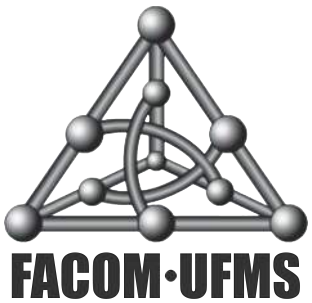
Determinação de trechos genômicos específicos para seleção de oligonucleotídeos iniciadores usando famílias de proteínas e sequências características

Dissertação de Mestrado

Rodrigo Andrade Cardoso

Orientação: Prof. Nalvo Franco de Almeida Jr.

Área de concentração: Bioinformática



Faculdade de Computação

Universidade Federal de Mato Grosso do Sul

Campo Grande, Outubro de 2015

Resumo

Com o sequenciamento mais frequente de genomas, várias técnicas de comparação têm sido propostas, com inúmeras aplicações. Este trabalho propõe uma técnica computacional baseada em sequências características e na construção de famílias de proteínas para determinar trechos de um genoma que contenham candidatos a oligonucleotídeos iniciadores específicos desse genoma, quando comparado com outros. Os trechos determinados pela metodologia proposta podem ser usados como entrada para ferramentas que encontram oligonucleotídeos iniciadores específicos. Testes revelaram que a metodologia se mostrou muito efetiva para genomas de espécies diferentes.

Palavras-chave: genômica, genômica comparativa, oligonucleotídeos iniciadores, famílias de proteínas, sequências características

Agradecimentos

Aos meus pais e irmãs pelo incentivo, apoio e dedicação que fizeram tornar possível mais esta conquista.

À Camila que, à sua maneira, me deu apoio e compreendeu as diversas madrugadas e os muitos finais de semana dedicados a este trabalho.

Ao João e à Silvia pelo incentivo, apoio e preocupação em todos os momentos.

Ao professor Dr. Nalvo pela orientação, por tudo que me ensinou nesses anos. Agradeço, também, pela confiança em mim depositada para a realização deste trabalho.

Agradeço à professora Dra. Luciana Montera pelas sugestões e correções.

Ao professor Dr. Flábio Araújo por tornar possível a realização deste trabalho de mestrado e por suas contribuições ao longo deste processo.

Aos colegas de trabalho do IFMS/Coxim sempre muito solícitos nas muitas vezes que precisei me ausentar do *Campus*.

Enfim, agradeço a todos que direta ou indiretamente contribuíram para a realização deste trabalho.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

Nas últimas décadas, a Biologia Computacional vem ganhando notoriedade e importância, contribuindo de maneira considerável para a melhoria e desenvolvimento de processos em problemas cruciais, relacionados à Biologia Molecular e Genômica, principalmente através das ferramentas de comparação de genomas.

Técnicas moleculares tem sido amplamente utilizadas para sequenciamento genético, evolução molecular, entre outros objetivos. Dentre elas, destaque para a Reação em Cadeia da Polimerase (PCR), concebida no início da década de 1980 por Kary Mullis e que tem sido amplamente utilizada em laboratórios, tanto médicos quanto biológicos, para as mais diversas tarefas.

Basicamente, a PCR amplifica regiões específicas de um genoma, para que isso seja possível, necessita dos chamados oligonucleotídeos iniciadores, ou *primers*, que irão delimitar a região que será amplificada. Oligonucleotídeo iniciador é um oligonucleotídeo sintético curto e sua sequência é correspondente ao complemento de uma região no DNA alvo [18], logo, para amplificar uma determinada região do genoma é necessário um par de oligonucleotídeos iniciadores.

Muitos avanços nas áreas de genômica comparativa, metagenômica e transcritômica estão sendo obtidos graças ao advento das técnicas de sequenciamento de nova geração (*Next Generation Sequencing*, NGS) [33]. Em particular, a partir do uso de NGS, é posível obter, a custos muito mais reduzidos do que no passado, sequências genômicas inteiras de bactérias, por exemplo. Com isso, torna-se possível usar essas sequências com o objetivo de compará-las e determinar aspectos funcionais específicos e comuns entre os organismos comparados.

O problema de comparar sequências genômicas torna-se complicado quando elas são muito similares, como é o caso de algumas espécies muito próximas evolutivamente, ou mais ainda quando se trata de comparar sequências genômicas de cepas de uma mesma espécie. Neste contexto, o problema de se determinar oligonucleotídeos iniciadores específicos para genomas próximos é um grande desafio na bioinformática.

A motivação deste trabalho está exatamente relacionada à necessidade de se obter uma ferramenta computacional que receba como entrada uma sequência genômica alvo e também um conjunto de outras sequências genômicas, e que seja capaz de determinar trechos da sequência alvo que contenham candidatos promissores a oligonucleotídeos iniciadores específicos da sequência alvo com relação às outras sequências. Importante salientar que não é objeto deste trabalho encontrar oligonucleotídeos iniciadores específicos, mas sim trechos que contenham potenciais oligonucleotídeos iniciadores específicos. Assim, os trechos encontrados devem ser usados como entrada para ferramentas que determinam tais oligonucleotídeos iniciadores, como por exemplo *Primer-BLAST* [44].

A metodologia proposta tem como diferencial a utilização de duas abordagens completamente independentes: famílias de proteínas e sequências características. A primeira é uma abordagem originada na bioinformática a partir da comparação

envolvendo todos os genes de proteínas dos genomas, cujo objetivo é determinar conjuntos de proteínas funcionalmente relacionadas. A segunda é uma técnica computacional de comparação de sequências que tenta encontrar sequências específicas, a partir de uma abordagem conhecida como "k diferenças".

Resumidamente, o método funciona da seguinte forma. A partir das famílias dos genes de proteínas dos genomas comparados, é possível determinar regiões do genoma alvo que contenham os chamados genes *singletons*, que são genes que não participam de qualquer família de genes de quaisquer dos genomas comparados e que portanto são regiões do genoma alvo potencialmente específicas. Para cada uma dessas regiões, nosso método procura blocos pequenos (esses blocos serão candidatos a oligonucleotídeos iniciadores) que acontecem no genoma alvo e que ocorrem nas outras sequências com pelo menos k diferenças. Esses blocos com k diferenças são o objeto de uma conhecida técnica computacional, que consiste na procura das tais sequências características. A partir da determinação desses blocos, o método finalmente retorna trechos contendo conjuntos de blocos próximos entre si e que sejam maximais com relação a isso, ou seja, dado um trecho o conjunto de blocos nele contido é único não existindo nenhum outro trecho com os mesmos blocos.

A avaliação da metodologia aqui proposta foi feita com o uso da ferramenta *Primer-BLAST* [44], uma das principais para se encontrar oligonucleotídeos iniciadores, tendo como entrada os trechos determinados pelo nosso método. Como estudos de caso, três conjuntos de sequências genômicas foram usados. O primeiro conjunto consiste de genomas do gênero *Mycobacterium*; o segundo contém genomas do gênero *Xanthomonas*; e o terceiro contém cepas da espécie *Mycobacterium bovis*. Os resultados obtidos mostraram que método é bastante eficaz para os casos de genomas de um mesmo gênero, enquanto que não obteve bons resultados para cepas de uma mesma espécie.

A principal contribuição deste trabalho consiste, portanto, na metodologia para se encontrar os trechos contendo bons candidatos a oligonucleotídeos iniciadores específicos, usando as abordagens de famílias de proteínas e de sequências características. Resultados preliminares foram publicados em [10].

O texto segue organizado da seguinte forma. O Capítulo 2 contextualiza o problema, trazendo ainda os conceitos básicos necessários para o entendimento do método proposto. O problema das k diferenças é apresentado no Capítulo 3. No Capítulo 4, a metodologia é descrita, enquanto que os resultados obtidos são apresentados no Capítulo 5. Finalmente, no Capítulo 6, considerações finais são feitas.

Capítulo 2

Contextualização

Neste capítulo serão apresentados alguns conceitos básicos e notações utilizadas ao longo desta dissertação. É assumido que o leitor já possui conhecimento básico acerca de Biologia Molecular. Contudo, são introduzidas, resumidamente, informações consideradas necessárias e importantes para a compreensão deste trabalho na Seção 2.1. Como este trabalho é centrado na análise da sequência do DNA de bactérias, a Seção 2.2 traz uma breve descrição da estrutura do genoma bacteriano, enquanto que a Seção 2.3 apresenta a definição de famílias de proteínas. Tais informações podem ser encontradas em [45]. Em 2.4 é apresentado uma breve descrição de oligonucleotídeos iniciadores. Por fim, na Seção 2.5, tem-se descrita a técnica de PCR que se utiliza dos oligonucleotídeos iniciadores e conta com uma vasta aplicabilidade, com por exemplo, testes de identificação genética, medicina forense, entre outras.

2.1 Conceitos Básicos de Biologia Molecular

Importante área da Biologia, tendo como base o estudo da vida em escalas moleculares, a Biologia Molecular compreende, principalmente, a relação entre genética, DNA, RNA e a produção de proteínas. Ela está intimamente ligada com outras áreas de estudo, dentre elas a bioquímica e a própria genética.

Qualquer matéria viva, com exceção dos vírus, é composta por pequenas estruturas denominadas células, onde está contido o objeto de estudo desta área da Biologia. É nessa estrutura complexa que estão presentes as características morfológicas e fisiológicas dos organismos vivos, formadas por um agregado de moléculas, organizadas conforme suas funções e delimitadas por uma membrana celular.

Apesar de possuírem características estruturais comuns, tais como a arquitetura de suas membranas, processos metabólicos, como por exemplo a replicação do DNA, síntese protéica e produção de energia química, os organismos mantêm diferenças em nível celular, podendo ser classificados em dois grandes grupos, procariotos e eucariotos, como pode ser visto na Figura 2.1.

As células de organismos procariotos não contêm núcleo organizado, como por exemplo as bactérias e cianobactérias, ficando o material genético disperso no citoplasma, juntamente com outras partículas. Já nas céculas de organismos eucariotos, como é o caso das células animal e vegetal, existe um núcleo organizado. No núcleo, o material genético encontra-se envolvido por uma membrana nuclear, também chamada de carioteca.

Além de água e alguns outros elementos químicos, as células são constituídas por moléculas pequenas, como por exemplo os nucleotídeos, aminoácidos, açúcares e,

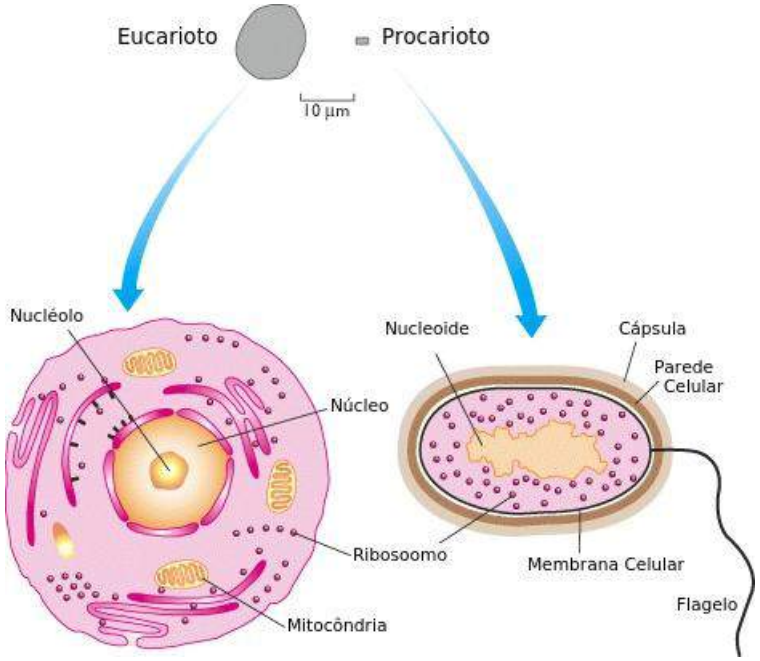


Figura 2.1: Célula de organismo eucarioto (esquerda) e de procarioto (direita). Figura adaptada de [8].

também, pelas macromoléculas denominadas polímeros biológicos que, por sua vez, podem ser de 3 tipos, ácidos nucléicos, proteínas e carboidratos.

Os carboidratos, dentre outras funções, são a principal fonte de energia celular e são constituídos basicamente por açúcares. Já as proteínas desempenham inúmeras funções biológicas, além de determinar a forma e estrutura da célula. Conhecidas também como moléculas que realizam o trabalho celular, as proteínas também são responsáveis pela função catalítica, controle de permeabilidade das membranas e, principalmente, controle da função gênica.

Cada uma das proteínas produzidas para desempenhar alguma função específica é formada por somente vinte aminoácidos diferentes, ou seja, existem somente vinte aminoácidos diferentes que, combinados, podem gerar todas as proteínas, como por exemplo a insulina que é formada por 51 aminoácidos e a hemoglobina que contém 574 aminoácidos.

Tendo em vista a importância das proteínas, é necessário que a célula possua algum mecanismo de controle sobre quais proteínas sintetizar, em que quantidade e em que situações. Tais informações estão armazenadas nos ácidos nucléicos, moléculas que estocam e transmitem a informação genética na célula.

São dois os tipos de ácidos nucléicos existentes, o ácido desoxirribonucléico (DNA) e o ácido ribonucléico (RNA) que, por sua vez, são formados por somente cinco tipos diferentes de nucleotídeos.

