.

Análise de Nocautes de Reações

A ANR tem o mesmo embasamento teórico da ABF. Seu modo de funcionamento é através da restrição do limite superior e do limite inferior de uma reação ao zero, isto é, o nocaute desta reação. Para cada nocaute é feita a ABF e visando testar se o modelo consegue crescer naquela condição, ou, em outras palavras, se a equação de formação de biomassa consegue suportar fluxo.

4.5 Resultados e Discussão

4.5.1 Reconstrução automática

A reconstrução automática pelo Model SEED resultou em 1.511 reações, 1168 metabólitos e 1545 peptídeos. Além disso, o modelo conta com três compartimentos: fora do sistema, espaço extracelular e espaço citoplasmático. Com relação a esses compartimentos, nota-se a ausência do espaço periplasmático, afinal *Burkholderia sacchari* é uma bactéria gram-negativa. A equação de formação de biomassa obtida é mostrada na figura 4.3.

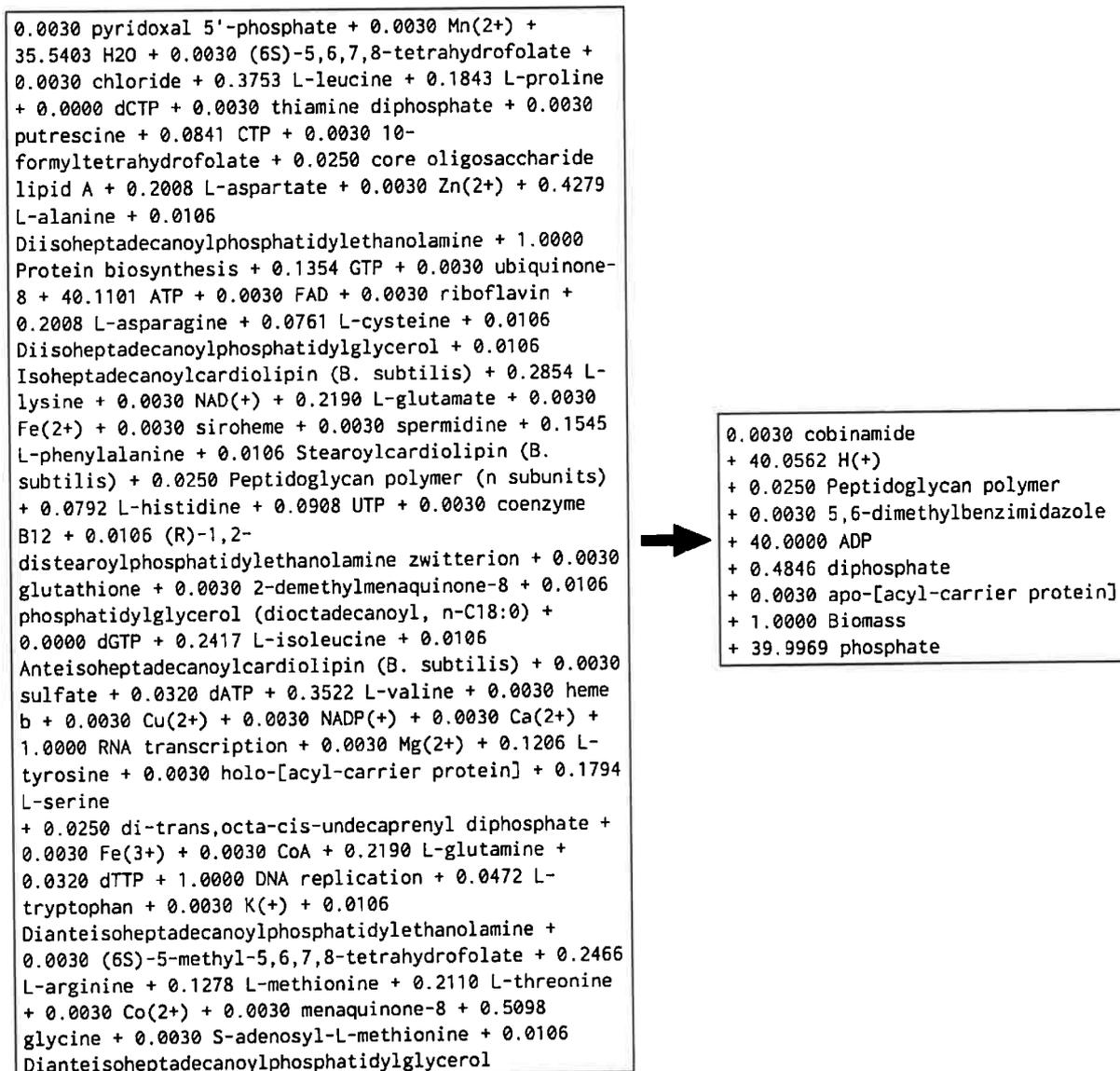


Figura 4.3: Representação esquemática da equação de formação de biomassa de *B. sacchari* inferida pela ferramenta Model SEED

4.5.2 Adequação de vocabulário

A adequação de vocabulário resultou no seguinte mapeamento. Com relação às reações, das 1.511 reações, 1.467 foram mapeadas à linguagem MNXref com identificador e descrição diferente. As 44

reações restantes foram parcialmente mapeadas. Com relação aos metabólitos, do total de 1.168, 1.081 foram mapeados com identificador e descrição diferente, 86 foram mapeados com identificador diferente, mas com a mesma descrição e um metabólito foi mapeado com ambos identificador e descrição iguais. Por último, com relação aos compartimentos, dos três compartimentos, dois foram mapeados com identificador diferente, mas com a mesma descrição e um foi mapeado com o mesmo identificador mas com descrição diferente. Além disso, a classificação das reações feita automaticamente pelo MetaNetX culminou nos resultados mostrados na figura 4.4:

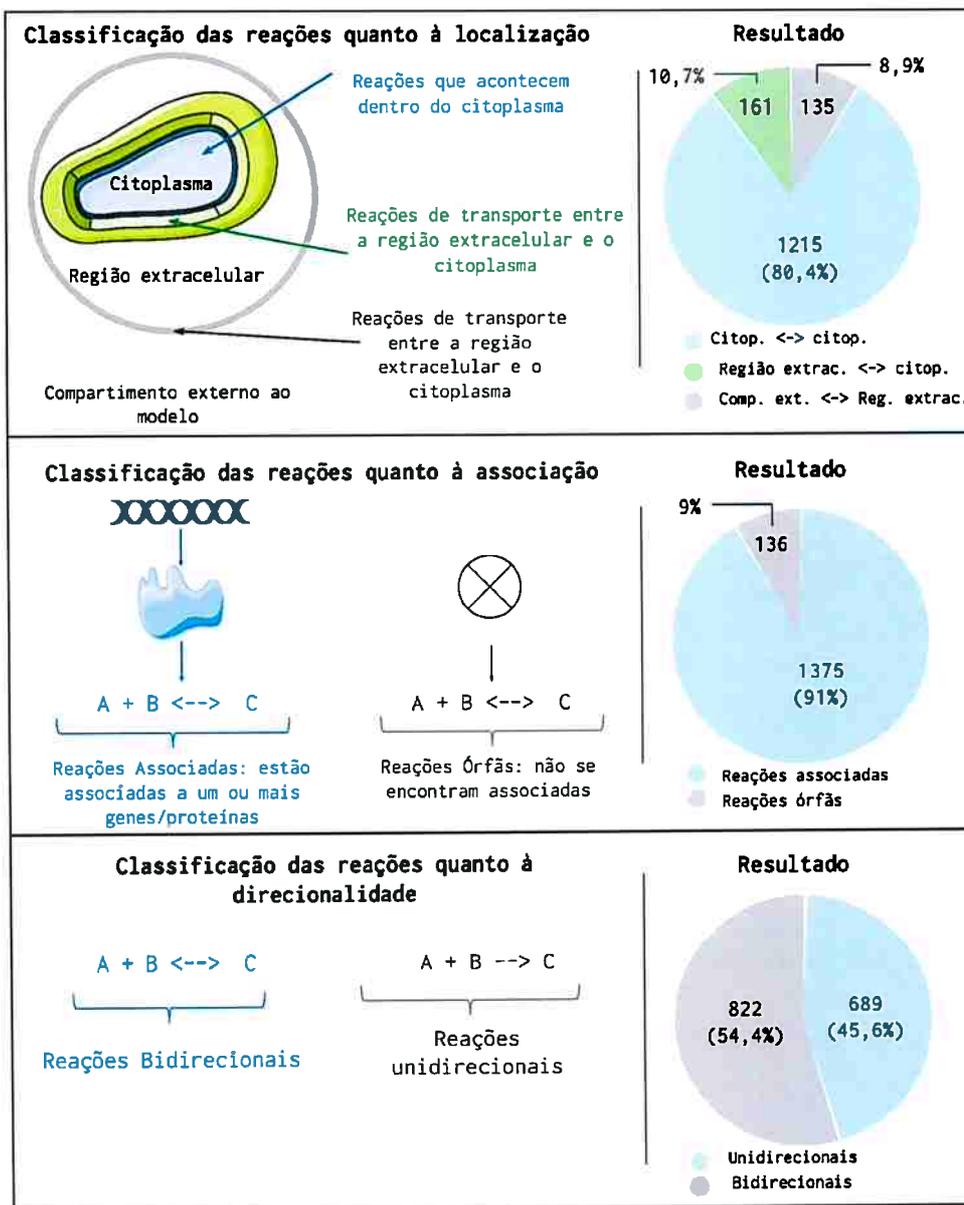


Figura 4.4: Classificação das reações de *B. sacchari* pelo MetaNetX quanto à localização, associação e direcionalidade.

4.5.3 Análise da rede reconstruída

Análise de crescimento

A análise de crescimento realizada através da Análise de Balanço de Fluxo resultou em crescimento celular. Isso indica que não há lacunas entre os precursores e a formação de biomassa.

Análise de nocautes de reações

A análise de nocautes de reações resultou em 293 nocautes letais. Isso quer dizer que cada presença é obrigatória para que to que os precursores da reação de formação de biomassa. Dentre esses nocautes, chamaram atenção aqueles que estavam envolvidos com o transporte entre compartimentos. A lista desse subconjunto de nocautes letais é mostrada na tabela 4.1. Nesta listagem, o que chamou atenção foi a presença do carboidrato trealose (3ª linha da tabela, α,α -trehalose). A presença da reação de transporte de trealose do meio extracelular para o meio intracelular implica que trealose é essencial para *B. sacchari* e que a bactéria não seria capaz de crescer na ausência desta fonte de carbono. Este resultado, porém, está em conflito com o que se sabe sobre a fisiologia da bactéria. No artigo que descreve *B. sacchari* como uma nova espécie, Brämer e colaboradores realizaram testes de assimilação de várias fontes de carbono e constataram que a bactéria não consegue crescer quando usa somente trealose como única fonte de carbono (Brämer *et al.*, 2001).

Tabela 4.1: Subconjunto dos 293 nocautes letais que envolvem o transporte de metabólitos entre o compartimento externo e a região extracelular

1 L-lisina [extracelular]	<==>	1 L-lisina [externo]
1 Co(2+) [extracelular]	<==>	1 Co(2+) [externo]
1 α,α -trealose [extracelular]	<--->	1 α,α -trealose [externo]
1 tiamina [extracelular]	<==>	1 tiamina [externo]
1 Mg(2+) [extracelular]	<==>	1 Mg(2+) [externo]
1 Fe(2+) [extracelular]	<==>	1 Fe(2+) [externo]
1 espermidina [extracelular]	<==>	1 espermidina [externo]
1 Fe(3+) [extracelular]	<==>	1 Fe(3+) [externo]
1 glicilasparagina [extracelular]	<==>	1 glicilasparagina [externo]
1 Zn(2+) [extracelular]	<==>	1 Zn(2+) [externo]
1 Cl(1-) [extracelular]	<==>	1 Cl(1-) [externo]
1 Cu(2+) [extracelular]	<==>	1 Cu(2+) [externo]
1 Mn(2+) [extracelular]	<==>	1 Mn(2+) [externo]
1 citosina [extracelular]	<==>	1 citosina [externo]
1 Ca(2+) [extracelular]	<==>	1 Ca(2+) [externo]
1 K(+) [extracelular]	<==>	1 K(+) [externo]

Desse modo, na tentativa de solucionar o conflito entre reconstrução e evidência biológica, foram percorridas todas as reações desde a entrada de trealose na célula até a formação de biomassa e foi constatado que o gargalo estava na produção de glicose-1-fosfato, afinal a reconstrução obtida apresenta esta molécula como precursora longínqua da formação de biomassa e só trealose conseguiria suprir a célula com glicose-1-fosfato. Assim, pesquisou-se dentro do próprio MetaNetX de onde poderia estar sendo providenciada a glicose-1-fosfato em outros organismos e foi constatado que G1P pode ser produzido através de G6P. Essa reação se mostrou relevante pois a conversão de

Tabela 4.2: Reações envolvendo G1P que estão presentes em outros modelos, mas que não o estão em *B. sacchari*. Os modelos usados foram *bigg_iJN746*, *bigg_iAF1260*, *seed_Seed272560_3* e *iso_iYO844_flux1* para *P. putida*, *E. coli*, *B. pseudomallei* e *B. subtilis*, respectivamente. Todos os modelos foram obtidos através do MetaNetX

Organismo	Reações
<i>Pseudomonas putida</i> KT2440	1 G6P [c] \rightleftharpoons 1 G1P [c]
<i>Escherichia coli</i> K12	1 G6P [c] \rightleftharpoons 1 G1P [c] 1 G1P [e] \rightarrow 1 G1P [x] 1 G1P [p] + 1 H ₂ O [p] \rightarrow 1 fosfato [p] + 1 D-glicose [p] 1 G1P [c] + 1 α -maltohexaose [c] \rightleftharpoons 1 fosfato [c] + 1 α -maltoheptaose [c] 1 α -D-galactose-1P [c] + 1 UDP-D-glicose [c] \rightleftharpoons 1 G1P [c] + 1 UDP- α -D-galactose [c]
<i>Burkholderia pseudomallei</i> K96243	1 CDP-D-glicose [c] + 1 difosfato [c] \rightleftharpoons 1 H(+) [c] + 1 G1P [c] + 1 CTP [c] 1 fosfato [c] + 1 sacarose [c] \rightleftharpoons 1 G1P [c] + 1 D-frutose [c] 1 G6P [c] \rightleftharpoons 1 G1P [c]
<i>Bacillus subtilis</i> 168	1 G6P [c] \rightleftharpoons 1 G1P [c] 2 fosfato [c] + 1 G1P [e] \rightarrow 2 fosfato [e] + 1 G1P [c] 1 G1P [e] \rightarrow 1 G1P [x] 2 G1P [c] \rightarrow 1 D-glicose [c] + 1 D-glicose-1,6-bifosfato [c]

G6P até G1P poderia fazer com que a célula não se limitasse à trealose.