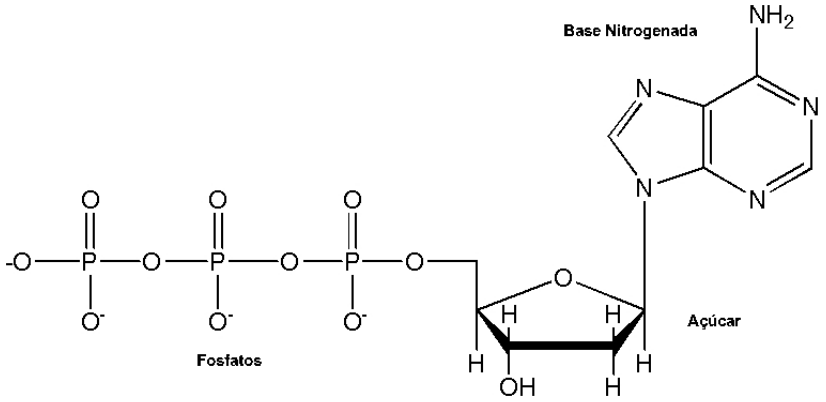


Figura 2.2: Estrutura dos nucleotídeos. Figura adaptada de [45].

Como pode ser observado na Figura 2.2, a estrutura dos nucleotídeos compreende um grupo fosfato, um açúcar (pentose) e uma base nitrogenada unidos por ligações covalentes. Basicamente, a diferença entre os dois ácidos se encontra no tipo de açúcar, como pode ser visto na Figura 2.3, desoxirribose no caso do DNA e ribose no RNA e, na composição de bases da molécula.

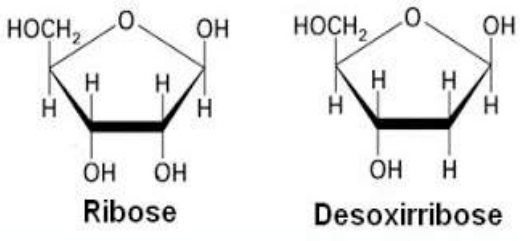


Figura 2.3: Detalhes das moléculas de açúcar que diferenciam RNA e DNA. Figura adaptada de [45].

As bases nitrogenadas adenina (A), guanina (G) e citosina (C) são encontradas em ambos os ácidos, porém, a base timina (T) está presente somente no DNA, enquanto que a base uracila (U) aparece somente no RNA.

Quando os nucleotídeos se ligam a fim de formar o ácido nucléico, estes possuem uma orientação química de extrema importância. Uma das extremidades de uma fita de DNA ou RNA, há um grupamento fosfato ligado ao carbono 5 (carbono 5') do açúcar (extremidade 5') e, na outra extremidade, há uma hidroxila ligada ao carbono 3 (carbono 3') do açúcar (extremidade 3'). Por convenção, a sequência nucleotídica é escrita e lida da esquerda para a direita, no sentido da extremidade 5' para a extremidade 3'.

2.1.1 DNA

O DNA é um aglomerado de nucletídeos que formam uma estrutura de dupla hélice, a partir de duas fitas que se enrolam em torno do eixo da hélice, ambas as fitas na direção 5' $\rightarrow$ 3', porém em sentidos opostos, como pode ser visto na Figura 2.4. As bases nitrogenadas ficam no interior da hélice ligadas por pontes de hidrogênio entre as duas fitas, mantendo desta forma a estrutura da molécula.

Devido às suas características, a base A pode se parear, no caso do DNA, com a base T e a base C com a base G, conferindo assim, a complementaridade da molécula de DNA. Sempre que houver uma base A numa fita haverá uma base T pareada na outra e quando houver uma base G, na outra fita haverá uma base C, sendo essa a propriedade fundamental do DNA, base para os processos celulares.

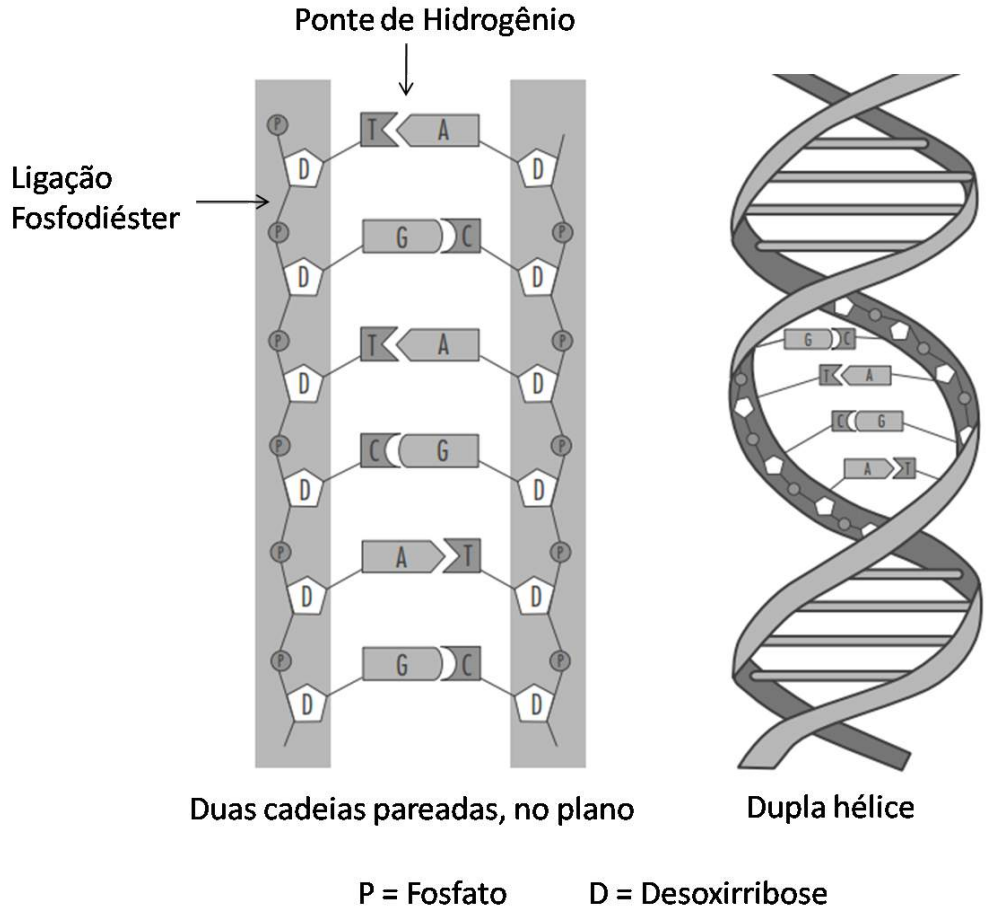


Figura 2.4: Ligações químicas entre os nucleotídeos, as bases nitrogenadas e o modelo em dupla hélice do DNA. Figura adaptada de [39].

2.1.2 RNA

O RNA é formado por uma estrutura de fita simples, como pode ser visto na Figura 2.5, que ocasionalmente pode se dobrar de tal modo que suas próprias bases se pareiam umas com as outras. Além das diferenças já citadas em relação ao DNA, diferentes tipos de RNA podem estar presentes na célula. Os principais são o RNA mensageiro (mRNA) o RNA transportador (tRNA) e o RNA ribossômico (rRNA).

O mRNA é responsável por transferir a informação genética do DNA até os ribossomos, a fim de ocorrer a síntese de uma determinada proteína, já o RNA transportador

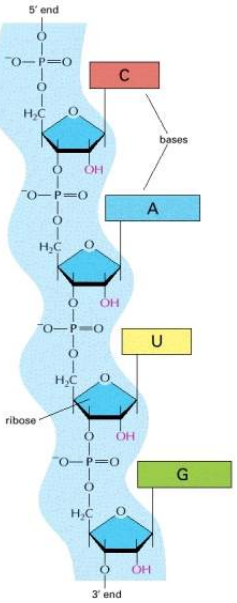


Figura 2.5: Estrutura química do RNA. Figura adaptada de [2].

é responsável por transportar aminoácidos até os ribossomos e o RNA ribossômico atua principalmente na construção dos ribossomos.

2.1.3 Síntese Protéica

Tanto o DNA quanto o RNA desempenham papel fundamental para a manutenção da atividade celular, fazem parte do mecanismo desenvolvido para a produção de proteínas que irão auxiliar em todo e qualquer processo biológico, sendo essenciais para o funcionamento celular.

São necessárias algumas definições para melhor compreensão da subseção. Uma região gênica ou codificadora compreende uma sequência de nucleotídeos que irão definir a ordem dos aminoácidos nas proteínas. Podem existir na região gênica sequências de nucleotídeos que não codificam parte alguma da proteína, essas sequências não codificantes são denominadas íntrons. A região codificadora é delimitada pelas sequências nucleotídicas reguladoras, que contém a sequência promotora, res-

ponsável por indicar o início da região codificadora e a sequência terminadora, cuja responsabilidade é indicar o fim da região codificadora.

Embora seja um procedimento complexo, resumidamente o processo de síntese protéica pode ser descrito seguindo duas etapas, a primeira denominada transcrição e a segunda conhecida como tradução.

A etapa de transcrição compreende a cópia de uma região gênica em uma molécula de mRNA. Uma região gênica compreende uma sequência específica de aminoácidos do DNA que codifica uma determinada proteína. Para que isto ocorra, a dupla fita de DNA rompe suas ligações de hidrogênio e se abre, servindo como um molde, fazendo com que a sequência gênica da região seja transcrita em uma sequência de mRNA através da complementaridade das bases.

A transcrição somente é possível devido ao fato de a enzima polimerase, mais especificamente neste caso a RNA polimerase, ser responsável por ligar os nucleotídeos livres na fita de DNA, tal classe de enzima é capaz de adicionar nucleotídeos na extremidade 3' de uma região pareada do DNA, movendo-se ao longo da cadeia de DNA no sentido 3' $\rightarrow$ 5' possibilitando a extensão da cadeia de mRNA no sentido 5' $\rightarrow$ 3' [45], conforme a Figura 2.6. Concluída a transcrição, a fita de mRNA é liberada e o DNA retorna à sua formação original.

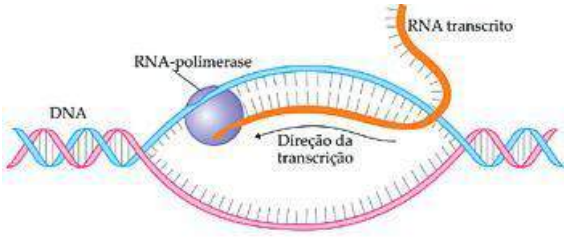


Figura 2.6: Esquema ilustrando a etapa de transcrição de uma sequência molde de mRNA em uma proteína. Figura adaptada de [22].

A próxima etapa, a de tradução, ilustrada na Figura 2.7, consiste na leitura da sequência do mRNA pelo ribossomo a cada três nucleotídeos, denominadas trinca de nucleotídeos, também conhecidas como códons, que especificam qual o aminoácido deverá ser transportado até o ribossomo pelo tRNA, de modo qua a sequência do mRNA se pareie com a sequência do tRNA, conhecida também como anti-códon. Uma vez pareado, o aminoácido ligado ao tRNA se desprende e se liga aos outros aminoácidos. Dessa forma, ao final do processo a proteína é sintetizada.

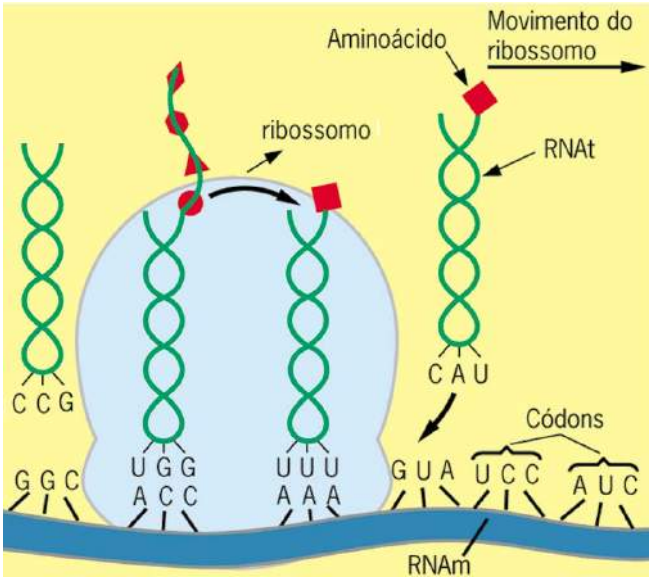


Figura 2.7: Tradução de uma fita de mRNA em proteína. Figura adaptada de [39].

Como cada códon é formado por uma sequência de 3 nucleotídeos que darão origem a um aminoácido e, como temos 4 possíveis nucleotídeos no DNA, logo, teríamos um total de 64 aminoácidos possíveis. Porém, temos somente 20 aminoácidos que, combinados, dão origem às proteínas. Conclui-se então que um mesmo aminoácido pode ser codificado por códons que apresentam sequências de nucleotídeos diferentes, como pode ser visto na Figura 2.8.

1ª Base		2ª BASE U		2ª BASE C		2ª BASE A		2ª BASE G		3ª Base
	U	U U U, U U C	Fenilalanina (Fen)	U C U, U C C, U C A, U C G	Serina (Ser)	U A U, U A C	Tirosina (Tir)	U G U, U G C	Cisteína (Cis)	U
		U U A, U U G	Leucina (Leu)			U A A, U A G	códon de parada	U G A	códon de parada	A
								U G G	Triptofani (Trp)	G
	C	C U U, C U C, C U A, C U G	Leucina (Leu)	C C U, C C C, C C A, C C G	Prolina (Pro)	C A U, C A C	Histidina (His)	C G U, C G C, C G A, C G G	Arginina (Arg)	U, C, A, G
						C A A, C A G	Glutamina (Glu)			
	A	A U U, A U C, A U A	Isoleucina (Ile)	A C U, A C C, A C A, A C G	Treonina (Ter)	A A U, A A C	Asparagina (Asn)	A G U, A G C	Serina (Ser)	U, C
		A U G	Metionina (Met) / códon de início			A A A, A A G	Lisina (Lis)	A G A, A G G	Arginina (Arg)	A, G
	G	G U U, G U C, G U A, G U G	Vitalina (Val)	G C U, G C C, G C A, G C G	Alanina (Ala)	G A U, G A C	Ácido Aspártico (Asp)	G G U, G G C, G G A, G G G	Glicina (Gli)	U, C, A, G
						G A A, G A G	Ácido Glutâmico (Glu)			

Figura 2.8: O conjunto de correspondências entre triplas de nucleotídeos no DNA e os aminoácidos. Figura adaptada de [22].

2.2 Genoma Bacteriano

Apesar de toda célula conter material genético (DNA) responsável por toda a informação necessária à sobrevivência do organismo e, este ser constituído por moléculas de fita dupla, os genomas das bactérias apresentam algumas características que os diferem dos eucariotos. Tais atributos compreendem, por exemplo, seu tamanho, forma, estrutura, distribuição dos genes, entre outros.

Os genes bacterianos são menos complexos, formados pela região reguladora, codificadora e, ao contrário dos eucariotos, genes de procariotos muito raramente apresentam íntrons. Tal característica, de exata equivalência entra a sequência nucleotídica do gene e a sequência de aminoácidos da proteína, é conhecida como colinearidade.

Com relação ao genoma, as células bacterianas contêm, em sua maioria, um único cromossomo, formado, como já dito anteriormente, por uma molécula de fita dupla de DNA. Na quase totalidade dos casos, sua estrutura é disposta de maneira circular fechada, como pode ser visto na Figura 2.9. Há, no entanto, casos em que o genoma possui arquitetura diferenciada. A Tabela 2.1, transcrita de [38], mostra

alguns exemplos de genomas de bactérias com diferentes números de cromossomos e outros replicons. Um replicon é uma molécula de DNA que pode ser um cromossomo ou um plasmídeo, que consiste em moléculas menores, capazes de se replicar independentemente.

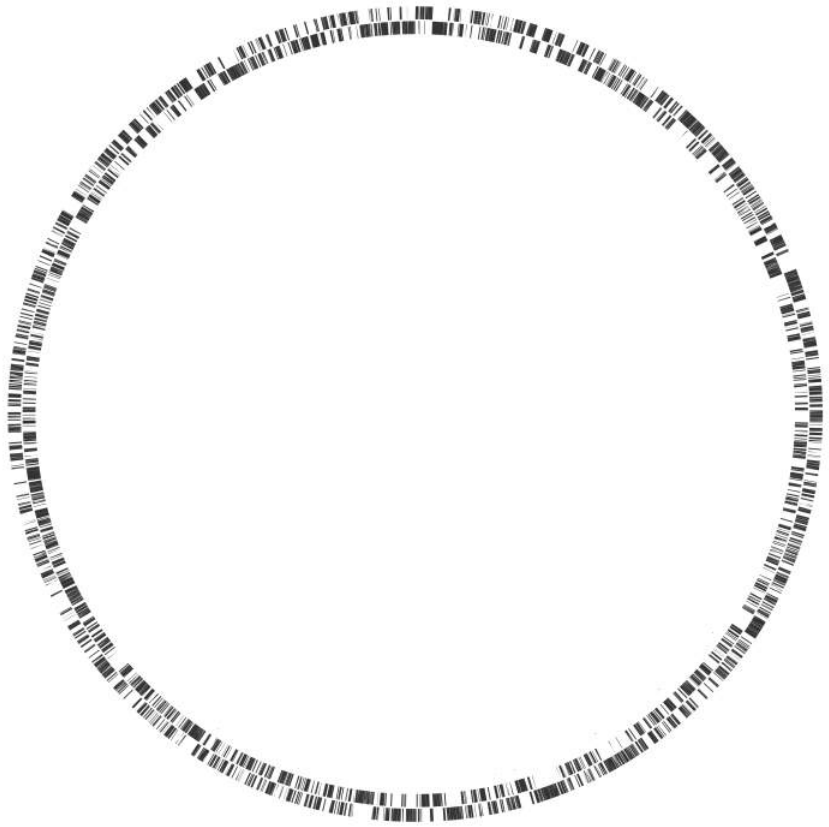


Figura 2.9: Cromossomo circular da bactéria *Agrobacterium tumefaciens* C58. Os círculos representam a posição relativa dos genes codificadores de proteínas em ambas as fitas. Figura construída usando o software GenomeViz [21].

Devido à sua menor complexidade, estes organismos foram amplamente estudados e uma grande quantidade de projetos para sequenciamento de seus genomas foram desenvolvidos. Com isso, o conhecimento acerca das células procarióticas permitiram a identificação e o mapeamento de um número expressivo de genes em diferentes espécias bacterianas, destacando assim padrões na organização desses genomas.

Com relação ao tamanho, em comparação com qualquer genoma eucarioto, os genomas de procariotos são muito pequenos, podendo variar de aproximadamente

Tabela 2.1: Alguns genomas bacterianos e seus tamanhos de replicons e cromossomos. A última coluna mostra o número de genes codificadores de proteínas.

Genome	**Replicon**	**Accession number**	**Tamanho**	**Genes**
Agrobacterium tumefaciens str. C58	Chromosome circular	NC_003062	2,841,580	2,765
	Chromosome linear	NC_003063	2,075,577	1,851
	Plasmid At	NC_003064	542,868	542
	Plasmid Ti	NC_003065	214,233	197
Borrelia burgdorferi ZS7 str. ZS7	Chromosome	NC_011728	906,707	808
	Plasmid ZS7_cp26	NC_011724	26,514	25
	Plasmid ZS7_cp32-1	NC_011731	30,330	36
	Plasmid ZS7_cp32-12	NC_011735	29,806	39
	Plasmid ZS7_cp32-3+10	NC_011720	48,168	60
	Plasmid ZS7_cp32-4	NC_011736	30,964	40
	Plasmid ZS7_cp32-9	NC_011722	30,467	35
	Plasmid ZS7_lp17	NC_011782	17,266	14
	Plasmid ZS7_lp25	NC_011783	24,326	17
	Plasmid ZS7_lp28-1	NC_011780	23,422	13
	Plasmid ZS7_lp28-2	NC_011779	29,758	31
	Plasmid ZS7_lp28-3	NC_011781	28,414	22
	Plasmid ZS7_lp28-4	NC_011785	28,885	21
	Plasmid ZS7_lp36	NC_011778	36,852	23
	Plasmid ZS7_lp54	NC_011784	53,615	55
Escherichia coli str. K-12 substr. MG1655	Chromosome	NC_000913	4,639,675	4,145
Mycoplasma genitalium G37	Chromosome	NC_000908	580,076	475
Streptomyces coelicolor A3(2)	Chromosome	NC_003888	8,667,507	7,768
	Plasmid SCP1	NC_003903	356,023	351
	Plasmid SCP2	NC_003904	31,317	34

$0,6 \times 10^6$ (*Mycoplasma genitalium*) até $9,5 \times 10^6$ (*Myxococcus xanthus*) pares de base. Devido ao seu tamanho reduzido, apresentam uma estrutura compacta e dedicada à codificação de proteínas, onde, praticamente todo o DNA apresenta função codificadora ou reguladora com poucos pares de bases entre estes. Em casos extremos de compactação, pode ocorrer a sobreposição de genes, onde, uma mesma sequência é capaz de codificar duas proteínas diferentes [45].

2.3 Famílias de proteínas

Nos últimos anos, uma grande quantidade de projetos para sequenciamento completo do genoma de organismos tem sido desenvolvidos, como pode ser visto na Figura 2.10. Com isso, temos um aumento na quantidade conhecida de informações codificadas no DNA. Tais informações são essenciais pois, características comuns entre dois organismos tendem à serem codificadas no DNA em porções conservadas, podendo ter sido mantidas à partir de um ancestral comum. Com isso, é possível comparar genes de diferentes genomas de maneira que seja possível inferir sobre a sua funcionalidade, bem como, relacionar como estes organismos evoluíram.

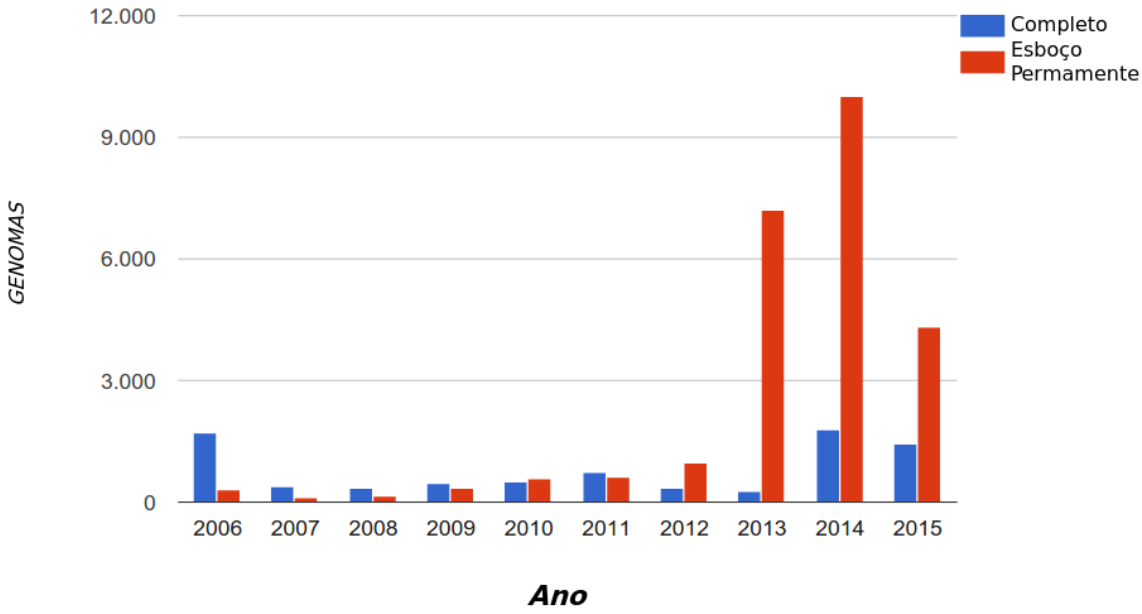


Figura 2.10: Estatística dos Projetos de Sequenciamento de Genoma nos últimos anos. Extraído de `www.genomesonline.org`. Acessado em 31 de Agosto de 2015.

À medida em que mais estudos são realizados, mais genomas são sequenciados, e mais informações são obtidas, tornando possível realizar, de forma mais eficiente, a comparação entre estes genomas, identificando regiões de forma a predizer suas funções e, por consequência, agrupá-las em famílias de acordo com a funcionalidade,

proteína que codifica ou, como é realizado neste trabalho, de acordo com o grau de similaridade através do alinhamento entre estas regiões.

Informações de similaridade entre proteínas a partir de genomas diferentes são muito utilizadas para investigar distâncias filogenéticas, como por exemplo em [19]. Porém, neste trabalho, o agrupamento de proteínas de diferentes genomas em famílias tem o intuito não de encontrar uma determinada família de proteínas presente em diferentes genomas, mas sim de usar os chamados *singletons*, proteínas de um dado organismo alvo que não se agrupam em família. Vamos fazer uso dos *singletons* com intuito de buscar uma região de um determinado organismo com baixa similaridade com relação aos demais.

Para que seja possível encontrar *singletons* de um determinado organismo com relação à outros, primeiramente é necessário encontrar todas as famílias de proteínas. Este processo é realizado através do cálculo da similaridade entre as sequências protéicas dos genomas. Posteriormente é realizada a clusterização das mesmas, ou seja, as proteínas de todos os organismos são comparadas e a partir de uma medida de similaridade são agrupadas de modo que, em cada grupo, estejam proteínas com elevado grau de similaridade entre si.

Ferramentas foram desenvolvidas para agrupar famílias de proteínas baseadas na similaridade. A seguir, será descrita uma delas, a ferramenta OrthoMCL [28]. Além de prover a clusterização de proteínas entre diversos genomas também lista todos os *singletons* destes e esta lista é uma das entradas necessárias para a execução da aplicação desenvolvida neste trabalho.

2.3.1 OrthoMCL

Com a crescente quantidade de informações genômicas sendo produzidas e, considerando que sequências similares tendem a apresentar funções similares, a ferramenta OrthoMCL [28] foi desenvolvida para que fosse possível identificar conjuntos de genes ortólogos, que são genes homólogos de espécies diferentes, separadas por um evento de especiação; e genes parálogos, que são genes de mesma espécie e que são resultantes de um evento de duplicação.

Devido aos custos computacionais de comparação de genomas inteiros, como por exemplo o alinhamento das múltiplas sequências e a interpretação desses alinhamentos, OrthoMCL utiliza uma estratégia baseada no Blast [4] para a comparação das sequências e o algoritmo de Clusterização de Markov (*Markov Cluster* - MCL) [14], que se utiliza de Probabilidade e Teoria dos Grafos, para permitir a classificação em grupos num espaço de similaridades.