Entretanto, era preciso ter mais respaldo antes de simplesmente adicionar a equação, então no intuito de procurar por alguma falha na anotação ou alguma anotação equivocada, recorreu-se à base de dados MetaCyc para procurar pela reação desejada. Foi verificado que ela é catalisada pela enzima fosfoglucomutase, de número EC 5.4.2.2., e o panorama encontrado é mostrado na figura 4.5. Nesta figura, nota-se que a enzima fosfoglucomutase de *Pseudomonas aeruginosa* é capaz de catalisar tanto a reação 5.4.2.2 desejada quanto a reação correlata 5.4.2.8, onde é usado o substrato manose, ao invés de glicose.

Foi visto, através do anotador RAST, que o número EC 5.4.2.8 está presente em *B. sacchari*. Desse modo a sequência de *Pseudomonas* e de *B. sacchari* foram alinhadas e o resultados indicou 55% de identidade e 72% de aminoácidos positivos. As sequências das duas enzimas, além do resultado do alinhamento, são mostrados na figura 4.6.

Esses resultados foram considerados como evidências de que a reação de conversão entre G6P e G1P pode ocorrer em *B. sacchari*. Assim, a reação foi adicionada à reconstrução. Para comprovar o sucesso das alterações, foi feita uma nova análise de nocaute, que revelou 288 reações letais. A tabela 4.3 mostra a disjunção entre os dois conjuntos. Nela consta a presença da equação de entrada de trealose na célula e isso significa que a alteração feita foi responsável por mudar a essencialidade do modelo quanto à trealose, fato compatível com os resultados de Brämer *et al.* (2001).

4.6 Conclusões

A rede metabólica em escala genômica de *Burkholderia sacchari* foi reconstruída e está disponível nos formatos SBML e arquivo TSV (Tab-Separated Values - arquivo cujos valores estão separados por tabulação). Além da reconstrução foi feita uma etapa de curadoria manual que resultou na inclusão de uma reação no modelo. Nessa etapa, foi usado como premissa a equação de formação de biomassa gerada pela ferramenta Model SEED. Por ser uma reação inferida com base na composição

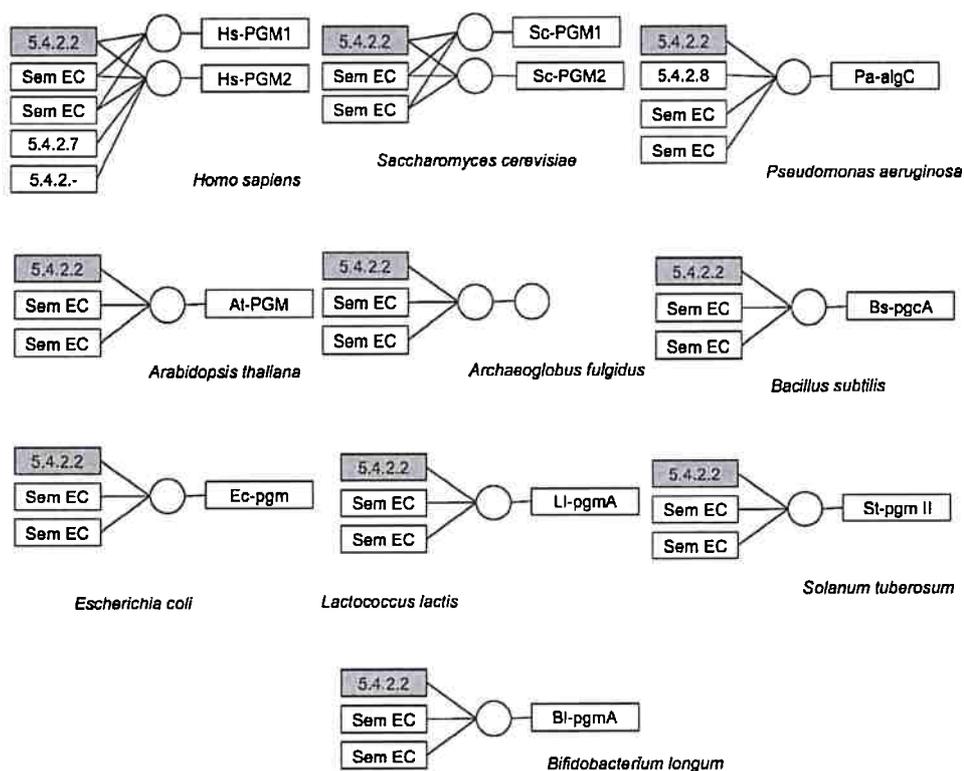
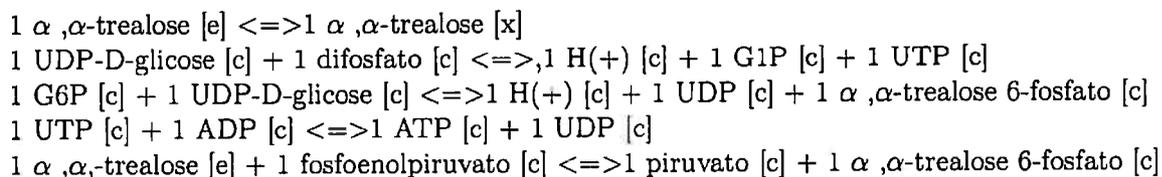


Figura 4.5: A enzima fosfoglucomutase em diferentes organismos. Esta figura mostra a relação. Extraído do banco de dados MetaCyc.

Tabela 4.3: Lista mostrando a disjunção entre os 293 nocautes letais obtidos antes da reanotação e os 288 nocautes letais obtidos na análise posterior



de outras bactérias, esta reação está sujeita a erros.

Por fim, é necessário mencionar que apesar das características de interesse da plataforma MetaNetX (principalmente a completude de sua padronização e a possibilidade de realizar análises da rede), foram observadas algumas lacunas que dificultaram o processo de curadoria manual da enzima fosfoglucomutase. Quando se leva em consideração que as etapas de refinamento do protocolo de Thiele e Palsson, mais especificamente a etapa 18, sugerem que cada gene seja inspecionado individualmente, estas lacunas identificadas podem tornar a curadoria da reconstrução de *B. sacchari*, nos moldes do MetaNetX, demasiadamente longa e enfadonha. Desse modo, essas lacunas foram endereçadas neste trabalho e serão pormenorizadas no capítulo seguinte.

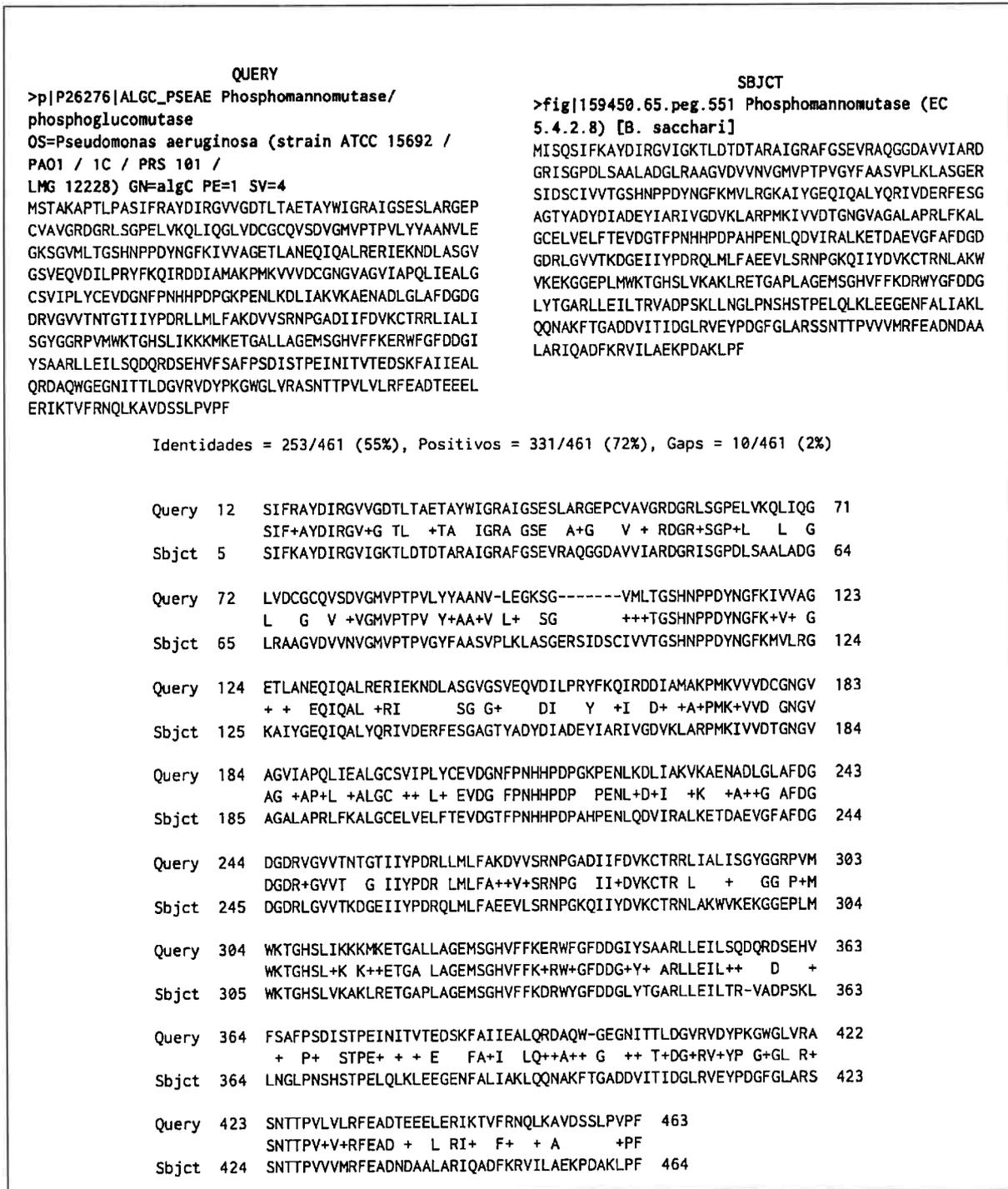


Figura 4.6: Alinhamento entre as duas enzimas fosfomanomutases usando a ferramenta BLAST.

Capítulo 5

Representação da rede metabólica reconstruída como um banco de dados orientado a grafo

5.1 Considerações Preliminares

As lacunas que foram identificadas na ferramenta MetaNetX com respeito às etapas de curadoria manual são tratadas no presente capítulo. Resolver tais lacunas, ou pelo menos apontar direções para resolvê-las, é importante pois trará mais robustez e agilidade ao extenso processo de curadoria manual que precisará ser feito em trabalhos futuros que tenham em vista o refinamento da reconstrução rascunho de *Burkholderia sacchari*. Assim, diante da importância da curadoria manual, este capítulo é dedicado à proposta de uma nova forma de representar a reconstrução metabólica para superar as lacunas encontradas. Assim, primeiro serão descritas as dificuldades encontradas, depois será proposta a nova forma de representação e, por fim, será mostrado como ficaria o processo de curadoria manual com a nova representação.

5.2 Lacunas encontradas no processo de curadoria manual

5.2.1 Dificuldade em se percorrer vias metabólicas

Durante as etapas de curadoria manual da enzima fosfoglucomutase, foi necessário inspecionar os componentes da rede metabólica de *B. sacchari* para descobrir qual era a razão de trealose ser uma fonte de carbono essencial, enquanto que os dados experimentais indicavam a situação contrária. Desde a entrada de trealose na célula, cada passo foi inspecionado segundo o procedimento descrito na figura 5.1. Nesse processo, o nome do metabólito era inserido no campo de buscas e os resultados exibidos mostravam em quais reações aquele metabólito aparecia. Então, repetia-se esse processo tantas vezes quanto fosse necessário. Dois problemas se mostram aqui: a dificuldade de se percorrer as vias metabólicas e a ausência de uma visão geral sobre elas.