A ferramenta pode ser vista como um processo fragmentado em dois passos. Veja a Figura 2.11, onde, a partir das sequências de proteínas para todos os organismos de interesse, num primeiro momento é realizada a comparação todos-contra-todos utilizando a ferramenta Blastp. Relacionamentos de homologia, isto é, ortologia e paralogia, são identificados quando *e-values* de alinhamento entre pares de proteínas são inferiores a 10^{-5} (este valor foi encontrado a partir de estudo experimental). Uma vez obtidos os dados dos alinhamentos, estes são então utilizados para a construção de um grafo onde os nós representam as proteínas e as arestas são ponderadas refletindo a similaridade entre elas.

O próximo passo é utilizar o algoritmo de Clusterização de Markov sobre o grafo de forma a encontrar *clusters*. A idéia central do algoritmo é a de que existe um *cluster* se existem arestas ligando os nós deste *cluster* tal que, dado um nó, ao escolher

aleatoriamente uma aresta, a probabilidade de atingir um outro nó pertencente ao mesmo *cluster* é maior do que a probabilidade de atingir um nó que não pertence ao *cluster*. Seguindo esta metodologia, os *clusters* são encontrados realizando uma sequência de caminhos randômicos no grafo (*random walks*), sendo possível descobrir onde o fluxo se concentra e, consequentemente, identificar os *cluster*.

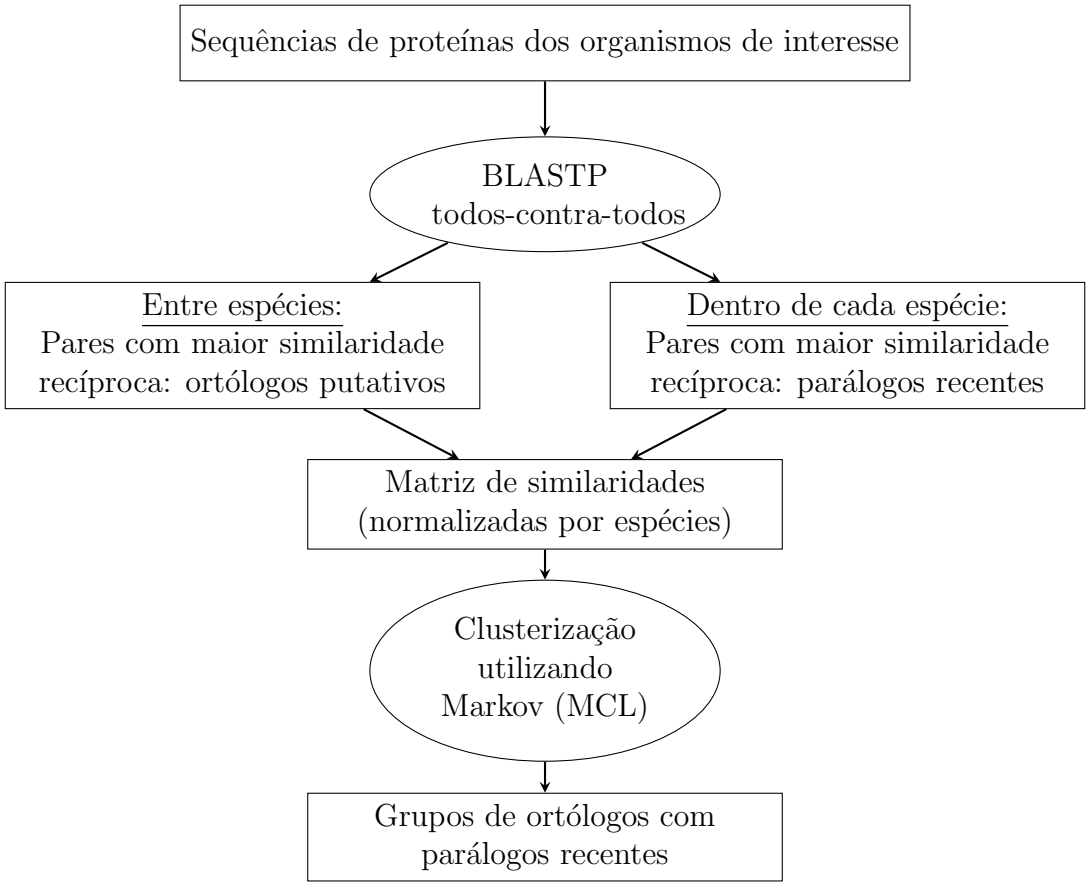


Figura 2.11: Fluxograma de execução da Ferramenta OrthoMCL. Figura adaptada de [28].

2.4 Oligonucleotídeos Iniciadores

Técnicas moleculares que se utilizam da PCR (Reação em Cadeia da Polimerase) para o sequenciamento genético, amplificação de regiões específicas de um genoma, diagnóstico, evolução molecular, dentre outras finalidades, necessitam dos chamados

oligonucleotídeos iniciadores, ou *primers*, para que regiões do DNA alvo sejam amplificadas. Um oligonucleotídeo iniciador nada mais é do que um oligonucleotídeo sintético curto concebido de tal forma que sua sequência é correspondente ao complemento de uma região no DNA alvo. Para uma dada região do DNA alvo que se deseja amplificar, são necessários dois oligonucleotídeos iniciadores, o complementar à região inicial de uma das fitas moldes (*primer forward*) e o complementar à região final da outra fita molde (*primer reverse*) [18], como pode ser visto na Figura 2.13(A).

Diante do exposto, é evidente que o oligonucleotídeo iniciador é de fundamental importância para que resultados satisfatórios sejam alcançados. Por isso, algumas características, como por exemplo o tamanho e especificidade, composição, extremidade 3' e a estrutura interna, entre outras, devem ser levadas em consideração na avaliação de um candidato a oligonucleotídeo iniciador.

Um oligonucleotídeo iniciador que somente se pareia com a região alvo para a qual foi projetado é dito específico e sua especificidade é proporcional ao seu tamanho, pois, quanto maior, mais específico. Logo, oligonucleotídeos iniciadores não podem ter tamanho muito reduzido, possuindo em geral no mínimo 18 bases [1]. Apesar de oligonucleotídeos iniciadores maiores serem mais específicos e de apresentarem uma maior estabilidade na ligação com a região alvo, apresentam também um custo mais elevado de produção e são suscetíveis a formação de estruturas secundárias. Por essas razões, o recomendável é que não ultrapassem 30 bases [1].

Outro fator relevante em um oligonucleotídeo iniciador é a sua composição, já que a quantidade de bases `C` e `G` tem relevante influência na temperatura que deve ocorrer o anelamento entre o oligonucleotídeo iniciador e a região alvo, pois, caso a quantidade de bases `GC` for menor que 50% a temperatura de anelamento será reduzida podendo

o oligonucleotídeo iniciador anelar em alvos similares e amplificar produtos não específicos [1].

Para o pareamento das bases `C` e `G`, são necessárias três pontes de hidrogênio, ao passo que para o pareamento das bases `A` e `T` são necessárias duas pontes de hidrogênio. Também por este motivo, quanto maior a quantidade de bases `C` e `G`, mais fortemente o oligonucleotídeo iniciador estará ligado à região alvo do DNA. Em geral, valores entre 40% e 60% de (`C`+`G`) em relação à quantidade de bases do oligonucleotídeo iniciador são recomendáveis.

Devido às características do pareamento entre `C` e `G`, é aconselhável que uma das duas bases componham a extremidade 3' do oligonucleotídeo iniciador, pois, como este é um pareamento mais estável é esperado que a DNA polimerase inicie o processo de síntese mais eficientemente neste caso [18].

De significativa importância para a realização da PCR, a enzima DNA-polimerase é responsável por ligar os nucleotídeos livres na fita de DNA por complementaridade, tal classe de enzima é capaz de adicionar nucleotídeos na extremidade 3' de uma região pareada do DNA, possibilitando a extensão da cadeia de DNA no sentido 5' $\rightarrow$ 3' [45].

Alguns cuidados também devem ser tomados com relação à estrutura interna do oligonucleotídeo iniciador, como por exemplo evitar que um conjunto de bases se repita (*repeat*) ou mesmo a sequência de repetições de uma única base (*runs*). O problema de *runs* e *repeats* é que um oligonucleotídeo iniciador que apresenta tais características pode se alinhar à uma região diferente da região alvo no DNA (oligonucleotídeo iniciador não específico).

Como os oligonucleotídeos iniciadores utilizados para amplificar uma dada região do DNA alvo trabalham em pares, também deve ser verificado se apresentam trechos complementares entre si, pois, poderá ocorrer um pareamento entre oligonucleotídeos iniciadores (*hetero-dimer*) ao invés do pareamento entre oligonucleotídeo iniciador e região alvo do DNA, também deve ser evitada a possibilidade de que duas cópias de um mesmo oligonucleotídeo iniciador se alinhem (*self-dimer*), outro fenômeno indesejado pode ocorrer em oligonucleotídeos iniciadores longos, onde esses podem se auto-parear, dando origem à estruturas denominadas *hairpins*, como podem ser vistos na Figura 2.12.

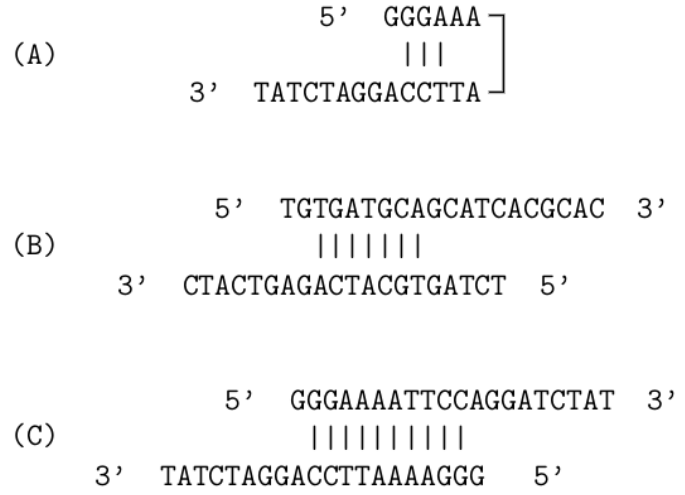


Figura 2.12: Exemplos de formação de estrutura secundária. Hairpin é representado em (A), hetero-dimer em (B) e self-dimer em (C). Extraída de http://bioweb.uwlax.edu/ acessado em 3 de setembro de 2015.

2.4.1 Ferramentas para projeto de oligonucleotídeos iniciadores

Apesar da complexidade de projetar um bom oligonucleotídeo iniciador, devido às características requeridas, esta tarefa pode ser facilmente realizada através de ferra-

manetas disponíveis na web que auxiliam nesta função. Dentre as principais, podem ser citadas a *Web Primer* [1], *Primer3* [42] e *Primer-BLAST* [44].

Com uma interface simplificada, a ferramenta *Web Primer* disponibiliza ao usuário duas formas de entrada de dados, através da sequência do DNA ou apenas informando o identificador no *GenBank*. É possível ainda, escolher entre oligonucleotídeo iniciador para PCR ou sequenciamento.

Munida de uma série de possibilidades, a ferramenta *Primer3* é considerada eficiente para o projeto de oligonucleotídeos iniciadores para PCR, permitindo um maior controle sobre a natureza do oligonucleotídeo iniciador à ser projetado.

Finalmente, a ferramenta *Primer-BLAST* utiliza a ferramenta *Primer3* para o projeto de oligonucleotídeos iniciadores e, uma vez definidos, é realizada uma comparação destes oligonucleotídeos iniciadores via *BLAST* com uma base de dados especificada pelo usuário. A finalidade da comparação é a de garantir que os pares de oligonucleotídeos iniciadores escolhidos não causem amplificação de outras sequências que não seja a sequência alvo.

2.5 Reação em Cadeia da Polimerase (PCR)

Nas últimas décadas tem ocorrido um significativo aumento de estudos que envolvem a Biologia Molecular, seja para o sequenciamento de genoma, para determinar a função de um determinado gene, análise de polimorfismo de DNA, diagnóstico de doenças, evolução molecular, entre outros. Com isso, técnicas têm surgido permitindo um avanço considerável na área.

[1] disponível em: http://www.yeastgenome.org/cgi-bin/web-primer

Devido à sua simplicidade, sensibilidade quanto aos resultados dos experimentos e aplicabilidade, a Reação em Cadeia da Polimerase, ou *Polymerase Chain Reaction* (PCR) tem sido amplamente utilizada, proporcionando enormes possibilidades em várias áreas do conhecimento.

Nesta técnica, onde todo o procedimento é realizado *in vitro*, moléculas de DNA são amplificadas milhares ou até milhões de vezes muito rapidamente, de forma a gerar quantidade suficiente de DNA para que, em seguida, possam ser analisados. Para que o processo possa ser realizado, é necessário o DNA da sequência alvo que se deseja amplificar, enzima DNA polimerase, dois oligonucleotídeos iniciadores, desoxirribonucleotídeos (dNTPs) e, uma correta concentração de $MgCl_2$ [45].

O processo tem início com a mistura dos elementos e, sem seguida, o composto é colocado em um termociclador. Tal equipamento realiza o aquecimento da solução a 95°C, fazendo com que a dupla fita de DNA se rompa, processo conhecido como desnaturação, possibilitando assim o anelamento dos oligonucleotídeos iniciadores com a região alvo no momento em que o termociclador ajusta a temperatura para entre 45°C e 65°C.

Por fim, com uma temperatura de 68-72°C, ocorre a extensão da cadeira de DNA, através da enzima DNA polimerase, a partir dos oligonucleotídeos iniciadores, em cada uma das fitas na direção 5' $\rightarrow$ 3'. Este processo, ilustrado na Figura 2.13, é repetido uma série de vezes (geralmente entre 20 e 30 ciclos) de tal forma que ao final, a região alvo do DNA esteja amplificada.