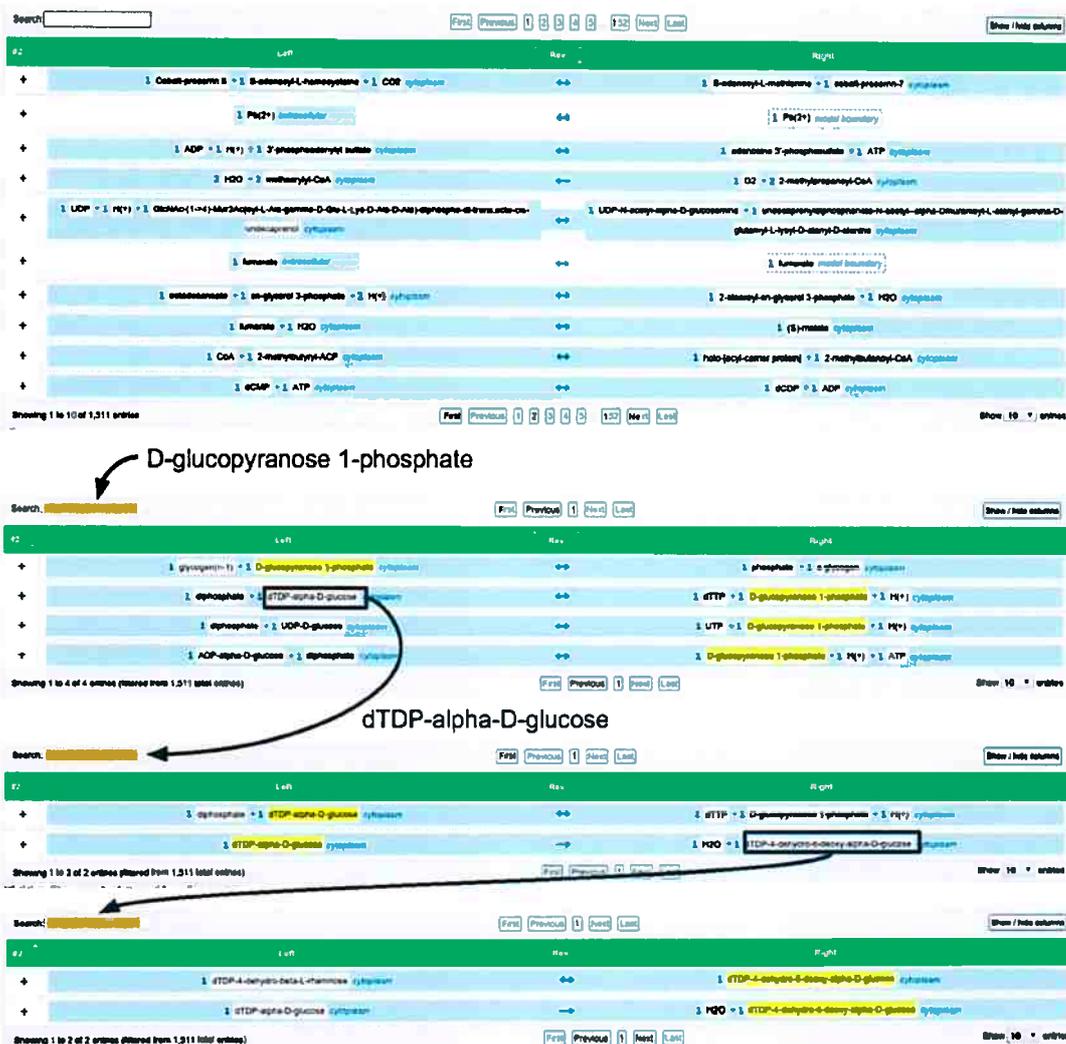


Figura 5.1: Exemplo ilustrando como vias metabólicas podem ser percorridas usando o MetaNetX. Para examinar os componentes de uma via metabólica, é necessário acessar a lista de todas as reações da reconstrução de interesse (primeiro screenshot), onde há um campo designado para buscas. Como exemplo, pesquisou-se por glicose-1-fosfato e a busca resultou em quatro reações que contém esse metabólito (segundo screenshot), onde glicose-1-fosfato foi destacado em amarelo. A seguir, supondo que o metabólito dTDP-alfa-D-glicose seja de interesse, seu nome deve ser inserido novamente no campo de buscas, que resultou em duas reações (terceiro screenshot). Por fim, esse procedimento se repete quantas vezes necessário.

5.2.2 Dificuldade na edição da reconstrução

Para inserir reações na reconstrução foi necessário escolher a reação desejada através da plataforma *web*, escrevê-la no campo indicado na figura 5.2 e, então, realizar a fusão (com a ferramenta *Split/Merge*) das duas listas de reações (a reconstrução que se está curando e a reação individual visada). Todo este procedimento descrito se mostrou excessivamente burocrático e demandou muito tempo para a execução de uma tarefa aparentemente simples de adicionar uma reação à reconstrução que se está curando. Igualmente burocrático é o procedimento de retirar reações de uma reconstrução. Apesar de não ter sido removida nenhuma reação durante as etapas de curadoria manual, foi feito um exemplo de como seriam estas etapas (figura 5.3).

My selection of metabolic networks and pathways

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#3	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
	Overall (non-redundant)	1511	1426	1168	3	1545	



Upload a list of reactions

Upload

Options

Name of the new model:

List of reactions

Mapping option medium



My selection of metabolic networks and pathways

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#3	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
#4	reacao_exemplo	1	2	2	1	0	BC
	Overall (non-redundant)	1512	1426	1168	3	1545	



Split / Merge

Create a new model made of the selected parts

Options

Name of the new model:



My selection of metabolic networks and pathways

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#3	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
#4	reacao_exemplo	1	2	2	1	0	BC
#5	reconstrucao_curada	1512	1426	1168	3	1545	BC
	Overall (non-redundant)	1512	1426	1168	3	1545	

Figura 5.2: Etapas necessárias para a inclusão de uma reação individual ou uma lista de reações na reconstruções em que se está curando pela plataforma MetaNetX. Considerando uma reconstrução exemplo que foi inserida no MetaNetX (primeiro screenshot), para adicionar uma lista de reações de interesse a essa reconstrução, primeiro é necessário inseri-la na plataforma do MetaNetX (segundo screenshot), onde ela será tratada como uma reconstrução (terceiro screenshot). Depois, é preciso mesclar as duas reconstruções (quarto screenshot) para finalmente obter uma reconstrução curada (quinto screenshot)

My selection of metabolic networks and pathways

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#3	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
	Overall (non-redundant)	1511	1426	1168	3	1545	



Upload a list of reactions

Upload

Options

Name of the new model:

List of reactions:

Mapping option: medium



My selection of metabolic networks and pathways

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#1	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
#2	reacao_exemplo	1	4	4	1	0	BC
	Overall (non-redundant)	1512	1426	1168	3	1545	



Combine logically

Logical combination of two models

Options

Model ID:

Select two models to combine:

Select M1	Select M2	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
<input checked="" type="radio"/>	<input type="radio"/>	reconstrucao_exemplo	1511	1426	1168	3	1545	BC
<input type="radio"/>	<input checked="" type="radio"/>	reacao_exemplo	1	4	4	1	0	BC

Operation on the reactions:

All user data are removed after 24h

Figura 5.3: Etapas necessárias para a remoção de uma reação individual ou uma lista de reações na reconstruções em que se está curando pela plataforma MetaNetX. Considerando uma reconstrução exemplo que foi inserida no MetaNetX (primeiro screenshot), para remover uma lista de reações de interesse dessa reconstrução, primeiro é necessário inseri-la na plataforma do MetaNetX (segundo screenshot), onde essa lista será tratada como uma reconstrução (terceiro screenshot). Depois, é preciso combinar logicamente as duas reconstruções por meio do operador lógico NÃO, que irá criar uma reconstrução final contendo as reações que só pertencem à reconstrução exemplo e não à lista adicionada (quarto screenshot)

5.3 Proposta de solução das lacunas identificadas

Diante das lacunas identificadas, elas poderiam ser resolvidas usando uma variedade de abordagens. Para restringir o espaço de soluções, foi usada como âncora uma outra lacuna identificada na literatura: o problema da representação das reconstruções.

Apesar de muitos trabalhos se referirem às reconstruções metabólicas como bases de conhecimento (Feist *et al.*, 2009; Hamilton e Reed, 2014; Heavner e Price, 2015; Thiele e Palsson, 2010a,b), a maioria das reconstruções continua sendo publicada como planilhas eletrônicas e/ou como arquivos no formato SBML (Ravikrishnan e Raman, 2015). O caso não é diferente para o MetaNetX. Nele, as reconstruções estão disponíveis tanto em arquivos TSV (que indiretamente se referem a tabelas) como em arquivos SBML nas versões 2 e 3.

Esses dois formatos, entretanto, não são os ideais para se representar reconstruções metabólicas. De acordo com Ravikrishnan e Raman (2015), representações em planilhas eletrônicas, por não serem estruturadas, são difíceis de serem validadas e sua análise pode ser complexa. Com relação aos arquivos SBML, Heavner e Price (2015) afirmam que este formato é útil para a representação de modelos metabólicos, mas não o é para a representação de reconstruções metabólicas. Para elas, os autores sugerem que a estrutura de bancos de dados seria a mais adequada:

“...current standards for structured data formats that enable publishing and exchanging models (such as the Systems Biology Markup Language (SBML) or the Minimal Information Required in the Annotation of Models (MIRIAM) standard are designed for model exchange, rather than reconstruction exchange. It is likely that exchange and annotation of a reconstruction requires a database schema definition, rather than a markup language, which may be more suitable for exchange of functional models.” (Heavner e Price, 2015)

Nesse contexto, a sugestão de Heavner de representar reconstruções como base de dados foi usada para guiar a escolha de uma solução para as lacunas encontradas no processo de curadoria manual. Com relação aos tipos de bancos de dados, Robinson *et al.* (2015) afirma que dados muito interconectados são melhor representados por meio de bancos de dados não relacionais, mais especificamente por meio de bancos de dados em grafo. Esse tipo de banco representa as entidades como nós e os relacionamentos como arestas que ligam os nós. De acordo com os autores, esse tipo de representação deixa mais intuitiva a modelagem conceitual dos dados e torna as pesquisas por buscas de caminho no grafo mais rápidas.

Como os dados de uma reconstrução metabólica são muito interconectados, seria interessante, então, de acordo com Robinson e colaboradores, que fossem representados através de um banco de dados em grafo. E dentre os vários bancos de dados em grafo existentes, foi visto que o OrientDB (<http://orientdb.com/>) apresenta uma interface visual que seria útil para o processo de curadoria manual. Assim, representar tanto as informações contidas no MetaNetX quanto a reconstrução de *B. sacchari* através do banco de dados em grafo OrientDB tem o potencial de resolver tanto as lacunas encontradas durante as etapas de curadoria manual quanto a lacuna da representação das reconstruções, identificadas por Ravikrishnan e Raman (2015) e Heavner e Price (2015), respectivamente.

5.4 Materiais e Métodos

5.4.1 Descrição do conjunto de dados

Três tipos de dados diferentes foram usados na nova proposta. Os dois primeiros se referem aos dados de 30.897 reações e de 124.835 compostos químicos, que formam a base do MetaNetX. Estas informações estão disponíveis para *download* através do endereço <http://metanetx.org/mnxdoc/mnxref.html> e estão formatadas em arquivos TSV.

Para ilustrar como o arquivo de reações está estruturado, suas duas primeiras linhas são mostradas abaixo:

```
MNX_ID Description Formula Charge Mass InChi SMILES Source
MNXM01 H(+) H 1 1.0079 InChI=1S/p+1 [H+] chebi:15378
```

E para ilustrar a estruturação do arquivo de compostos químicos, também são mostradas suas duas primeiras linhas:

```
MNX_ID Equation Description Balance EC Source
MNXR01 1 MNXM01 = 1 MNXM1 1 'H(+)' = 1 'H(+)' true MNXR01
```

O terceiro tipo de dado que foi usado no banco de dados em grafo diz respeito às reações que estão presentes em *Burkholderia sacchari*. Este arquivo também está em formato TSV e um exemplo de sua entrada é mostrado a seguir.

```
RE5447DA3_LR 1 MNXM10368@MNXC3 --> 1 MNXM2@MNXC3 + 1 MNXM8197@MNXC3
RE5447DA3_LR MNXR27639 4.2.1.0 seed:rxn05365
```

5.4.2 Modelagem conceitual do banco de dados

Baseado no conteúdo dos três arquivos supracitados foi desenhado um banco de dados em que há três tipos de classes de nós: Metabólitos, Reações Gerais e Reações Específicas. As reações gerais são reações em que a direcionalidade não está descrita. As reações específicas se diferenciam das gerais não só pela presença de direcionalidade, mas também por conter informações a respeito das enzimas que as catalisam. Cada classe de nós criados contém uma lista de propriedades, como mostra a figura 5.4, e as informações usadas para preencher essas listas foram retiradas dos arquivos mencionados no item acima.

A modelagem dos relacionamentos entre os nós procurou ser o mais intuitiva possível. Há três tipos de relacionamento, todos unidirecionais: IS_CONSUMED (é consumido) é a conexão que parte de um metabólito substrato e vai até a reação geral que o consome. PRODUCES (produz) é o relacionamento que parte de uma reação geral e chega em um substrato produto. E, por fim, IS_INSTANCE_OF (é uma instância de) é o relacionamento que se origina de uma reação específica e vai até sua respectiva reação geral.