Como pode ser visto na Figura 2.13(A), a PCR é iniciada com um DNA de cadeia dupla, e cada ciclo da reação inicia-se com um breve tratamento de aquecimento para separar as duas cadeias (etapa 1). Após a separação das cadeias, o arrefecimento do DNA na presença de um grande excesso de oligonucleotídeos iniciadores permite que

estes hibridizem com sequências complementares das duas cadeias de DNA (etapa 2). Em seguida, através da ação da enzima DNA polimerase, as novas cadeias de DNA são sintetizadas, a partir dos dois iniciadores (etapa 3). O ciclo completo é então reiniciado, elevando novamente a temperatura, para separar as cadeias de DNA recém-sintetizadas (Figura 2.13(A)).

À medida em que o procedimento é realizado, os fragmentos sintetizados servem novamente como modelo, e em poucos ciclos o DNA predominante é idêntico ao da sequência delimitada pelos oligonucleotídeos iniciadores. Do DNA colocado na reação inicial, apenas a sequência entre os dois oligonucleotídeos iniciadores é amplificada, porque não existem oligonucleotídeos iniciadores ligados em qualquer outro lugar. No exemplo ilustrado na Figura 2.13(B), três ciclos de reação produzem 16 cadeias de DNA, 8 dos quais (em caixa amarelo) são do mesmo comprimento e correspondem exatamente a uma ou a outra cadeia da sequência original que se desejava amplificar. Depois de mais três ciclos, 240 das cadeias de DNA das 256 correspondem exatamente à região da sequência alvo, e depois de mais vários ciclos, essencialmente todas as novas cadeias de DNA terão o comprimento pretendido inicialmente.

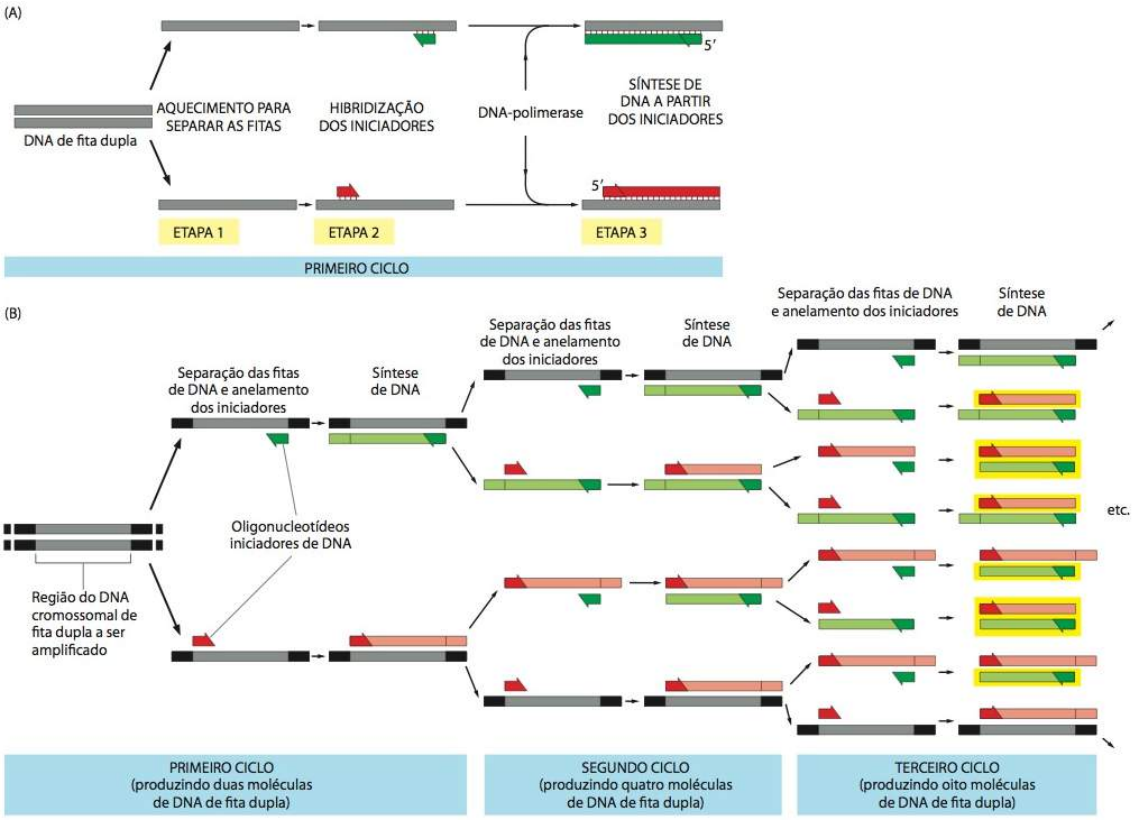


Figura 2.13: Descrição suscinta de uma reação PCR. Figura adaptada de [2].

Capítulo 3

Abordagem baseada no problema das k diferenças

De acordo com o que foi apresentado na Seção 2.4, um oligonucleotídeo iniciador específico deve ser capaz de parear unicamente à região alvo do genoma para a qual foi projetado. Analisando o problema computacionalmente, pode-se representar o genoma como uma sequência de caracteres. O objetivo é então encontrar duas subcadeias (*primer forward* e *primer reverse*) que somente ocorrem numa determinada região da sequência alvo e, quando se trata da identificação de genomas, tais subcadeias, além de somente ocorrerem na sequência alvo, não podem ocorrer em qualquer região nas demais sequências pertencentes ao conjunto de sequências a serem diferenciadas.

Neste capítulo é descrita a abordagem utilizada no trabalho para encontrar regiões com possíveis candidatos a oligonucleotídeos iniciadores, baseada nos trabalhos desenvolvidos em [26, 29]. O capítulo é assim dividido: na Seção 3.1 é descrita a abordagem mais geral de comparação com a permissão de diferenças. Na Seção 3.2

as sequências características são definidas e o algoritmo utilizado neste trabalho é descrito.

3.1 Comparação permitindo diferenças

Uma abordagem para este problema de encontrar marcadores moleculares específicos, em particular oligonucleotídeos iniciadores, quando analisado computacionalmente, requer a solução do Problema da Caracterização de Sequências, que pode ser definido como a busca por uma subcadeia que caracterize uma dada sequência alvo quando comparada com outras.

Formalmente, o Problema da Caracterização de Sequências pode ser definido da seguinte forma: dado um conjunto $\mathcal{S}$ de sequências e uma sequência $T \not\subseteq \mathcal{S}$, encontrar a menor sequência u tal que u é subcadeia de T e não é uma subcadeia de qualquer sequência em $\mathcal{S}$. Com isso, temos uma sequência u que caracteriza a sequência T com relação ao conjunto $\mathcal{S}$.

No caso em que as sequências em $T \cup \mathcal{S}$ apresentam uma elevada similaridade entre si e, como descrito na Seção 2.4, a especificidade de um oligonucleotídeo iniciador depende, principalmente, do seu tamanho e da temperatura necessária para o anelamento com a região alvo, a sequência u, subcadeia que caracteriza T, deve apresentar a menor similaridade possível com relação às sequências em $\mathcal{S}$. Dessa forma, faz-se necessário buscar uma sequência u, que apresenta a maior quantidade de *mismatches*, quando realizado um alinhamento local com as sequências em $\mathcal{S}$, com o intuito de maximizar sua especificidade em relação à T.

Como consequência, o problema passa ser a busca por subcadeias da sequência alvo que possuam um número mínimo de diferenças, daí o termo "k diferenças",

em relação aos genomas do conjunto $\mathcal{S}$. Daqui em diante no texto, o problema é referenciado como *o problema das k diferenças* e o tópico a seguir trata da busca por subcadeias com no mínimo k diferenças.

3.2 Abordagem utilizando sequências características

Devido à especificidade do problema tratado neste trabalho, faz-se necessária uma nova abordagem para o Problema da Caracterização de Sequências. Com o intuito de garantir que determinadas regiões do genoma possam conter bons candidatos a oligonucleotídeos iniciadores, o problema computacional agora se traduz em encontrar uma subcadeia da sequência alvo que, se encontrada nas sequências do conjunto não alvo, o seja com pelo menos k diferenças.

Formalizando o problema, seja $\mathcal{S}$ um conjunto de sequências e T uma sequência tal que $T \notin \mathcal{S}$, é necessário encontrar a menor subcadeia u de T, tal que u não corresponde à qualquer subcadeia das sequências em $\mathcal{S}$ com menos do que k caracteres distindos. Em outras palavras, $\forall w \in \mathcal{S}$, se w' é uma subcadeia de w, então $dist_{lev}(w', u) \geq k$, onde $dist_{lev}(u, v)$ é denominada *Distância de Levenshtein* entre as sequências u e v. Também conhecida como distância de edição, a Distância de Levenshtein é uma métrica que calcula o número mínimo de inserções, remoções e substituições de caracteres necessárias para transformar u em v, onde u,$v \in \Sigma^*$, sendo que Σ^* denota o conjunto finito de todas as sequências sobre o alfabeto Σ (incluindo a sequência vazia ϵ).

Assumindo as sequências u=u_1 ... u_n e v=v_1 ... v_m, a Distância de Levenshtein é definida recursivamente como:

$$dist_{lev}(u_1...u_n, v_1...v_m) = \begin{cases} max\{n,m\} & se\ n=0\ ou\ m=0 \\ min\begin{cases} dist_{lev}(u_2...u_n,v_2...v_m)+1\ se\ [u_1 \neq v_1]), \\ dist_{lev}(u_2...u_n,v_2...v_m)\ se\ [u_1 = v_1]), \\ dist_{lev}(u_1...u_n,v_2...v_m)+1, \\ dist_{lev}(u_2...u_n,v_1...v_m)+1 \end{cases} & se\ n,m>0. \end{cases}$$

onde a primeira linha trata os casos limites de sequências vazias, a segunda representa a substituição, a terceira a casamento de dois símbolos iguais, a quarta linha representa a remoção e a quinta a inserção.

A seguir é apresentado o algoritmo utilizado neste trabalho para encontrar subcadeias de T que não são encontradas com menos do que k diferenças em qualquer subcadeia das sequências do conjunto não alvo S.

Algoritmo para encontrar sequência característica com pelo menos *k* diferenças

De acordo com o que foi descrito em [26], para encontrar uma subcadeia de uma sequência com pelo menos k diferenças com relação à qualquer subcadeia de outra sequência, basta encontrar a maior subcadeia com (k–1) diferenças e acrescentar um símbolo. Com isso, temos a menor subcadeia com k diferenças.

Um exemplo é mostrado na Tabela 3.1. Considere uma sequência p = "TTGAT"e uma sequência q = "GAATAATAGGC", o que se deseja é encontrar uma subcadeia p' de p, que contenha pelo menos 2 diferenças de todas as subcadeias de q. Como

dito no parágrafo anterior, basta encontrar a maior subcadeia com (k–1) diferenças e acrescentar um símbolo a ela.

Tabela 3.1: Tabela correspondente à execução do Algoritmo 1 tendo como entrada a sequência p = "TTGAT", a sequência q = "GAATAATAGGC"e o valor de $k = 2$. Estão em destaque as linha com valores de i iguais a 1, 2 e 3, uma vez que nestas linhas são encontrados valores de $k - 1$.

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p \q		G	A	A	T	A	A	T	A	G	G	C
0		0	0	0	0	0	0	0	0	0	0	0	0
1	T	1	1	1	1	0	1	1	0	1	1	1	1
2	T	2	2	2	2	1	1	2	1	1	2	2	2
3	G	3	2	3	3	2	2	2	2	2	1	2	3
4	A	4	3	2	3	3	2	2	3	2	2	2	3
5	T	5	4	3	3	3	3	3	2	3	3	3	3

A busca por p' se dá da seguinte forma, assuma p'' = "T", tal subcadeia apresenta 1 diferença com relação a algumas subcadeias de q, como por exemplo as subcadeias "G"e "A"conforme ilustrado na Tabela 3.2, porém, sendo p' formada por p'' com o acréscimo do próximo símbolo de p, de tal forma que p' = "TT", verifica-se que p' permanece com 1 diferença com relação a algumas subcadeias de q, como por exemplo "T"e "TA", logo, a subcadeia p'' não é a maior com 1 diferença.

Como a subcadeia p'' = "T"não satisfez a condição de maior subcadeia com 1 diferença, seja p'' = "TT". Como visto anteriormente, a subcadeia "TT"contém 1 diferença com relação a subcadeias em q, assumindo que p'' seja a maior subcadeia de p com 1 diferença, logo p' = "TTG", contudo, como pode ser visualizado na Ta-

Tabela 3.2: Busca por subcadeia com 2 diferenças de p = "TTGAT" com relação a q = "GAATAATAGGC", sendo p'' = "T".

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p \q		G	A	A	T	A	A	T	A	G	G	C
0		0	0	0	0	0	0	0	0	0	0	0	0
1	T	1	1	1	1	0	1	1	0	1	1	1	1
2	T	2	2	2	2	1	1	2	1	1	2	2	2

Tabela 3.3: Busca por subcadeia com 2 diferenças de p = "TTGAT" com relação a q = "GAATAATAGGC", sendo p'' = "TT".

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p \q		G	A	A	T	A	A	T	A	G	G	C
0		0	0	0	0	0	0	0	0	0	0	0	0
1	T	1	1	1	1	0	1	1	0	1	1	1	1
2	T	2	2	2	2	1	1	2	1	1	2	2	2
3	G	3	2	3	3	2	2	2	2	2	1	2	3

Tabela 3.4: Busca por subcadeia com 2 diferenças de p = "TTGAT" com relação a q = "GAATAATAGGC", sendo p'' = "TTG".