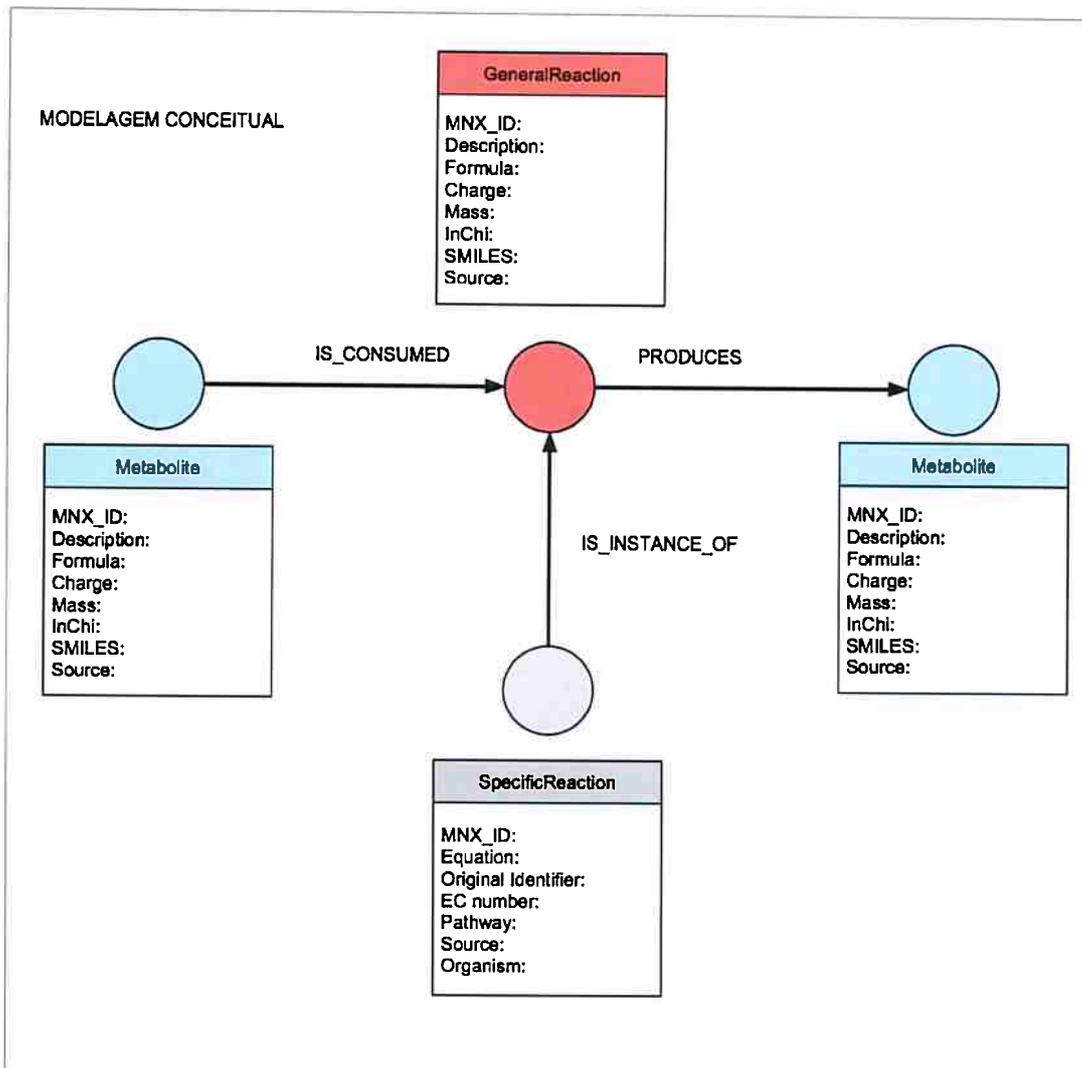


Figura 5.4: Modelagem conceitual do banco de dados. Os nós pertencem a três tipos de classes: Metabólitos, Reações Gerais e Reações Específicas. As propriedades de cada classe foram extraídas dos arquivos no formato TSV do MetaNetX.

5.5 Resultados e Discussão

5.5.1 Percorrendo vias metabólicas com a nova proposta

Foi feito um exemplo de como seria a investigação dos elementos de uma via metabólica de uma maneira sequencial na nova proposta de representação. A figura 5.5 mostra o passo-a-passo dessa investigação por meio de seis *screenshots*. Os resultados indicam que a representação da maneira proposta traz mais agilidade à tarefa de percorrer vias metabólicas. No banco de dados em grafo, partindo de um metabólito de interesse, com dois cliques se chega a um metabólito produto. Já na curadoria usando a plataforma MetaNetX, para obter um metabólito produto é necessário realizar uma busca acerca do metabólito de interesse.

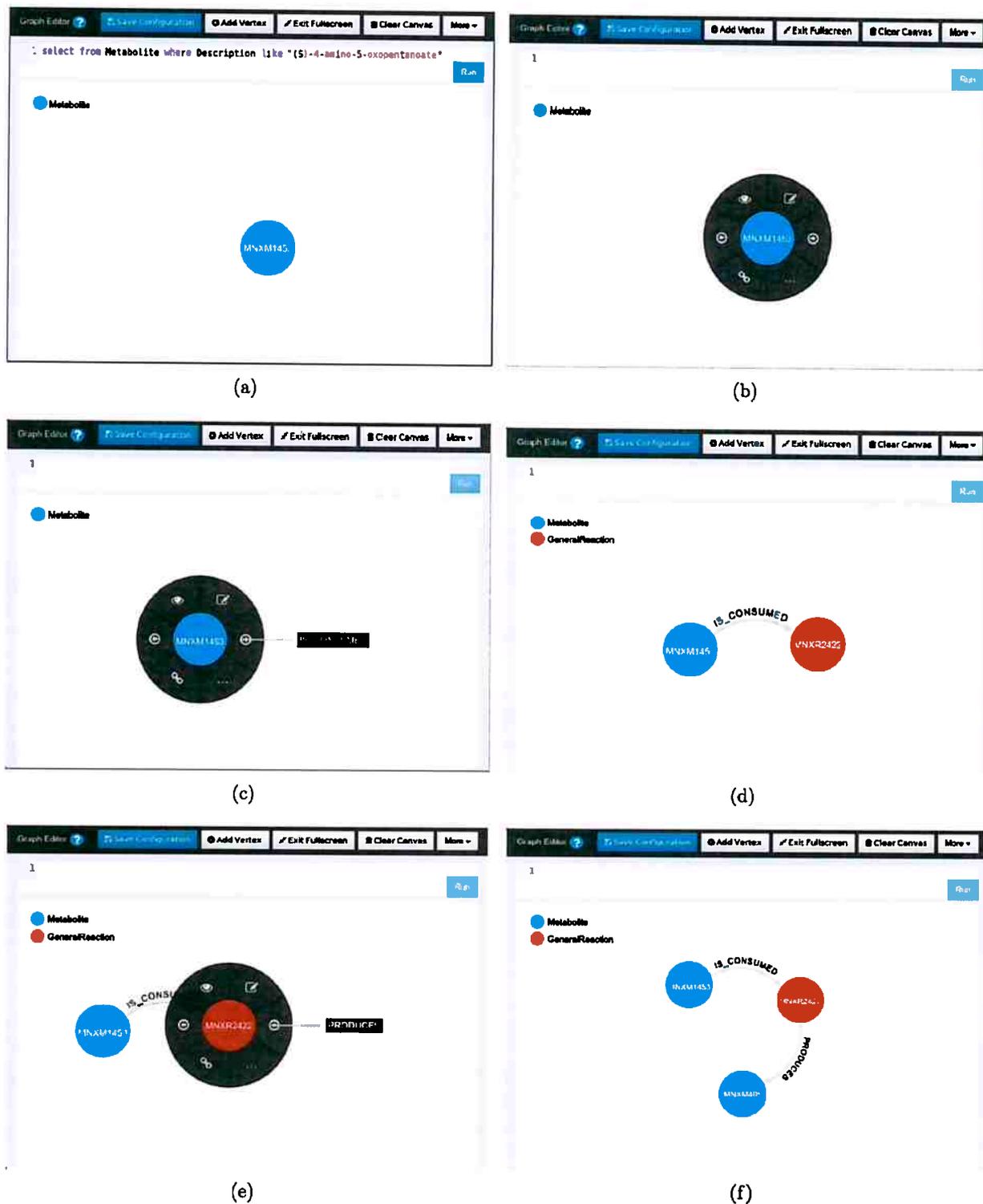


Figura 5.5: Exemplo ilustrando como vias metabólicas podem ser percorridas usando o banco de dados criado. Em (a), um metabólito de interesse é selecionado. Ao clicar nesse metabólito, um menu de opções se abre (b) e, dentre elas, é possível verificar com quais outros nós da rede este metabólito tem relacionamento IS_CONSUMED (c). Ao clicar nessa opção, é mostrado a reação que consome este metabólito (d). De maneira análoga, é possível inquirir a reação para verificar quais metabólitos são produzidos por ela. Nesse caso, clica-se na reação (e) e o resultado é mostrado na subfigura f.

5.5.2 Edição da reconstrução com a nova proposta

Em razão da modelagem conceitual do banco de dados criado, há duas formas de adicionar reações a ele. A primeira é a adição de reações gerais, que é mostrada na figura 5.6. Nela, primeiro se adiciona um nó representando a reação geral e, em seguida, são adicionadas as relações deste nó para com os nós que representam os metabólitos. Já a segunda forma de adicionar reações é através da adição de reações específicas. Este caso é similar ao caso acima pois também começa com a adição de um nó representando a reação, mas se diferencia no passo posterior, onde a reação específica adicionada é ligada a uma reação geral já existente. Com relação à remoção de uma reação, quer seja ela geral ou específica, o procedimento a ser empregado é mostrado na figura 5.7 e consiste em clicar na reação que se deseja excluir, clicar no ícone lixeira no menu que aparecer e, finalmente, confirmar a remoção através de uma caixa de diálogo.

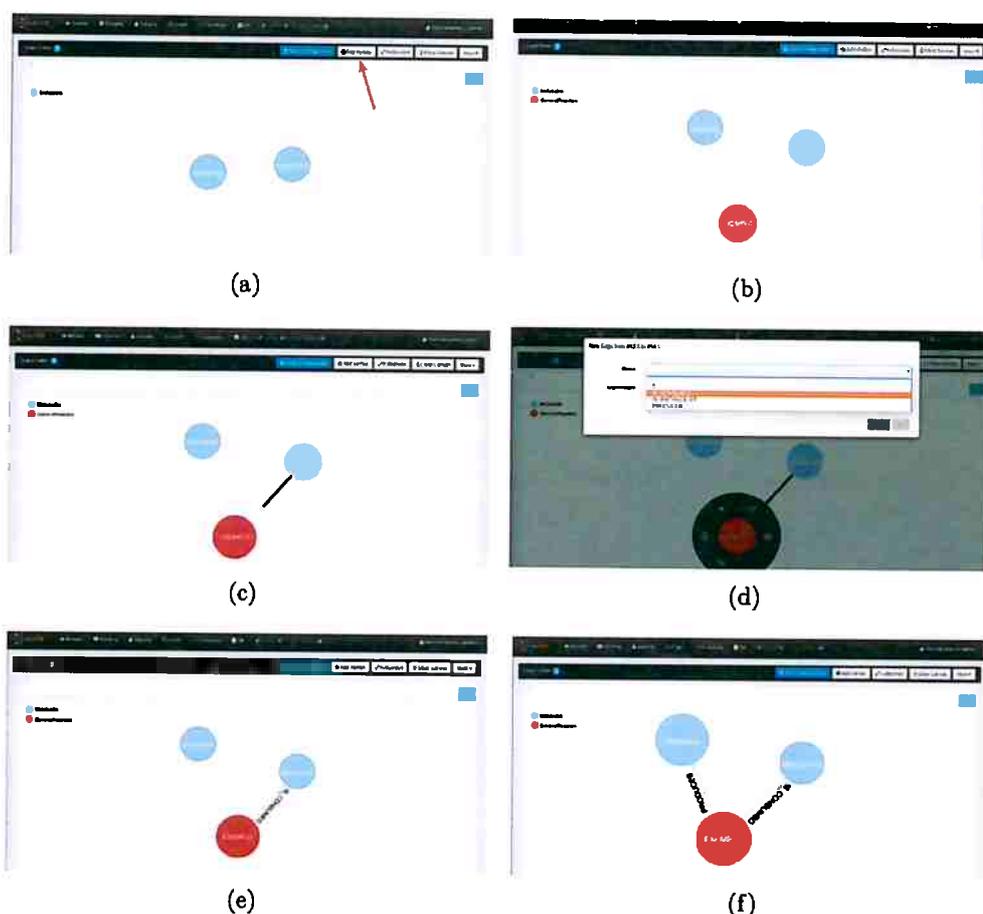
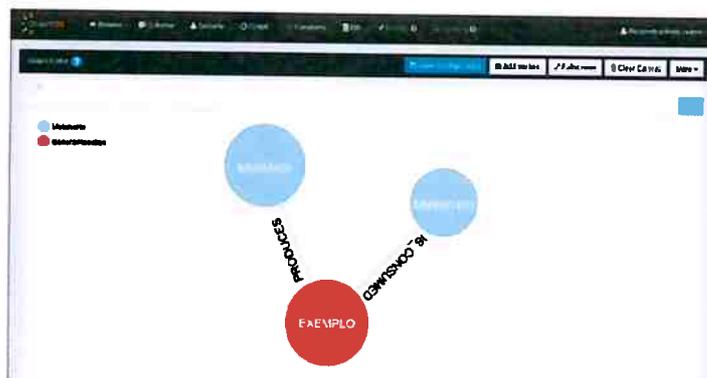
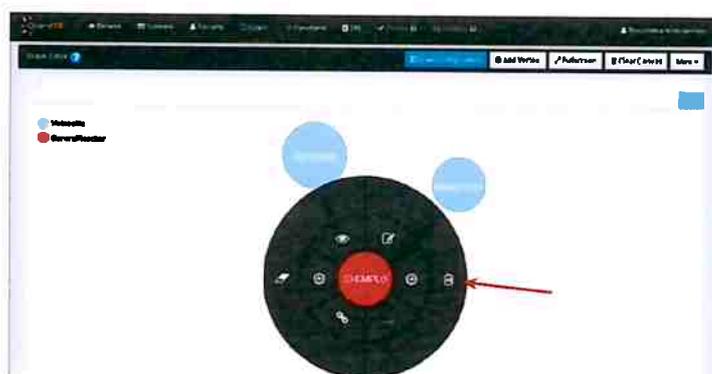


Figura 5.6: Exemplo ilustrando como uma reação é adicionada ao banco de dados. Considerando dois metabólitos de interesse, para adicionar uma reação entre eles é necessário clicar no botão *Add Vertex*, indicado pela seta vermelha na subfigura *a*. Depois de preenchidas as informações sobre a classe de nó a ser adicionada e os atributos desse nó, o banco de dados mostra o elemento adicionado em *b*. Então, para indicar que o metabólito *MNXM1453* é consumido pela reação adicionada, clica-se sobre o metabólito e seleciona-se a opção *adicionar aresta*. Dessa maneira, é preciso arrastar a aresta até o nó de interesse (*c*) e selecionar qual a classe deste relacionamento. Por ser uma relação de consumo, foi selecionada a classe *IS_CONSUMED* (*d*) e o resultado dessa inclusão de relacionamento é mostrado em (*e*). Por fim, repete-se o procedimento para o relacionamento de produção entre a reação exemplo e o metabólito *MNXM405* (*e*)



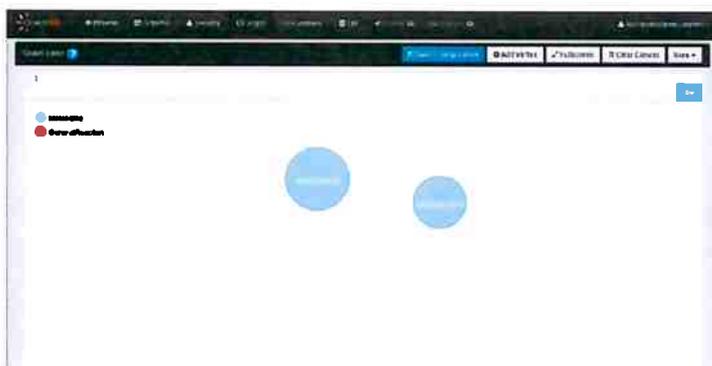
(a)



(b)



(c)



(d)

Figura 5.7: Exemplo ilustrando a remoção de uma reação do banco de dados. Considerando uma reação que deseja ser removida (a), ao clicar sobre esta reação, um menu de opções é exibido (b). Neste menu, o ícone lixeira indica a opção de deleção do componente da rede. Ao clicar nesse ícone e confirmar a deleção (c), a reação é excluída da reconstrução (d).

5.6 Conclusões

Foi obtida uma nova forma de representar e armazenar a reconstrução de *B. sacchari*. Esta nova forma, o banco de dados em grafo, resolve a questão proposta por Heavner e Price (2015) da ausência de formatos robustos para reconstruções, traz mais intuitividade no processo de percorrer os elementos de uma via metabólica e torna a questão da alteração da reconstrução mais ágil. Entretanto, a solução proposta ainda deixa em aberto o gargalo de transformar a reconstrução editada para um formato simulável, como, por exemplo, os arquivos SBML.

Capítulo 6

Considerações Finais

6.1 Contribuições

Este trabalho tem três contribuições principais: primeiro, o conhecimento do conteúdo genômico da bactéria *Burkholderia sacchari*; segundo, o conhecimento de sua rede metabólica; terceiro, uma proposta de representação de reconstruções metabólicas através de um banco de dados orientado a grafos.

Com relação às duas primeiras contribuições, sua importância está relacionada ao modo como facilitam o ciclo de engenharia metabólica da bactéria, onde o conhecimento do conteúdo genômico e a relação deste com o reactoma é extremamente útil tanto para identificar gargalos na produção de algum produto de interesse quanto para atacá-los por meio de estratégias de modificação genética.

Entretanto, apesar da importância deste trabalho ter sido, desde o começo, atrelada à pesquisa aplicada de *B. sacchari*, este trabalho também pode ser visto do ponto de vista da pesquisa básica. Sawana *et al.* (2014) fizeram uma análise filogenômica do gênero *Burkholderia* e seus resultados mostram que algumas assinaturas moleculares seriam o indício de que este gênero seria composto por dois subgrupos distintos. Os autores propõe então que o gênero *Burkholderia* deveria ser dividido em dois gêneros: *Burkholderia* e *Paraburkholderia*. No gênero que foi mantido estariam os organismos patogênicos. No novo gênero, estariam espécies ambientais, que é o caso de *B. sacchari*. Nesse contexto, a reconstrução que foi feita para *Burkholderia sacchari* representa apenas a segunda reconstrução de uma espécie de *Burkholderia* e a primeira do novo gênero proposto *Paraburkholderia* (figura 6.1). Desse modo, a reconstrução de *B. sacchari* é importante também para a pesquisa básica, onde pode ser usada em análises comparativas entre o metabolismo desta espécie e o metabolismo das espécies patogênicas.

Com relação à terceira contribuição, sua importância está no modo como ele facilita as etapas de curadoria manual que precisarão ser feitas nas etapas de refinamento da rede. Nesse sentido, a importância do banco de dados desenvolvido não se restringe ao escopo deste trabalho e pode se estendida para qualquer reconstrução metabólica.

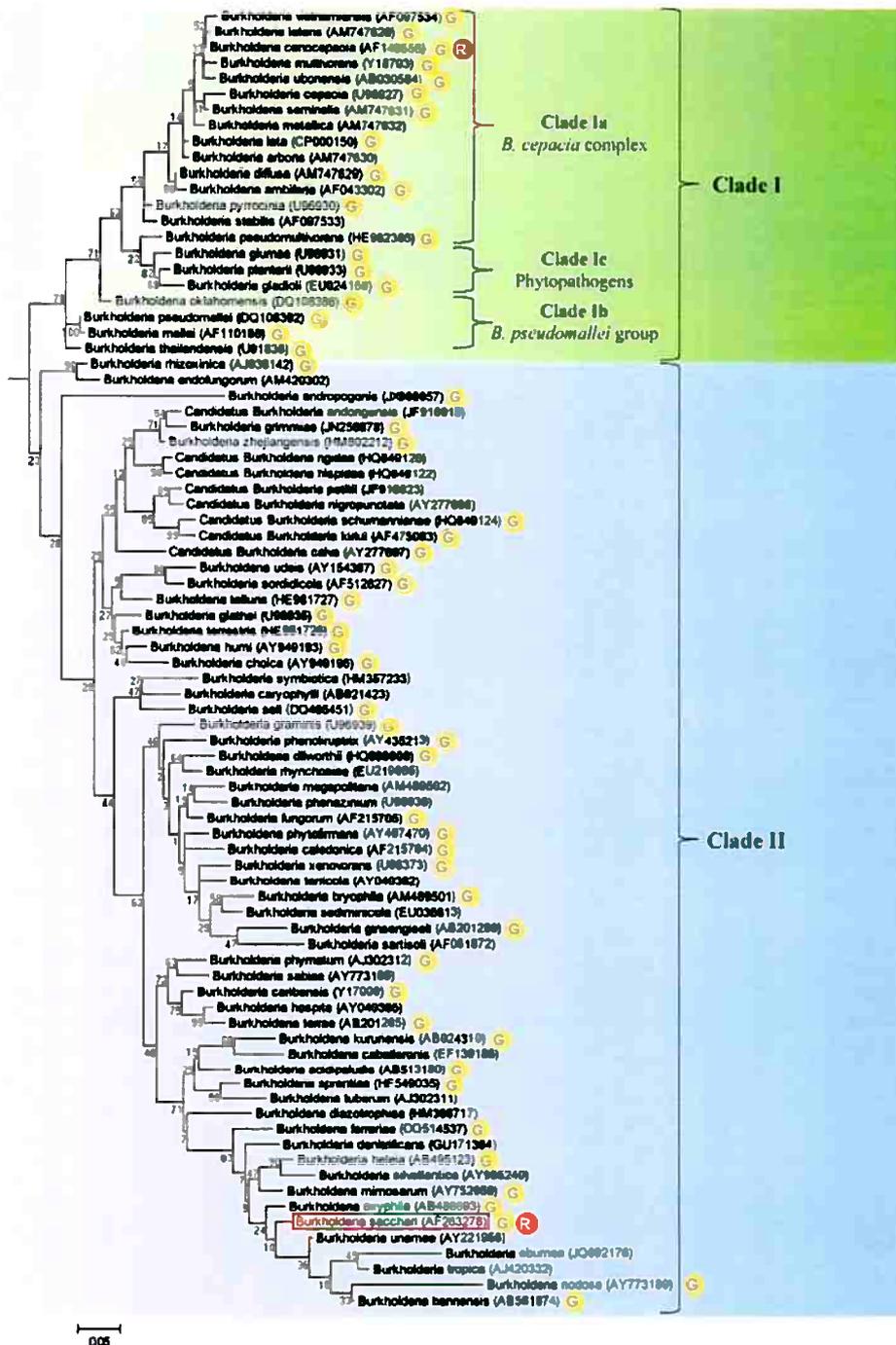


Figura 6.1: Proposta de divisão do gênero *Burkholderia*. Em verde estão as espécies patogênicas do gênero *Burkholderia* mantido e em azul estão as espécies ambientais do novo gênero proposto *Paraburkholderia*. O círculo amarelo indica que a espécie em questão apresenta genoma sequenciado e o círculo vermelho indica que a espécie apresenta rede metabólica em escala genômica reconstruída. Adaptado de Sawana et al. (2014)

6.2 Sugestões para Pesquisas Futuras

O genoma de *B. sacchari* foi recentemente ressequenciado, desta vez usando tecnologia Illumina MiSeq com *reads* pareados. Esse novo sequenciamento poderá ajudar na redução do número de *contigs* obtido com a montagem usando a tecnologia 454. Um exemplo disso é o trabalho de Utturkar *et al.* (2014), que mostrou que a montagem híbrida contendo dados 454 e Illumina supera a montagem individual.

Com relação à anotação e reconstrução metabólica, é necessário que estudos futuros revisem a equação de biomassa que foi proposta computacionalmente. Isso poderia, num primeiro momento, ser feito usando as descrições sobre a composição bioquímica de *B. sacchari* presente no estudo de Brämer *et al.* (2001), mas também deveria ser complementado com outros experimentos. Neste último caso, uma sugestão é seguir as diretrizes do trabalho de Tervo e Reed (2013), onde é proposto um arcabouço computacional para guiar o desenho de experimentos que visam corrigir a equação de formação de biomassa.

Uma outra sugestão é acerca da curadoria manual da reconstrução. Com base nos estudos de Brämer *et al.* (2001); Mendonça *et al.* (2014), seria prontamente possível curar manualmente as vias do propionato, de catabolismo de xilose, e vias de assimilação dos coprodutos usados.

Com relação à representação como um banco de dados em grafo, as sugestões para pesquisas futuras seriam colocar o banco de dados em grafo criado numa aplicação *web* que poderia ser usada por diversos usuários. Seria de grande valia nesse caso tornar as funcionalidades mais intuitivas a usuários que tem pouco conhecimento sobre programação e bancos de dados. Além disso, seria importante para a transparência da reconstrução (Heavner e Price, 2015) que as inferências feitas por homologia pudessem ser distinguidas daquelas inferidas por meio de evidência bioquímica. Nesse sentido, Thiele e Palsson (2010a) propõe um critério de confiança que atribui notas às reações de acordo com o tipo de evidência encontrada. Nessa proposta, o valor 4 se refere à evidência bioquímica, 3 à dados genéticos experimentais, 2 tanto à dados fisiológicos quanto à dados obtidos por inferência de função de produtos gênicos, 1 a inferência mediante modelagem da rede reconstruída e 0 se refere à ausência de evidência. Desse modo, a inclusão desse sistema de classificação poderia ser de fato implementada na base de dados em grafo através da criação de um par chave-valor para a entidade Reação.

Apêndice A

Artigo publicado

Comunicação do sequenciamento e montagem do genoma de *Burkholderia sacchari* na revista *Genome Announcements*.