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p \q		G	A	A	T	A	A	T	A	G	G	C
0		0	0	0	0	0	0	0	0	0	0	0	0
1	T	1	1	1	1	0	1	1	0	1	1	1	1
2	T	2	2	2	2	1	1	2	1	1	2	2	2
3	G	3	2	3	3	2	2	2	2	2	1	2	3
4	A	4	3	2	3	3	2	2	3	2	2	2	3

bela 3.3, a subcadeia "TAG"de q apresenta 1 diferença com relação a p', não sendo verdade a subcadeia p'' = "TT"como a maior com 1 diferença.

Assumindo agora p'' = "TTG", tal subcadeia contém 1 diferença com relação a subcadeia "TAG" de q, admitindo ser a maior subcadeia com 1 diferença, por conseguinte p' = "TTGA". De acordo com a linha 4 da Tabela 3.4, p' contém 2 ou mais diferenças com relação a qualquer subcadeia da sequência q, foi encontrada então uma subcadeia de p com pelo menos 2 diferenças a partir da maior subcadeia de p com 1 diferença.

Dito isto, o Algoritmo 1 ilustra a adaptação da Distância de Levenshtein que, ao invés de calcular a distância entre duas sequências, calcula a distância entre uma sequência p e as subcadeias de uma sequência q, sendo a linha 7 responsável por essa alteração na característica do algoritmo (normalmente essa linha teria o comando $D(0, j) \leftarrow j$). Como a busca é pela maior subcadeia com $(k-1)$ diferenças, para

otimizar a execução do Algoritmo 1, foi adicionada uma variável (I_{max}) que retorna o índice da maior linha da matriz que contenha o valor $(k-1)$, não sendo necessário percorrer a matriz posteriormente em busca da maior linha com a característica desejada.

Algoritmo 1: Algoritmo k diferenças

```
   Entrada: sequências p e q; inteiro k
   Saída: inteiro I_max
   início
1  |  I_max ← −1
2  |  m ← |p|
3  |  n ← |q|
4  |  para i ← 0, m faça
5  |  |  D[i, 0] ← i
   |  fim para
6  |  para j ← 0, n faça
7  |  |  D[0, j] ← 0
   |  fim para
8  |  para i ← 1, m faça
9  |  |  para j ← 1, n faça
10 |  |  |  se p_i = q_j então
11 |  |  |  |  D[i, j] ← D[i − 1, j − 1]
12 |  |  |  senão
13 |  |  |  |  D[i, j] ← min(D[i, j − 1] + 1, D[i − 1, j] + 1, D[i − 1, j − 1]) + 1
   |  |  |  fim se
   |  |  fim para
14 |  |  se D[i, j] = (k − 1) então
15 |  |  |  I_max ← i
   |  |  fim se
   |  fim para
16 |  retorna I_max
   fim
```

Uma vez que o teste da linha 14 garante que, ao final do processamento, a variável de retorno I_{max} contém o maior índice de p que satisfaz a condição, é possível encontrar, caso exista, a maior subcadeia $p' = p_1 \dots p_{I_{max}}$ de p, que somente é encontrado na sequência q com pelo menos (k–1) diferenças. A Tabela 3.5, mostra um exemplo de execução do Algoritmo 1.

Tabela 3.5: Tabela correspondente à execução do Algoritmo 1 tendo como entrada a sequência p = "MONARCH", a sequência q = "RHETORICIAN"e o valor de k = 2. Está em destaque a linha com $i = 2$, uma vez que esta é a maior linha com valor $k - 1$, portanto, este será o valor de retorno da variável I_{max}.

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p \q		R	H	E	T	O	R	I	C	I	A	N
0		0	0	0	0	0	0	0	0	0	0	0	0
1	M	1	1	1	1	1	1	1	1	1	1	1	1
2	O	2	2	2	2	2	1	2	2	2	2	2	2
3	N	3	3	3	3	3	2	2	3	3	3	3	2
4	A	4	4	4	4	4	3	3	3	4	4	3	3
5	R	5	4	5	5	5	4	3	4	4	5	4	4
6	C	6	5	5	6	6	5	4	4	4	5	5	5
7	H	7	6	5	6	7	6	5	5	5	5	6	6

Algumas informações são relevantes para uma melhor compreensão do algoritmo. O inteiro k e as sequências p e q são informações necessárias de entrada para sua execução, onde k representa a quantidade de diferenças da subcadeia a ser buscada em p com relação a todas as subcadeias em q. O tamanho das sequências p e q, com relação a quantidade de símbolos, é definido por $|p|$ e $|q|$ respectivamente.

Uma estrutura semelhante aquela apresentada no exemplo da Tabela 3.5 é criada para que possam ser calculadas as diferenças entre as subcadeias, tal estrutura é denominada de matriz e sendo representada por D no algoritmo em questão. Como uma matriz é formada por células, estas são caracterizadas através das linhas e colunas da matriz, por isso, para identificar cada posição é necessário especificar o número da linha e da coluna, como exemplos, a última célula da matriz é identificada como $D[|p|, |q|]$ e, a primeira célula da matriz é definida como $D[0, 0]$.

O algoritmo apresentado utiliza um método de resolução de problemas denominada Programação Dinâmica (PD). Tal técnica é aplicável a problemas onde a solução ótima pode ser computada a partir da solução ótima previamente calculada e armazenada de subproblemas do problema original. É verificado o emprego dessa

técnica nas linhas 11 e 13 do Algoritmo 1, onde o cálculo do valor $D[i, j]$ leva em consideração posições anteriormente calculadas e armazenadas.

Com relação à complexidade, observando as linhas 8 e 9, temos que são calculadas $m \times n$ células da matriz, devido ao fato de o bloco do *loop* aninhado sempre executar $m \times n$ vezes, logo, é possível concluir que a complexidade de tempo é $O(mn)$. Para a execução do algoritmo, o espaço necessário para armazenar os custos de edição corresponde ao armazenamento de uma matriz de tamanho $(m+1) \times (n+1)$ (linhas 4 e 6). Portanto, a complexidade de espaço, assim como a de tempo, é de $O(mn)$.

Considerando o Algoritmo 1, a solução para o problema original formalizado na Seção 3.1, onde tem-se uma sequência T, para a qual se deseja encontrar uma subcadeia que a caracterize, com pelo menos k diferenças, com relação à um conjunto de sequências S, é apresentada no Algoritmo 2.

Algoritmo 2: Busca subcadeias de T com pelo menos k diferenças (adaptado de [26])

Entrada: sequência alvo T, de comprimento $|T|$; conjunto não alvo de sequências $\mathcal{S}$ com $|\mathcal{S}|$ sequências; inteiro k; inteiro b
Saída: subcadeias de T com k diferenças em relação à $\mathcal{S}$
início

```
1   para i ← 0 ... |T| − b faça
2     J_max ← −1
3     para j ← 0 ... |S| − 1 faça
4       J_maxaux ← −1
5       J_maxaux ← ALGORITMO1((T[i, i + b]);S_j;k)
6       se (J_maxaux = −1)||(J_maxaux = b) então
7         vá para linha 1
8       se J_maxaux > J_max então
9         J_max ← J_maxaux
        fim para
10    armazenar a subcadeia (T[i, i + J_max + 1])
    fim para
11  retorna subcadeias armazenadas na linha 10
fim
```

O Algoritmo 2 recebe como entrada uma sequência T, a qual se deseja encontrar uma subcadeia característica, um conjunto de sequências $\mathcal{S}$, dito conjunto não alvo, um inteiro k, que indica a quantidade mínima de diferenças que a subcadeia de T deve conter em relação às sequências do conjunto não alvo e, por fim, um inteiro b que indica o tamanho máximo admitido para a subcadeia de T. Como saída, o algoritmo encontra, com tamanho máximo igual a b, todas as subcadeias de T com pelo menos k diferenças com relação à todas as subcadeias das sequências do conjunto não alvo.

Para otimizar o espaço de armazenamento em memória, à cada iteração do laço na linha 1, apenas um prefixo de T, com início na posição i e tamanho b é então selecionado para, dentro da estrutura de repetição da linha 3, ser calculada sua k diferença com relação a todas as subcadeias das sequências presentes no conjunto $\mathcal{S}$, utilizando para isso o Algoritmo 1, chamado na linha 5.

O teste realizado na linha 6 tem a finalidade de, caso o prefixo de T selecionado não apresente a diferença mínima exigida (k símbolos), a execução retorna à linha 1, pois, não faz-se necessário que o mesmo prefixo seja testado com as demais sequências de S, uma vez que é necessário que a diferença mínima seja com relação a todas elas.

Como o que se deseja é a menor subcadeia de T com pelo menos k diferenças com relação à todo o conjunto não alvo, a linha 8 é responsável por testar se o retorno do Algoritmo 1 é maior do que o valor armazenado em J_{max}. Caso seja, é necessário atualizar o valor de J_{max}, pois, a subcadeia de T deve ser maior para que seja possível as k diferenças com relação à sequência atual S_j.

Por fim, a linha 10 é responsável por armazenar a menor subcadeia de T encontrada com pelo menos k diferenças com relação à todo o conjunto não alvo. Importante salientar que o Algoritmo 1 retorna a maior subcadeia com $(k-1)$ diferenças e,

portanto, se esta é a maior com $(k-1)$ diferenças ao inserir o próximo caracter da sequência, esta não terá mais $(k-1)$ diferenças e sim k diferenças, pois, caso não o fosse, a subcadeia inicial não seria a maior com $(k-1)$ diferenças. É por esta razão que a subcadeia a ser armazenada na linha 10 corresponde ao intervalo $T[i, i + J_{maxaux} + 1]$.

O esforço computacional com essa abordagem basicamente é determinado pela complexidade do Algoritmo 1, que é chamado j vezes pelo Algoritmo 2, onde j é igual a quantidade de genomas do conjunto $\mathcal{S}$, conforme linha 3 do Algoritmo 2. Uma vez que a complexidade de tempo do Algoritmo 1 é definido pelos laços das linhas 8 e 9, ao chamar o Algoritmo 1 na linha 5 o Algoritmo 2 passa como parâmetro uma subcadeia de T com tamanho b e a cadeia S_j, assumindo que o maior genoma em $\mathcal{S}$ tem tamanho n, a chamada ao Algoritmo 1 na linha 5 tem complexidade de $O(b \times n)$, como esta linha é chamado j vezes temos uma complexidade de $O(j \times b \times n)$ e, por sua vez, este trecho de código é executado $|T| - b$ vezes de acordo com a linha 1, logo, para o Algoritmo 2 temos uma complexidade de tempo de $O(|T| \times j \times b \times n)$.

Para exemplificar a execução do Algoritmo 2, tome a sequência T={MONARCH} e o conjunto $\mathcal{S}$ = {ARCHITECT, CHEMIST, RHETORICIAN}, k=2 e b=5. A Tabela 3.6 ilustra a primeira rodada de execução do Algoritmo referente ao laço da linha 1. Como já dito anteriormente, para otimizar a execução, note que a sequência p não corresponte à toda sequência T, mas sim, à uma subcadeia de T com tamanho b, reduzindo assim espaço de armazenamento e tempo de processamento, sem afetar na corretude do Algoritmo.

Com relação à linha 3, a Tabela 3.6-A, 3.6-B e 3.6-C exemplificam o cálculo das $(k-1)$ diferenças do prefixo de T com relação à todas as subcadeias das sequências em S à partir do Altoritmo 1. Note que em 3.6-A, o retorno do Altoritmo 1 corresponde

ao valor 1, em 3.6-B ao valor 2 e em 3.6-C também ao valor 1. Logo, ao final da primeira rodada de execução da estrutura de repetição da linha 1, o valor da variável J_{max} será igual a 2 (ver linha 8), resultando na subcadeia "MON".

A segunda iteração do laço na linha 1 do Algoritmo 2 é ilustrada na Tabela 3.7. Assim como visto na Tabela 3.6, é possível identificar que a busca realizada na linha 5 retorna, para cada uma das sequências em $\mathcal{S}$, a maior linha contendo o valor $k-1$. Novamente, ao final da iteração na linha 1, o valor da variável J_{max} será igual a 2, resultando, neste caso, na subcadeia "ONA" a ser armazenada na linha 10.

Por fim, a terceira repetição do laço na linha 1 (Tabela 3.8), exemplifica o caso em que a subcadeia de T contém uma diferença com relação a sequência S_1. Uma vez que a subcadeia não contém duas diferenças com relação a uma das sequências em S, não é necessário realizar o teste com as sequências restantes. Ao final da execução do algoritmo, a linha 11 retorna as subcadeias "MON" e "ONA", dado que a condição de parada na linha 1 é satisfeita.

Tabela 3.6: As Tabelas A, B e C correspondem à execução da primeira iteração da linha 1 do Algoritmo 2 tendo como entrada a sequência T={MONARCH} e o conjunto $\mathcal{S}$ = {ARCHITECT, CHEMIST, RHETORICIAN}, sendo k=2 e b=5.

A

	j	0	1	2	3	4	5	6	7	8	9
i	p\q		A	R	C	H	I	T	E	C	T
0		0	0	0	0	0	0	0	0	0	0
1	M	1	1	1	1	1	1	1	1	1	1
2	O	2	2	2	2	2	2	2	2	2	2
3	N	3	3	3	3	3	3	3	3	3	3
4	A	4	3	4	4	4	4	4	4	4	4
5	R	5	4	3	4	5	5	5	5	5	5

B

	j	0	1	2	3	4	5	6	7
i	p\q		C	H	E	M	I	S	T
0		0	0	0	0	0	0	0	0
1	M	1	1	1	1	0	1	1	1
2	O	2	2	2	2	1	1	2	2
3	N	3	3	3	3	2	2	2	3
4	A	4	4	4	4	3	3	3	3
5	R	5	5	5	5	4	4	4	4

C

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p\q		R	H	E	T	O	R	I	C	I	A	N
0		0	0	0	0	0	0	0	0	0	0	0	0
1	M	1	1	1	1	1	1	1	1	1	1	1	1
2	O	2	2	2	2	2	1	2	2	2	2	2	2
3	N	3	3	3	3	3	2	2	3	3	3	3	2
4	A	4	4	4	4	4	3	3	3	4	4	3	3
5	R	5	4	5	5	5	4	3	4	4	5	4	4

Tabela 3.7: As Tabelas A, B e C correspondem à execução da segunda iteração da linha 1 do Algoritmo 2 tendo como entrada a sequência T={MONARCH} e o conjunto $\mathcal{S}$ = {ARCHITECT, CHEMIST, RHETORICIAN}, sendo k=2 e b=5.

A

	j	0	1	2	3	4	5	6	7	8	9
i	p\q		A	R	C	H	I	T	E	C	T
0		0	0	0	0	0	0	0	0	0	0
1	O	1	1	1	1	1	1	1	1	1	1
2	N	2	2	2	2	2	2	2	2	2	2
3	A	3	2	3	3	3	3	3	3	3	3
4	R	4	3	2	3	4	4	4	4	4	4
5	C	5	4	3	2	3	4	5	5	4	5

B

	j	0	1	2	3	4	5	6	7
i	p\q		C	H	E	M	I	S	T
0		0	0	0	0	0	0	0	0
1	O	1	1	1	1	1	1	1	1
2	N	2	2	2	2	2	2	2	2
3	A	3	3	3	3	3	3	3	3
4	R	4	4	4	4	4	4	4	4
5	C	5	4	5	5	5	5	5	5

C

	j	0	1	2	3	4	5	6	7	8	9	10	11
i	p\q		R	H	E	T	O	R	I	C	I	A	N
0		0	0	0	0	0	0	0	0	0	0	0	0
1	O	1	1	1	1	1	0	1	1	1	1	1	1
2	N	2	2	2	2	2	1	1	2	2	2	2	1
3	A	3	3	3	3	3	2	2	2	3	3	3	2
4	R	4	3	4	4	4	3	3	3	4	4	4	3
5	C	5	4	4	5	5	4	4	4	3	4	5	4

Tabela 3.8: A Tabela corresponde à execução da terceira iteração da linha 1 do Algoritmo 2 tendo como entrada a sequência T={MONARCH} e o conjunto $\mathcal{S}$ = {ARCHITECT, CHEMIST, RHETORICIAN}, sendo k=2 e b=5.

	j	0	1	2	3	4	5	6	7	8	9
i	p \q		A	R	C	H	I	T	E	C	T
0		0	0	0	0	0	0	0	0	0	0
1	N	1	1	1	1	1	1	1	1	1	1
2	A	2	1	2	2	2	2	2	2	2	2
3	R	3	2	1	2	3	3	3	3	3	3
4	C	4	3	2	1	2	3	4	4	3	4
5	H	5	4	3	2	1	2	3	4	4	4

Capítulo 4

Metodologia

Neste capítulo são descritos os passos propostos neste trabalho, que devem ser executados para que, ao final do processo, trechos com possíveis candidatos a oligonucleotídeos iniciadores possam ser identificados no genoma alvo. O Capítulo está estruturado da seguinte forma: a Seção 4.1 trata da abordagem utilizada para otimizar o dentre as regiões encontradas na etapa anterior, a busca por subcadeias com k diferenças, em 4.3 é apresentado um algoritmo para seleção de trechos do genoma baseado na etapa apresentada na Seção 4.2.

4.1 Regiões específicas a partir dos *singletons*

Como visto na Seção 2.4, encontrar um bom oligonucleotídeo iniciador não é uma tarefa trivial, este deve conter os parâmetros citados para que se alinhe à região alvo do genoma para o qual foi projetado e, tratando-se da capacidade de identificar um dado genoma dentre outros, soma-se ainda a particularidade de somente poder alinhar com o genoma alvo, logo, há a necessidade de buscar no genoma alvo um

par de oligonucleotídeos iniciadores com os parâmetros necessários e garantir que este não é encontrado nos demais genomas, pois, caso contrário não se trata de um oligonucleotídeo iniciador específico.

Devido ao custo computacional da busca por oligonucleotídeos iniciadores como descrito no parágrafo anterior, a metodologia proposta visa reduzir tal esforço e realizar esta busca não a partir de todo o genoma alvo, mas sim, em regiões que apresentam características que levam a crer que ali podem ser encontrados oligonucleotídeos iniciadores que o caracterizem.

Para determinar tais regiões, a abordagem utilizada visa realizar uma busca por *singletons* no genoma alvo. Os *singletons* são proteínas isoladas de um determinado organismo que não se agrupam em nenhuma família de proteínas [19], de acordo com sua funcionalidade, como descrito na Seção 2.3.

Uma vez encontrado um *singleton*, a partir da sua posição no genoma, uma **região candidata** de tamanho M, do genoma alvo é um segmento que possui, como ponto médio, exatamente o ponto médio do *singleton* que a originou. A Figura 4.1 mostra dois exemplos de regiões candidatas.

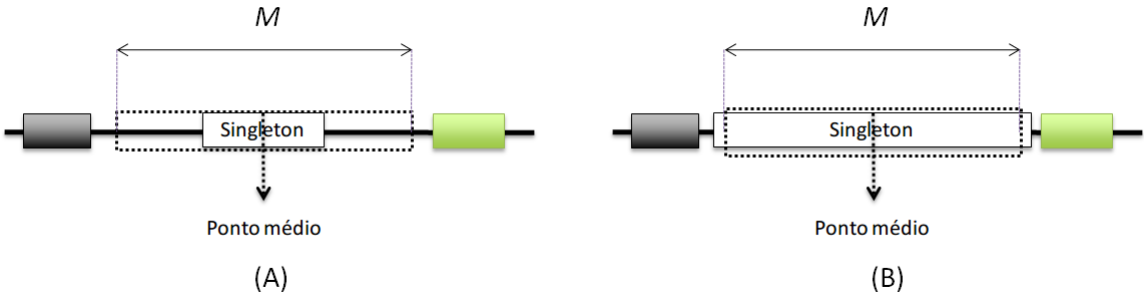


Figura 4.1: Figura com dois exemplos de região candidata a partir de um *singleton*. Os exemplos (A) e (B) ilustram que não importa o tamanho do *singleton*, a região candidata tem tamanho M e mesmo ponto médio do *singleton* que a originou.

O valor de M é definido pelo usuário avaliando a quantidade média de bases que compõem os genes *singletons* do genoma alvo, pois, o ideal seria que o valor de M

correspondesse ao tamanho de cada um dos genes *singletons*, contudo, alguns genes são formados por uma grande quantidade de bases o que implica numa maior região de busca e, por consequência, maior tempo de processamento, logo, o usuário deverá encontrar um valor de M que equilibre o tamanho médio dos *singletons* com o tempo de processamento que será demandado na busca.

Os *singletons* são assumidos como entrada na metodologia proposta. O fluxograma da Figura 4.2 ilustra a sequência de passos necessárias para sua obtenção. A partir de um arquivo para cada genoma, contendo os aminoácidos de cada uma de suas proteínas, a ferramenta descrita na Seção 2.3.1 encontra grupos de famílias de proteínas ortólogas entre os genomas.

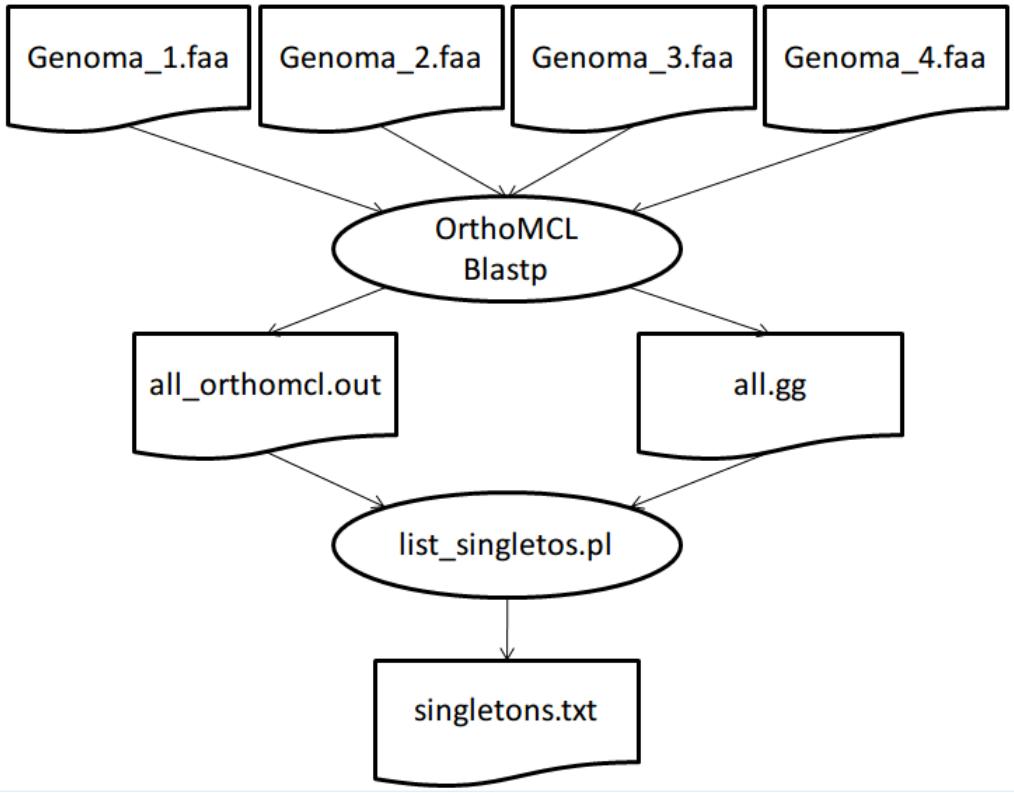


Figura 4.2: Fluxograma que ilustra, a partir dos arquivos contendo os aminoácidos de cada proteína presente em quatro genomas, como são obtidos os *singletons*.

Uma vez que todas as proteínas que apresentam semelhanças quanto a sua funcionalidade são agrupadas, aquelas proteínas que não formam grupos, ou seja, estão isoladas, são os *singletons* que estão sendo buscados.

A ferramenta OrthoMCL não retorna um arquivo com todos os *singletons* de cada genoma. Para isso é necessário utilizar um *script* denominado *list_singletons.pl* que, a partir do arquivo de saída da ferramenta OrthoMCL, mais precisamente o arquivo *all_orthomcl.out*, que contém em cada linha os identificadores das proteínas que pertencem a uma mesma família e, com o arquivo *all.gg*, que contém todos os identificadores para cada uma das proteínas de cada genoma, encontra, caso houver, toda proteína para cada um dos genomas que está em *all.gg* e não está em *all_orthomcl.out*, gerando assim o arquivo *singletons.txt*.

Uma vez encontrado o arquivo *singletons.txt*, para determinar as regiões específicas do genoma alvo, é necessário criar o arquivo *Genoma_alvo.sng* que deve conter apenas os identificadores dos *singletons* para o genoma alvo, como pode ser visto no exemplo ilustrado na Figura 4.3.

Mbovis_AN5_BRA.sng

```
1 Mbovis_AN5_BRA: gi|91770458 gi|91771033
  gi|91771481 gi|91771034 gi|91772903 gi|91773656
  gi|91773165 gi|91774162 gi|91771170 gi|91771495
  gi|91771409 gi|91771590 gi|91772913 gi|91771331
  gi|91772149 gi|91772701 gi|91773408 gi|91771024
  gi|91773174 gi|91772134 gi|91773289 gi|91772120
  gi|91773412 gi|91771494
```

Figura 4.3: Imagem que ilustra um exemplo de arquivo que contém os identificadores das proteínas classificadas como *singletons* para um dado genoma alvo.

Além do arquivo com os *singletons* para o genoma alvo, também é necessário o arquivo *Genoma_alvo.ptt*, uma vez que contém, além de outras informações, o identificador da proteína, suas coordenadas no genoma, além da fita onde foi expressa.

A partir dos arquivos *Genoma_alvo.sng* e *Genoma_alvo.ptt*, é possível encontrar as regiões específicas do genoma alvo e, para isso foi desenvolvido o Algoritmo 3, capaz de, dados uma lista contando o identificador da proteína, sua posição de início e fim, o índice da proteína *singleton* na lista e o tamanho da região pretendida pelo usuário, retornar a posição inicial e final da região candidata formada a partir do *singleton*.

Algoritmo 3: Retorna Região Candidata para um dado *singleton*

Entrada: Lista L com posição inicial (ini) e final (fim) para cada proteína de T em *Genoma_alvo.ptt*; inteiro M correspondente ao tamanho da Região; inteiro i que indica o índice da proteína *singleton* na Lista L; Inteiros *start* e *end*, variáveis globais compartilhadas com outros procedimentos que correspondem ao início e fim de uma região candidata

Saída: Inteiros globais *start* e *end* atualizados

início

```
1  | inteiro: tam, pM;
2  | tam ← |T|
3  |  se (tam < M) então
4  |  |  start ← −1
5  |  |  end ← −1
6  | retorna start, end
7  | pM ← (L[id].ini + L[id].fim)/2
8  |  se (pM <= (M/2)) então
9  |  |  start ← 0
10 |  |  end ← M − 1
11 | senão
12 |  |  se (pM >= |T| − (M/2)) então
13 |  |  |  start ← |T| − (M/2)
14 |  |  |  end ← |T| − 1
15 |  |  senão
16 |  |  |  start ← pM − (M/2)
17 |  |  |  end ← pM − (M/2) − 1
```

fim

O Algoritmo 3 é responsável por encontrar as posições corretas de início e término da região candidata, respeitando os limites do início e fim da sequência do genoma

alvo, linhas 8 e 12, bem como verificando o tamanho da sequência, de acordo com a linha 3.