Draft Genome Sequence of the Polyhydroxyalkanoate-Producing Bacterium *Burkholderia sacchari* LMG 19450 Isolated from Brazilian Sugarcane Plantation Soil

Paulo Moises Raduan Alexandrino,^{a,b} Thatiane Teixeira Mendonça,^a Linda Priscila Guamán Bautista,^b Juliano Cherix,^b Gabriela Cazonato Lozano-Sakalauskas,^b André Fujita,^a Edmar Ramos Filho,^b Paul Long,^{c,d} Gabriel Padilla,^b Marilda Keico Taciro,^b José Gregório Cabrera Gomez,^b Luiziana Ferreira Silva^b

Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil^a; Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil^b; Institute of Pharmaceutical Science, King's College London, London, United Kingdom^c; Faculty of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil^d

***Burkholderia sacchari* LMG 19450, isolated from the soil of a sugarcane plantation in Brazil, accumulates large amounts of polyhydroxyalkanoates from sucrose, xylose, other carbohydrates, and organic acids. We present the draft genome sequence of this industrially relevant bacterium, which is 7.2 Mb in size and has a G+C content of 64%.**

Received 23 March 2015 Accepted 26 March 2015 Published 7 May 2015

Citation Alexandrino PMR, Mendonça TT, Guamán Bautista LP, Cherix J, Lozano-Sakalauskas GC, Fujita A, Ramos Filho E, Long P, Padilla G, Taciro MK, Gomez JCG, Silva LF. 2015. Draft genome sequence of the polyhydroxyalkanoate-producing bacterium *Burkholderia sacchari* LMG 19450 isolated from Brazilian sugarcane plantation soil. *Genome Announc* 3(3):e00313-15. doi:10.1128/genomeA.00313-15.

Copyright © 2015 Alexandrino et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Luiziana Ferreira Silva, lukneif@usp.br.

B*Burkholderia sacchari* LMG 19450 was isolated in the 1990s from the soil of a sugarcane plantation in Brazil (1) and was described as a new species (2). This bacterium has attracted interest from industry because of its capability to metabolize different carbon sources (sucrose, xylose, organic acids, etc.), reach high cell densities, and accumulate high levels of polyhydroxyalkanoates (PHA) (3–5) and also because it is sensitive to a large number of antibiotics (3). The aim of sequencing the genome was to identify genes involved in the catabolism of xylose and other sugars derived from biomass, as well as genes involved in PHA metabolism.

Genomic DNA was extracted using the DNeasy blood and tissue kit (Qiagen). The DNA was concentrated to 353.9 ng/μL, and the quality was assessed by agarose gel electrophoresis and on a NanoDrop spectrophotometer (Thermo Scientific). Whole-genome sequencing was performed by Macrogen (Seoul, South Korea) using the 454 GS FLX sequencing platform, which generated 785,669 reads. The sequencing reads were assembled with Newbler (Roche), and contigs annotation was carried out using the Rapid Annotation Server Subsystem Technology (RAST) (6).

The draft genome is composed of 121 contigs with a total length of 7,265,069 bp (depth of coverage, 49×) and a G+C content of 64.03%. The mean size of the contigs is 60,042 bp, and the N_{50} is 208,943 bp. RAST identified 6,741 coding regions, among which there were genes related to carbohydrate catabolism, e.g., xylose transporter ATP-binding subunit (*xylG*), xylose transporter substrate-binding protein (*xylF*), xylose isomerase (*xylA*), xylulokinase (*xylB*), and xylose operon regulatory protein (*xylR*) for xylose catabolism; to fatty acid catabolism, e.g., acyl-CoA synthetase (*fadD*), acyl-coenzyme A (CoA) dehydrogenase (*fadE*), enoyl-CoA hydratase-S-specific (*fadB-fadJ*); and to PHA metabolism, e.g., polyhydroxyalkanoic acid synthase (*phaC*) and

3-ketoacyl-CoA thiolase (*phaA*). No pathogenesis-related gene was found in *B. sacchari* LMG 19450. Only one gene associated with resistance to antibiotics, encoding undecaprenyl diphosphatase, was found, and it has been associated with resistance to bacitracin (7).

Considering that this bacterium has the potential to convert a wide range of carbon sources, *B. sacchari* LMG 19450 represents a promising candidate for the production of PHA and other bio-based products.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number JTDB00000000. The version described in this paper is the first version, JTDB01000000.

ACKNOWLEDGMENTS

This work was supported by the São Paulo Research Foundation (FAPESP) (grants 10/51989-4 and 12/51533-6), Coordination for the Improvement of Higher Education Personnel (CAPES), and the National Research Council (CNPq), Brazil.

REFERENCES

- Gomez JCG, Rodrigues MFA, Alli RCP, Torres BB, Netto CLB, Oliveira MS, da Silva LF. 1996. Evaluation of soil Gram-negative bacteria yielding polyhydroxyalkanoic acids from carbohydrates and propionic acid. *Appl Microbiol Biotechnol* 45:785–791. <http://dx.doi.org/10.1007/s002530050763>.
- Brämer CO, Vandamme P, da Silva LF, Gomez JG, Steinbüchel A. 2001. Polyhydroxyalkanoate-accumulating bacterium isolated from soil of a sugar-cane plantation in Brazil. *Int J Syst Evol Microbiol* 51:1709–1713. <http://dx.doi.org/10.1099/00207713-51-5-1709>.
- Silva LF, Gomez JCG, Oliveira MS, Torres BB. 2000. Propionic acid metabolism and poly-3-hydroxybutyrate-co-3-hydroxyvalerate, (p 3HB-co-3HV) production by *Burkholderia* sp. *J Biotechnol* 76:165–174. [http://dx.doi.org/10.1016/S0168-1656\(99\)00184-4](http://dx.doi.org/10.1016/S0168-1656(99)00184-4).

4. Silva LF, Taciro MK, Michelin Ramos ME, Carter JM, Pradella JG, Gomez JG. 2004. Poly-3-hydroxybutyrate (P3HB) production by bacteria from xylose, glucose and sugarcane bagasse hydrolysate. *J Ind Microbiol Biotechnol* 31:245–254. <http://dx.doi.org/10.1007/s10295-004-0136-7>.
5. Mendonça TT, Gomez JG, Buffoni E, Sánchez Rodríguez RJ, Schripsema J, Lopes MS, Silva LF. 2014. Exploring the potential of *Burkholderia sacchari* to produce polyhydroxyalkanoates. *J Appl Microbiol* 116:815–829. <http://dx.doi.org/10.1111/jam.12406>.
6. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9:75. <http://dx.doi.org/10.1186/1471-2164-9-75>.
7. El Ghachi M, Derbise A, Bouhss A, Mengin-Lecreulx D. 2005. Identification of multiple genes encoding membrane proteins with undecaprenyl pyrophosphate phosphatase (UppP) activity in *Escherichia coli*. *J Biol Chem* 280:18689–18695. <http://dx.doi.org/10.1074/jbc.M412277200>.

Apêndice B

Créditos das imagens

As imagens foram feitas pelo próprio autor, mas se basearam em desenhos prontos, disponibilizados pela *Servier Medical Art*, sob licença *Creative Commons 3.0*, através do endereço eletrônico <http://www.servier.com/Powerpoint-image-bank>

Referências Bibliográficas

- Alexandrino et al.(2015)** Paulo Moises Raduan Alexandrino, Thatiane Teixeira Mendonça, Linda Priscila Guamán Bautista, Juliano Cherix, Gabriela Cazonato Lozano-Sakalauskas, André Fujita, Edmar Ramos Filho, Paul Long, Gabriel Padilla, Marilda Keico Taciro e Others. Draft Genome Sequence of the Polyhydroxyalkanoate-Producing Bacterium *Burkholderia sacchari* LMG 19450 Isolated from Brazilian Sugarcane Plantation Soil. *Genome Announcements*, 3(3):e00313—15. doi: 10.1128/genomeA.00313-15. Copyright. Citado na pág. 3, 14
- Anderson e Dawes(1990)** A J Anderson e E A Dawes. Occurrence, metabolism, metabolic role, and industrial uses of bacterial polyhydroxyalkanoates. *Microbiological reviews*, 54(4): 450–72. ISSN 0146-0749. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=372789&tool=pmcentrez&rendertype=abstract>. Citado na pág. 1
- Aziz et al.(2008)** Ramy K Aziz, Daniela Bartels, Aaron a Best, Matthew DeJongh, Terrence Disz, Robert a Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M Glass, Michael Kubal, Folker Meyer, Gary J Olsen, Robert Olson, Andrei L Osterman, Ross a Overbeek, Leslie K McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke e Olga Zagnitko. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9:75. ISSN 1471-2164. doi: 10.1186/1471-2164-9-75. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265698&tool=pmcentrez&rendertype=abstract>. Citado na pág. 19
- Baker(2012)** Monya Baker. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4):333–337. ISSN 1548-7091. doi: 10.1038/nmeth.1935. URL <http://dx.doi.org/10.1038/nmeth.1935>. Citado na pág. 8
- Blankenberg et al.(2010)** Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guru-prasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko e James Taylor. Galaxy: a web-based genome analysis tool for experimentalists. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 19:Unit 19.10.1–21. ISSN 1934-3647. doi: 10.1002/0471142727.mb1910s89. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4264107&tool=pmcentrez&rendertype=abstract>. Citado na pág. 14
- Blazeck e Alper(2010)** John Blazeck e Hal Alper. Systems metabolic engineering: Genome-scale models and beyond. *Biotechnology Journal*, 5(7):647–659. ISSN 1860-7314. doi: 10.1002/biot.200900247. URL <http://dx.doi.org/10.1002/biot.200900247>. Citado na pág. 3
- Borodina et al.(2005)** Irina Borodina, Preben Krabben e Jens Nielsen. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome research*, 15(6):820–9. ISSN 1088-9051. doi: 10.1101/gr.3364705. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1142472&tool=pmcentrez&rendertype=abstract>. Citado na pág. 30
- Bramer et al.(2002)** C. O. Bramer, L. F. Silva, J. G. C. Gomez, H. Priefert e A. Steinbuechel. Identification of the 2-Methylcitrate Pathway Involved in the Catabolism of Propionate in the Polyhydroxyalkanoate-Producing Strain *Burkholderia sacchari* IPT101T and Analysis of a Mutant Accumulating a Copolyester with Higher 3-Hydroxyvalerate Content. *Applied and Envi-*

ronmental Microbiology, 68(1):271–279. ISSN 0099-2240. doi: 10.1128/AEM.68.1.271-279.2002. URL <http://aem.asm.org/content/68/1/271.long>. Citado na pág. 2, 19

- Brämer et al.(2001)** Christian O Brämer, Peter Vandamme, Luiziana F da Silva, J G Gomez e Alexander Steinbüchel. Polyhydroxyalkanoate-accumulating bacterium isolated from soil of a sugar-cane plantation in Brazil. *International journal of systematic and evolutionary microbiology*, 51(5):1709–1713. Citado na pág. 19, 36, 37, 53
- Brettin et al.(2015)** Thomas Brettin, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Gary J Olsen, Robert Olson, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, James A Thomason, Rick Stevens, Veronika Vonstein, Alice R Wattam e Fangfang Xia. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, 5:8365. ISSN 2045-2322. doi: 10.1038/srep08365. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4322359&tool=pmcentrez&rendertype=abstract>. Citado na pág. 19
- Earl et al.(2011)** Dent Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R Zerbino, Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne, Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T Simpson, Nuno A Fonseca, ?nanç Birol, T Roderick Docking, Isaac Y Ho, Daniel S Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis, Delphine Naquin, Nicolas Maillet, Michael C Schatz, David R Kelley, Adam M Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu, Wen-Chi Chou, Anuj Srivastava, Timothy I Shaw, J Graham Ruby, Peter Skewes-Cox, Miguel Betegon, Michelle T Dimon, Victor Solovyev, Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez, Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li, Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall, Paul J Kersey, Richard Durbin, Shaun D Jackman, Jarrod A Chapman, Xiaoqiu Huang, Joseph L DeRisi, Mario Caccamo, Yingrui Li, David B Jaffe, Richard E Green, David Haussler, Ian Korf e Benedict Paten. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, 21(12):2224–41. ISSN 1549-5469. doi: 10.1101/gr.126599.111. URL <http://genome.cshlp.org/content/21/12/2224.short>. Citado na pág. 13
- Edwards e Holt(2013)** David J Edwards e Kathryn E Holt. Beginner’s guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial informatics and experimentation*, 3(1):2. ISSN 2042-5783. doi: 10.1186/2042-5783-3-2. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3630013&tool=pmcentrez&rendertype=abstract>. Citado na pág. 14
- Edwards e Palsson(1999)** J S Edwards e B O Palsson. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *The Journal of biological chemistry*, 274(25):17410–6. ISSN 0021-9258. URL <http://www.ncbi.nlm.nih.gov/pubmed/10364169>. Citado na pág. 29
- Edwards e Palsson(2000a)** J. S. Edwards e B. O. Palsson. The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533. ISSN 0027-8424. doi: 10.1073/pnas.97.10.5528. URL <http://www.pnas.org/content/97/10/5528.full>. Citado na pág. 29
- Edwards e Palsson(2000b)** J S Edwards e B O Palsson. Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC bioinformatics*, 1:1. ISSN 1471-2105. doi: 10.1186/1471-2105-1-1. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=29061&tool=pmcentrez&rendertype=abstract>. Citado na pág. 29
- Edwards et al.(2001)** J S Edwards, R U Ibarra e B O Palsson. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature biotechnology*, 19(2):125–30. ISSN 1087-0156. doi: 10.1038/84379. URL <http://www.ncbi.nlm.nih.gov/pubmed/11175725>. Citado na pág. 29

- Eklblom e Wolf(2014)** Robert Eklblom e Jochen B. W. Wolf. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, 7(9):n/a–n/a. ISSN 17524571. doi: 10.1111/eva.12178. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4231593&tool=pmcentrez&rendertype=abstract>. Citado na pág. 13
- Estrada-de los Santos et al.(2013)** Paulina Estrada-de los Santos, Pablo Vinuesa, Lourdes Martínez-Aguilar, Ann M Hirsch e Jesús Caballero-Mellado. Phylogenetic analysis of burkholderia species by multilocus sequence analysis. *Current microbiology*, 67(1):51–60. ISSN 1432-0991. doi: 10.1007/s00284-013-0330-9. URL <http://www.ncbi.nlm.nih.gov/pubmed/23404651>. Citado na pág. 19
- Feist et al.(2009)** Adam M Feist, Markus J Herrgård, Ines Thiele, Jennie L Reed e Bernhard Ø Pals-son. Reconstruction of biochemical networks in microorganisms. *Nature reviews. Microbiology*, 7(2):129–43. ISSN 1740-1534. doi: 10.1038/nrmicro1949. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3119670&tool=pmcentrez&rendertype=abstract>. Citado na pág. 29, 30, 44
- Finotello et al.(2012)** Francesca Finotello, Enrico Lavezzo, Paolo Fontana, Denis Peruzzo, Alessandro Albiero, Luisa Barzon, Marco Falda, Barbara Di Camillo e Stefano Toppo. Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data. *Briefings in Bioinformatics*, 13(3):269–280. ISSN 14675463. doi: 10.1093/bib/bbr063. Citado na pág. 13, 15, 16, 17
- Fleischmann et al.(1995)** R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty e J M Merrick. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512. ISSN 0036-8075. URL <http://www.ncbi.nlm.nih.gov/pubmed/7542800>. Citado na pág. 29
- Galens et al.(2011)** Kevin Galens, Joshua Orvis, Sean Daugherty, Heather H Creasy, Sam Angiuoli, Owen White, Jennifer Wortman, Anup Mahurkar e Michelle Gwinn Giglio. The IGS Standard Operating Procedure for Automated Prokaryotic Annotation. *Standards in genomic sciences*, 4(2):244–51. ISSN 1944-3277. doi: 10.4056/sigs.1223234. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3111993&tool=pmcentrez&rendertype=abstract>. Citado na pág. 8
- Gallant et al.(1980)** John Gallant, David Maier e James Astorer. On finding minimal length superstrings. *Journal of Computer and System Sciences*, 20(1):50–58. ISSN 00220000. doi: 10.1016/0022-0000(80)90004-5. URL <http://www.sciencedirect.com/science/article/pii/0022000080900045>. Citado na pág. 11
- Ganter et al.(2013)** Mathias Ganter, Thomas Bernard, Sébastien Moretti, Joerg Stelling e Marco Pagni. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics (Oxford, England)*, 29(6):815–6. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt036. URL <http://bioinformatics.oxfordjournals.org/content/29/6/815>. Citado na pág. 31
- Ghodsí et al.(2013)** Mohammadreza Ghodsí, Christopher M Hill, Irina Astrovskaya, Henry Lin, Dan D Sommer, Sergey Koren e Mihai Pop. De novo likelihood-based measures for comparing genome assemblies. *BMC research notes*, 6(1):334. ISSN 1756-0500. doi: 10.1186/1756-0500-6-334. URL <http://bmcresearchnotes.biomedcentral.com/articles/10.1186/1756-0500-6-334>. Citado na pág. 12
- Giardine et al.(2005)** Belinda Giardine, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, Daniel Blankenberg, Istvan Albert, James Taylor, Webb Miller, W James Kent e Anton Nekrutenko. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451–5. ISSN 1088-9051. doi: 10.1101/gr.4086505. URL <http://genome.cshlp.org/content/15/10/1451.long>. Citado na pág. 14

- Gilles et al.(2011)** André Gilles, Emese Megléc, Nicolas Pech, Stéphanie Ferreira, Thibaut Malausa e Jean-François Martin. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, 12(1):245. ISSN 1471-2164. doi: 10.1186/1471-2164-12-245. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3116506&tool=pmcentrez&rendertype=abstract>. Citado na pág. 17
- Goecks et al.(2010)** Jeremy Goecks, Anton Nekrutenko e James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86. ISSN 1465-6914. doi: 10.1186/gb-2010-11-8-r86. URL <http://genomebiology.com/2010/11/8/R86>. Citado na pág. 14
- Gombert e Nielsen(2000)** Andreas Karoly Gombert e Jens Nielsen. Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11(2):180–186. ISSN 09581669. doi: 10.1016/S0958-1669(00)00079-3. Citado na pág. 10
- Gomez et al.(1996)** J. G. C. Gomez, M. F. A. Rodrigues, R. C. P. Alli, B. B. Torres, C. L. Bueno Netto, M. S. Oliveira e L. F. da Silva. Evaluation of soil gram-negative bacteria yielding polyhydroxyalkanoic acids from carbohydrates and propionic acid. *Applied Microbiology and Biotechnology*, 45(6):785–791. ISSN 0175-7598. doi: 10.1007/s002530050763. URL <http://link.springer.com/10.1007/s002530050763>. Citado na pág. 1, 2
- Gurevich et al.(2013)** Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi e Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075. ISSN 13674803. doi: 10.1093/bioinformatics/btt086. Citado na pág. 19
- Hamilton e Reed(2014)** Joshua J. Hamilton e Jennifer L. Reed. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, 16(1):49–59. ISSN 14622912. doi: 10.1111/1462-2920.12312. Citado na pág. 32, 44
- Heavner e Price(2015)** Benjamin D Heavner e Nathan D Price. Transparency in metabolic network reconstruction enables scalable biological discovery. *Current Opinion in Biotechnology*, 34:105–109. ISSN 09581669. doi: 10.1016/j.copbio.2014.12.010. URL <http://linkinghub.elsevier.com/retrieve/pii/S0958166914002250>. Citado na pág. 44, 50, 53
- Heiner et al.(2013)** Cheryl Heiner, Susana Wang, Meredith Ashby, Yan Guo, Jason Underwood e Primo Baybayan. Greater than 10 kb Read Lengths Routine when Sequencing with Pacific Biosciences' XL Release. *Journal of Biomolecular Techniques : JBT*, 24(Suppl):S43–S43. ISSN 1524-0215. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3635392/>. Citado na pág. 7
- Hernandez et al.(2008)** David Hernandez, Patrice François, Laurent Farinelli, Magne Osterås e Jacques Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research*, 18(5):802–9. ISSN 1088-9051. doi: 10.1101/gr.072033.107. URL <http://genome.cshlp.org/content/18/5/802.short>. Citado na pág. 7
- Howison et al.(2013)** Mark Howison, Felipe Zapata e Casey W Dunn. Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics (Oxford, England)*, 29(23):2959–63. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt525. URL <http://www.ncbi.nlm.nih.gov/pubmed/24021385>. Citado na pág. 12
- Huse et al.(2007)** Susan M Huse, Julie a Huber, Hilary G Morrison, Mitchell L Sogin e David Mark Welch. Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biology*, 8(7):R143. ISSN 1465-6906. doi: 10.1186/gb-2007-8-7-r143. URL [http://genomebiology.com/2007/8/7/R143/abstract%delimit%026E30F%nd%delimit%026E30F\\$journals%delimit%026E30F\\$huse2007.pdf](http://genomebiology.com/2007/8/7/R143/abstract%delimit%026E30F%nd%delimit%026E30F$journals%delimit%026E30F$huse2007.pdf). Citado na pág. 16

- Kaiser et al.(2003)** Olaf Kaiser, Daniela Bartels, Thomas Bekel, Alexander Goesmann, Sebastian Kespohl, Alfred Pühler e Folker Meyer. Whole genome shotgun sequencing guided by bioinformatics pipelines—an optimized approach for an established technique. *Journal of Biotechnology*, 106(2-3):121-133. ISSN 01681656. doi: 10.1016/j.jbiotec.2003.08.008. URL <http://www.sciencedirect.com/science/article/pii/S0168165603002293>. Citado na pág. 7
- Karp et al.(2002)** Peter D Karp, Suzanne Paley e Pedro Romero. The Pathway Tools software. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S225-32. ISSN 1367-4803. URL <http://www.ncbi.nlm.nih.gov/pubmed/12169551>. Citado na pág. 30
- Karp et al.(2010)** Peter D Karp, Suzanne M Paley, Markus Krummenacker, Mario Latendresse, Joseph M Dale, Thomas J Lee, Pallavi Kaipa, Fred Gilham, Aaron Spaulding, Liviu Popescu, Tomer Altman, Ian Paulsen, Ingrid M Keseler e Ron Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11(1): 40-79. ISSN 1477-4054. doi: 10.1093/bib/bbp043. URL <http://bib.oxfordjournals.org/content/11/1/40>. Citado na pág. 30
- Kisand e Lettieri(2013)** Veljo Kisand e Teresa Lettieri. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. *BMC genomics*, 14(1): 211. ISSN 1471-2164. doi: 10.1186/1471-2164-14-211. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3618134&tool=pmcentrez&rendertype=abstract>. Citado na pág. 7
- Knuf e Nielsen(2012)** Christoph Knuf e Jens Nielsen. Aspergilli: systems biology and industrial applications. *Biotechnology journal*, 7(9):1147-55. ISSN 1860-7314. doi: 10.1002/biot.201200169. URL <http://www.ncbi.nlm.nih.gov/pubmed/22890866>. Citado na pág. 3, 5
- Koren et al.(2014)** Sergey Koren, Todd J Treangen, Christopher M Hill, Mihai Pop e Adam M Phillippy. Automated ensemble assembly and validation of microbial genomes. *BMC bioinformatics*, 15(1):126. ISSN 1471-2105. doi: 10.1186/1471-2105-15-126. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4030574&tool=pmcentrez&rendertype=abstract>. Citado na pág. 12, 13
- Kumar et al.(2012)** Akhil Kumar, Patrick F Suthers e Costas D Maranas. MetRxn: a knowledge-base of metabolites and reactions spanning metabolic models and databases. *BMC bioinformatics*, 13:6. ISSN 1471-2105. doi: 10.1186/1471-2105-13-6. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3277463&tool=pmcentrez&rendertype=abstract>. Citado na pág. 31
- Kumar e Blaxter(2010)** Sujai Kumar e Mark L Blaxter. Comparing de novo assemblers for 454 transcriptome data. *BMC genomics*, 11(1):571. ISSN 1471-2164. doi: 10.1186/1471-2164-11-571. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3091720&tool=pmcentrez&rendertype=abstract>. Citado na pág. 15
- Lee et al.(2011)** Jeong Wook Lee, Tae Yong Kim, Yu Sin Jang, Sol Choi e Sang Yup Lee. Systems metabolic engineering for chemicals and materials. *Trends in Biotechnology*, 29(8):370-378. ISSN 01677799. doi: 10.1016/j.tibtech.2011.04.001. URL <http://dx.doi.org/10.1016/j.tibtech.2011.04.001>. Citado na pág. 3
- Lemaire et al.(2012)** Benny Lemaire, Sandra Van Oevelen, Petra De Block, Brecht Verstraete, Erik Smets, Els Prinsen e Steven Dessein. Identification of the bacterial endosymbionts in leaf nodules of *Pavetta* (Rubiaceae). *International journal of systematic and evolutionary microbiology*, 62(Pt 1):202-9. ISSN 1466-5034. doi: 10.1099/ijs.0.028019-0. URL <http://ijs.microbiologyresearch.org/content/journal/ijssem/10.1099/ijs.0.028019-0>. Citado na pág. 19
- Lemoigne(1926)** Maurice Lemoigne. Products of dehydration and of polymerization of β -hydroxybutyric acid. *Bull Soc Chem Biol*, 8:770-782. Citado na pág. 1

- Lepidi(1972) A.A. Lepidi. Metabolism of poly-beta-hydroxybutyrate in the soil and in the rhizosphere. *Agricoltura Italiana*, 72:166–184. Citado na pág. 1
- Llaneras e Picó(2008) Francisco Llaneras e Jesús Picó. Stoichiometric modelling of cell metabolism. *Journal of Bioscience and Bioengineering*, 105(1):1–11. ISSN 13891723. doi: 10.1263/jbb.105.1. URL <http://linkinghub.elsevier.com/retrieve/pii/S1389172308700173>. Citado na pág. 10
- Loman e Pallen(2015) Nicholas J. Loman e Mark J. Pallen. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology*, 13(12):787–794. ISSN 1740-1526. doi: 10.1038/nrmicro3565. URL <http://dx.doi.org/10.1038/nrmicro3565>. Citado na pág. 7
- Lopes *et al.*(2009) Mateus Schreiner Garcez Lopes, José Gregório Cabrera Gomez e Luiziana Ferreira Silva. Cloning and overexpression of the xylose isomerase gene from *Burkholderia sacchari* and production of polyhydroxybutyrate from xylose. *Canadian journal of microbiology*, 55(8):1012–5. ISSN 1480-3275. doi: 10.1139/w09-055. URL <http://www.nrcresearchpress.com/doi/abs/10.1139/w09-055?url{ }ver=Z39.88-2003{&rfr{ }id=ori{ }3Arid{ }3Acrossref.org{&rfr{ }dat=cr{ }pub{ }3Dpubmed{&}{#}.VnLI{ }HYrLCI>. Citado na pág. 2, 5, 19
- Maier(1978) David Maier. The Complexity of Some Problems on Subsequences and Supersequences. *J. ACM*, 25(2):322–336. ISSN 0004-5411. doi: 10.1145/322063.322075. URL <http://doi.acm.org/10.1145/322063.322075>. Citado na pág. 11
- Maxam e Gilbert(1977) A M Maxam e W Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4. ISSN 0027-8424. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=392330{& }tool=pmcentrez{& }rendertype=abstract>. Citado na pág. 7
- Medvedev *et al.*(2007) Paul Medvedev, Konstantinos Georgiou, Gene Myers e Michael Brudno. *Computability of models for sequence assembly*, chapter Computabil, páginas 289–301. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-74126-8. doi: 10.1007/978-3-540-74126-8{ }27. URL <http://dx.doi.org/10.1007/978-3-540-74126-8{ }27>. Citado na pág. 12
- Mendonça *et al.*(2014) T. T. Mendonça, J. G C Gomez, E. Buffoni, R. J. Sánchez Rodriguez, J. Schripsema, M. S G Lopes e L. F. Silva. Exploring the potential of *Burkholderia sacchari* to produce polyhydroxyalkanoates. *Journal of Applied Microbiology*, 116:815–829. ISSN 13652672. doi: 10.1111/jam.12406. Citado na pág. 53
- Merrick e Doudoroff(1964) J. M. Merrick e M. Doudoroff. Depolymerization of poly-beta-hydroxybutyrate by intracellular enzyme system. *Journal of bacteriology*, 88:60–71. ISSN 0021-9193. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=277257{& }tool=pmcentrez{& }rendertype=abstract>. Citado na pág. 1
- Miller *et al.*(2010) Jason R Miller, Sergey Koren e Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327. Citado na pág. 7
- Monk *et al.*(2014) Jonathan Monk, Juan Nogales e Bernhard O Palsson. Optimizing genome-scale network reconstructions. *Nature Biotechnology*, 32(5):447–452. ISSN 1087-0156. doi: 10.1038/nbt.2870. URL <http://www.nature.com/doi/10.1038/nbt.2870>. Citado na pág. 30, 31
- Nagarajan e Pop(2009) Niranjan Nagarajan e Mihai Pop. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 16(7):897–908. ISSN 1557-8666. doi: 10.1089/cmb.2009.0005. URL <http://www.ncbi.nlm.nih.gov/pubmed/19580519>. Citado na pág. 11, 12

- Nagarajan e Pop(2013)** Niranjan Nagarajan e Mihai Pop. Sequence assembly demystified. *Nature reviews. Genetics*, 14(3):157–67. ISSN 1471-0064. doi: 10.1038/nrg3367. URL <http://dx.doi.org/10.1038/nrg3367>. Citado na pág. 12, 13
- Narzisi e Mishra(2011)** Giuseppe Narzisi e Bud Mishra. Comparing de novo genome assembly: the long and short of it. *PloS one*, 6(4):e19175. ISSN 1932-6203. doi: 10.1371/journal.pone.0019175. URL <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019175>. Citado na pág. 12, 13
- Oberhardt et al.(2009)** Matthew a Oberhardt, Bernhard Ø Palsson e Jason a Papin. Applications of genome-scale metabolic reconstructions. *Molecular systems biology*, 5(320):320. ISSN 1744-4292. doi: 10.1038/msb.2009.77. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2795471&tool=pmcentrez&rendertype=abstract>. Citado na pág. 3
- Overbeek et al.(2014)** Ross Overbeek, Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, Veronika Vonstein, Alice R Wattam, Fangfang Xia e Rick Stevens. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research*, 42(Database issue):D206–14. ISSN 1362-4962. doi: 10.1093/nar/gkt1226. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965101&tool=pmcentrez&rendertype=abstract>. Citado na pág. 8, 19
- Pereira et al.(2009)** Erica Mendes Pereira, Sonia Regina Silva-Queiroz, José Gregório Cabrera Gomez e Luiziana Ferreira Silva. Disruption of the 2-methylcitric acid cycle and evaluation of poly-3-hydroxybutyrate-co-3-hydroxyvalerate biosynthesis suggest alternate catabolic pathways of propionate in *Burkholderia sacchari*. *Canadian journal of microbiology*, 55(6):688–97. ISSN 1480-3275. doi: 10.1139/w09-018. URL <http://www.ncbi.nlm.nih.gov/pubmed/19767840>. Citado na pág. 2
- Pfau et al.(2015)** Thomas Pfau, Maria Pires Pacheco e Thomas Sauter. Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond. *Briefings in Bioinformatics*, (October):1–10. doi: 10.1093/bib/bbv100. Citado na pág. 31
- Phillippy et al.(2008)** Adam M Phillippy, Michael C Schatz e Mihai Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome biology*, 9(3):R55. ISSN 1474-760X. doi: 10.1186/gb-2008-9-3-r55. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2397507&tool=pmcentrez&rendertype=abstract>. Citado na pág. 13
- Pop(2009)** Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366. Citado na pág. 7
- Räihä e Ukkonen(1981)** Kari-Jouko Räihä e Esko Ukkonen. The shortest common supersequence problem over binary alphabet is NP-complete. *Theoretical Computer Science*, 16(2): 187–198. ISSN 03043975. doi: 10.1016/0304-3975(81)90075-X. URL <http://www.sciencedirect.com/science/article/pii/030439758190075X>. Citado na pág. 11
- Ravikrishnan e Raman(2015)** Aarthi Ravikrishnan e Karthik Raman. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Briefings in bioinformatics*, páginas bbv003–. ISSN 1477-4054. doi: 10.1093/bib/bbv003. URL <http://bib.oxfordjournals.org/content/early/2015/02/27/bib.bbv003.full>. Citado na pág. 30, 31, 44
- Reed et al.(2006)** Jennifer L. Reed, Iman Famili, Ines Thiele e Bernhard O. Palsson. Towards multidimensional genome annotation. *Nature Reviews Genetics*, 7(2):130–141. ISSN 1471-0056. doi: 10.1038/nrg1769. URL <http://www.nature.com/doi/10.1038/nrg1769>. Citado na pág. 8, 9, 30

- Robinson et al.(2015)** Ian Robinson, Jim Webber e Emil Eifrem. *Graph Databases*. O'Reilly. ISBN 9781491930892. Citado na pág. 44
- Rolfsson e Pálsson(2015)** Óttar Rolfsson e Bernhard Ø Pálsson. Decoding the jargon of bottom-up metabolic systems biology. *BioEssays*, 37(6):588–591. ISSN 02659247. doi: 10.1002/bies.201400187. URL <http://doi.wiley.com/10.1002/bies.201400187>. Citado na pág. 10
- Salzberg et al.(2012)** Steven L Salzberg, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, Michael C Schatz, Arthur L Delcher, Michael Roberts, Guillaume Marçais, Mihai Pop e James A Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22(3):557–67. ISSN 1549-5469. doi: 10.1101/gr.131383.111. URL <http://genome.cshlp.org/content/22/3/557.long>. Citado na pág. 13
- Sanger e Coulson(1975)** F Sanger e A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–8. ISSN 0022-2836. URL <http://www.ncbi.nlm.nih.gov/pubmed/1100841>. Citado na pág. 7
- Sawana et al.(2014)** Amandeep Sawana, Mobolaji Adeolu e Radhey S Gupta. Molecular signatures and phylogenomic analysis of the genus Burkholderia: proposal for division of this genus into the emended genus Burkholderia containing pathogenic organisms and a new genus Paraburkholderia gen. nov. harboring environmental species. *Frontiers in genetics*, 5:429. ISSN 1664-8021. doi: 10.3389/fgene.2014.00429. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4271702&tool=pmcentrez&rendertype=abstract>. Citado na pág. 51, 52
- Schmieder e Edwards(2011)** R. Schmieder e R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr026. URL <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr026>. Citado na pág. 14, 16
- Seemann(2014)** Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14):2068–9. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu153. URL <http://www.ncbi.nlm.nih.gov/pubmed/24642063>. Citado na pág. 8
- Shendure e Ji(2008)** Jay Shendure e Hanlee Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45. ISSN 1546-1696. doi: 10.1038/nbt1486. URL <http://dx.doi.org/10.1038/nbt1486>. Citado na pág. 7, 8
- Silva(2000)** L Silva. Propionic acid metabolism and poly-3-hydroxybutyrate-co-3-hydroxyvalerate (P3HB-co-3HV) production by Burkholderia sp. *Journal of Biotechnology*, 76(2-3):165–174. ISSN 01681656. doi: 10.1016/S0168-1656(99)00184-4. URL <http://www.sciencedirect.com/science/article/pii/S0168165699001844>. Citado na pág. 2, 5
- Simpson e Pop(2015)** Jared T Simpson e Mihai Pop. The Theory and Practice of Genome Sequence Assembly. *Annual review of genomics and human genetics*, 16:153–72. ISSN 1545-293X. doi: 10.1146/annurev-genom-090314-050032. URL <http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-090314-050032>. Citado na pág. 7, 12
- Staden(1979)** R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–10. ISSN 0305-1048. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=327874&tool=pmcentrez&rendertype=abstract>. Citado na pág. 7
- Stein(2001)** L Stein. Genome annotation: from sequence to biology. *Nature reviews. Genetics*, 2(7):493–503. ISSN 1471-0056. doi: 10.1038/35080529. URL <http://www.ncbi.nlm.nih.gov/pubmed/11433356>. Citado na pág. 7

- Steinbüchel e Fuchtenbusch(1998)** Alexander Steinbüchel e Bernd Fuchtenbusch. Bacterial and other biological systems for polyester production. *Trends in Biotechnology*, 16(10):419–427. ISSN 01677799. doi: 10.1016/S0167-7799(98)01194-9. URL <http://www.sciencedirect.com/science/article/pii/S0167779998011949>. Citado na pág. 1
- Sudesh et al.(2000)** K Sudesh, H Abe e Y Doi. Synthesis, structure and properties of polyhydroxyalkanoates: biological polyesters. *Progress in Polymer Science*, 25(10):1503–1555. ISSN 00796700. doi: 10.1016/S0079-6700(00)00035-6. URL <http://www.sciencedirect.com/science/article/pii/S0079670000000356>. Citado na pág. 1
- Tee et al.(2014)** Ting Wei Tee, Anupam Chowdhury, Costas D. Maranas e Jacqueline V. Shanks. Systems metabolic engineering design: Fatty acid production as an emerging case study. *Biotechnology and Bioengineering*, 111(5):849–857. ISSN 10970290. doi: 10.1002/bit.25205. Citado na pág. 3
- Tervo e Reed(2013)** Christopher J Tervo e Jennifer L Reed. BioMog: a computational framework for the de novo generation or modification of essential biomass components. *PLoS one*, 8(12): e81322. ISSN 1932-6203. doi: 10.1371/journal.pone.0081322. URL <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081322>. Citado na pág. 53
- Thiele e Palsson(2010a)** Ines Thiele e Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121. ISSN 1750-2799. doi: 10.1038/nprot.2009.203. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125167&tool=pmcentrez&rendertype=abstract>. Citado na pág. 3, 30, 44, 53
- Thiele e Palsson(2010b)** Ines Thiele e Bernhard Ø Palsson. Reconstruction annotation jamborees: a community approach to systems biology. *Molecular systems biology*, 6:361. ISSN 1744-4292. doi: 10.1038/msb.2010.15. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872611&tool=pmcentrez&rendertype=abstract>. Citado na pág. 44
- Utturkar et al.(2014)** Sagar M Utturkar, Dawn M Klingeman, Miriam L Land, Christopher W Schadt, Mitchel J Doktycz, Dale A Pelletier e Steven D Brown. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics (Oxford, England)*, 30(19):2709–16. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu391. URL <http://bioinformatics.oxfordjournals.org/content/30/19/2709.abstract>. Citado na pág. 53
- Van Domselaar et al.(2005)** Gary H Van Domselaar, Paul Stothard, Savita Shrivastava, Joseph A Cruz, AnChi Guo, Xiaoli Dong, Paul Lu, Duane Szafron, Russ Greiner e David S Wishart. BASys: a web server for automated bacterial genome annotation. *Nucleic acids research*, 33(Web Server issue):W455–9. ISSN 1362-4962. doi: 10.1093/nar/gki593. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160269&tool=pmcentrez&rendertype=abstract>. Citado na pág. 8
- Vezi et al.(2012a)** Francesco Vezi, Giuseppe Narzisi e Bud Mishra. Feature-by-feature-evaluating de novo sequence assembly. *PLoS one*, 7(2):e31002. ISSN 1932-6203. doi: 10.1371/journal.pone.0031002. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3272011&tool=pmcentrez&rendertype=abstract>. Citado na pág. 13, 19
- Vezi et al.(2012b)** Francesco Vezi, Giuseppe Narzisi e Bud Mishra. Reevaluating Assembly Evaluations with Feature Response Curves: GAGE and Assemblathons. *PLoS ONE*, 7(12):1–11. ISSN 19326203. doi: 10.1371/journal.pone.0052210. Citado na pág. 13
- Wallen e Rohwedder(1974)** Lowell L. Wallen e William K. Rohwedder. Poly- β -hydroxyalkanoate from activated sludge. *Environmental Science & Technology*, 8(6):576–579. ISSN 0013-936X. doi: 10.1021/es60091a007. URL <http://dx.doi.org/10.1021/es60091a007>. Citado na pág. 1