Portanto, nesta primeira etapa, é realizada uma busca, no genoma alvo, de porções que o diferem dos demais genomas, limitando assim o campo de busca à regiões específicas no genoma alvo que contenham *singletons*.

4.2 Determinação dos blocos com k diferenças

Uma vez que é possível delimitar uma região específica a partir dos *singletons* e do arquivo com as posições de cada gene do genoma como descrito na Seção 4.1, o passo seguinte consiste na busca do que chamamos de *blocos com k diferenças* nestas regiões. Um **bloco com k diferenças**, ou um simplesmente um **bloco**, quando k é implícito no contexto, é definido da seguinta forma: dado um conjunto $\mathcal{S}$ de sequências e uma sequência $T \not\subseteq \mathcal{S}$, um bloco é uma sequência u com tamanho máximo b, tal que u é subcadeia de T, que somente pode ser encontrada em qualquer sequência em $\mathcal{S}$ com pelo menos k diferenças.

Nesta etapa, o Algoritmo 2, apresentado na Seção 3.2 foi modificado com o intuito de realizar a busca somente por blocos com k diferenças nas regiões específicas definidas na Seção 4.1, gerando o Algoritmo 4. Além da sequência alvo T, o conjunto não alvo de sequências $\mathcal{S}$ e as variáveis correspondentes as características dos blocos procurados, como tamanho e quantidade mínima de diferenças, também são necessárias as informações como o arquivo com os *singletons*, como descrito na Seção anterior, o arquivo com as informações das proteínas no genoma alvo e variáveis globais que indicam o início e fim de cada uma das regiões específicas utilizadas na busca por blocos.

Algoritmo 4: Busca blocos em regiões específicas de T com pelo menos k diferenças e tamanho no máximo b

Entrada: sequência alvo T, de comprimento $|T|$; conjunto não alvo de sequências $\mathcal{S}$ com $|\mathcal{S}|$ sequências; inteiro k; inteiro b; inteiro M; *Genoma_alvo.ptt*; *Genoma_alvo.sng*; inteiros globais *start* e *end*
Saída: lista com blocos de T com k diferenças
início

```
1   ListaPTT: Lp
2   ListaBlocos: Lb
3   Inteiro: sing
4   Inteiro: id
5   lerArquivoPTT(Lp, Genoma_alvo.ptt)
6   sing ← lerProximoSing(Genoma_alvo.sng)
7   Enquanto (sing != −1) faça
8       id ← RETORNAINDICE(Lp, sing)
9       ALGORITMO3(Lp, M, id)
10      se start != −1 então
11          para i ← start ... end − b faça
12              J_max ← −1
13              para j ← 0 ... |S| − 1 faça
14                  J_maxaux ← −1
15                  J_maxaux ← ALGORITMO1((T[i, i + b]);S_j;k)
16                  se (J_maxaux = −1)||(J_maxaux = b) então
17                      sing ← lerProximoSing(Genoma_alvo.sng)
18                      vá para linha 7
19                  se J_maxaux > J_max então
20                      J_max ← J_maxaux
                fim para
21              armazenarOrdenado (Lb, T[i, i + J_max + 1], 0)
            fim para
22      sing ← lerProximoSing(Genoma_alvo.sng)
    fim enquanto
23  retorna Lb
fim
```

Foram necessários acrescentar procedimentos que realizam a leitura dos arquivos de entrada, como *lerArquivoPTT* responsável por ler e armazenar em uma lista o identificador, posição inicial e final para cada proteína do genoma alvo e *lerProximoSing* que realiza a leitura de um *singleton* por vez do arquivo com os *singletons* do genoma alvo, como ilustrado na Figura 4.3.

Uma vez que a busca por blocos não é realizada no genoma todo, o Algoritmo 3 é chamado na linha 9 para que, no laço da linha 13 as variáveis que limitam a região específica possam ser utilizadas, por fim, para sua utilização pelo algoritmo da próxima etapa, uma lista ordenada com todos os blocos de acordo com a posição inicial é retornada. A Figura 4.4 ilustra uma região específica com blocos após a execução do Algoritmo 4.

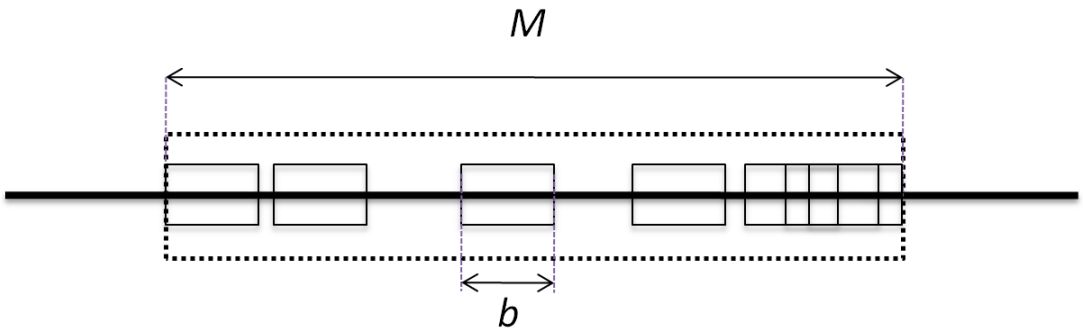


Figura 4.4: Figura que ilustra os blocos, retângulos, de tamanho b que podem ser encontrados dentro de uma dada região candidata de tamanho M.

Assim sendo, nesta etapa, são encontrados blocos com k diferenças, através do Algoritmo 4, a partir de regiões específicas candidatas obtidas com o Algoritmo 3. Com isso, ao invés de realizar uma busca por todo o genoma alvo, somente segmentos que apresentem um potencial de diferenças, em termos funcionais, são tomadas como entrada. No próximo passo da metodologia, finalmente encontraremos o que chamamos de *trecho*. Um trecho é de fato a saída final da nossa metodologia, ou seja, bons trechos representarão segmentos do genoma com bons candidatos a oligonucleotídeos iniciadores, que são os blocos.

4.3 Seleção de trechos candidatos do genoma alvo

Uma vez definida uma região específica e realizada a busca por blocos com k diferenças, trechos do genoma alvo que contenham pelo menos dois blocos são então selecionados como saída da metodologia proposta.

Um **trecho candidato** do genoma alvo, ou simpesmente **trecho**, é definido como um segmento de tamanho c do genoma T que contém no mínimo dois blocos com k diferenças, encontrados pelo Algoritmo 4, apresentado na Seção 4.2, e ainda maximais com relação a esta propriedade. Os trechos candidatos são maximais com relação aos blocos neles contidos, isto é, dados quaisquer dois trechos, a sequência de blocos abrangidos num trecho nunca estará contida em outro. A Figura 4.5 ilustra um trecho maximal com relação ao conjunto de blocos.

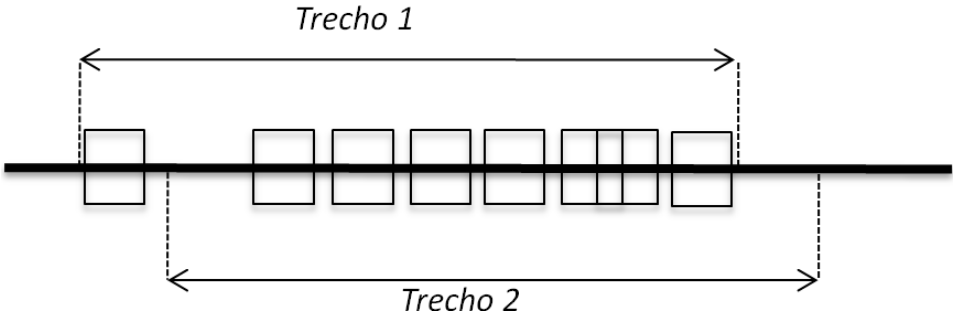


Figura 4.5: Nesta Figura, tem-se o exemplo onde, com relação aos blocos, o trecho 2 está contido no trecho 1, portanto, o trecho 1 é maximal sendo dado como saída da metodologia propostas.

O Algoritmo 5 é responsável por retornar todos os trechos maximais do genoma, por isso, os testes das linhas 14 e 24, além de verificar se dois ou mais blocos foram selecionados, analisam se os blocos que estão sendo selecionados no trecho em questão já foram utilizados anteriormente.

4.4 Etapas da Metodologia

Diante do exposto nas Seções anteriores deste capítulo, para a execução da metodologia faz-se necessário seguir os seguintes passos:

1. Encontrar os *singletons* de acordo com a Figura 4.2;
2. Criar um arquivo com os *singletons* somente do genoma alvo;
3. Utilizar o Algoritmo 4 para encontrar os blocos com k diferenças;
4. Executar o Algoritmo 5 para retornar os trechos candidatos.

Algoritmo 5: Busca trechos maximais com tamanho c de T que contenham blocos com k diferenças

Entrada: sequência alvo T; Lista L de blocos ordenados de T; inteiro c
Saída: Arquivo com trechos de T com tamanho c
início

```
1  | inteiro: teste, vlr, contBlocos, verInt
2  | Lista: *aux,*aux1
3  | teste ← 1
4  | contBlocos ← verInt ← 0
5  | Enquanto ( (L != NULL) && (teste == 1) ) faça
6  | | contBlocos + +
7  | | verInt ← verInt + L → verInt
8  | | L → verInt ← 1
9  | | aux ← L → prox
10 | | se (|T| − L → ini <= c) && (aux != NULL) então
11 | | | verInt ← verInt + aux → verInt
12 | | | aux → verInt ← 1
13 | | | contBlocos + +
14 | | | se (verInt < contBlocos) && contBlocos > 1 então
15 | | | | armazenarFasta (T[|T| − c, |T| − 1], contBlocos)
16 | | | teste ← 0
17 | | senão
18 | | | Enquanto ( (aux != NULL) && (aux → fim − L → ini < c) )
   | | | faça
19 | | | | verInt ← verInt + aux → verInt
20 | | | | contBlocos + +
21 | | | | aux → verInt ← 1
22 | | | | aux1 ← aux
23 | | | | aux ← aux → prox
   | | | fim enquanto
24 | | | se (verInt < contBlocos) && contBlocos > 1 então
25 | | | | vlr ← (L → ini + aux1 → fim)/2
26 | | | | se (vlr > (c/2)) então
27 | | | | | armazenarFasta (T[vlr − (c/2), (vlr + (c/2) − 1], contBlocos)
28 | | | | senão
29 | | | | | armazenarFasta (T[0, c − 1], contBlocos)
30 | | L ← L → prox
31 | | contBlocos ← verInt ← 0
   | fim enquanto
32 | retorna Arquivo com os Trechos
   fim
```

Capítulo 5

Experimentos

Neste capítulo, são descritos os experimentos realizados a partir da metodologia proposta, objetivando avaliá-la. Na Seção 5.1 são apresentados os genomas utilizados nos experimentos. Em seguida, na Seção 5.2, é detalhada a sistemática utilizada durante a fase de testes. Os resultados são apresentados na Seção 5.3. Por fim, uma breve discussão dos resultados obtidos é feita na Seção 5.4.

5.1 Estudos de Caso

Durante os experimentos realizados, três diferentes casos de teste foram utilizados, cada um composto por um conjunto de genomas e dentre eles o genoma alvo. O primeiro conjunto de teste é formado por diferentes espécies do gênero *Mycobacterium*. O segundo é formado por diferentes espécies do gênero *Xanthomonas*. O terceiro e último conjunto de teste é formado por cepas da espécie *Mycobacterium bovis*. A descrição e justificativa para estas escolhas são mostradas nesta Seção.

Espécies do Gênero *Mycobacterium*

O gênero *Mycobacterium* é constituído por dois grupos de espécies, as que provocam tuberculose em humanos e animais, o complexo *Mycobacterium tuberculosis* (CMT), e as denominadas micobactérias não-tuberculosas (MNT). O CMT é formado por *M. tuberculosis*, *M. bovis*, *M. microti*, *M. africanum*, *M. canettii*, *M. caprae*, *M. pinnipedii*, *M. mungi* e *M. orygis* [3, 5, 7, 25, 11], enquanto que mais de 140 espécies formam as (MNT), dentre as quais podemos citar *M. avium* e *M. smegmatis*, sendo a primeira um exemplo de MNT potencialmente patogênica e a segunda um exemplo de MNT raramente patogênica [41].

Existem relatos de infecção pelo *Mycobacterium sp.* descrito em textos históricos desde a antiguidade. Evidências de *M. tuberculosis* e infecção humana pelo *M. bovis* (agente da tuberculose bovina) foram descritos em esqueletos do Sul da Sibéria, com datas que vão de cerca de 1.761 a 2.199 anos antes atrás [24]. A transmissão zoonótica de *M. bovis* tem sido associada ao consumo de leite não pasteurizado ou cozido adequadamente e com a ingestão de alimentos contaminados.

Já as MNT encontram-se dispersas na natureza e, ao contrário das espécies do CMT, sua patogenicidade é variável. A capacidade das MNT em produzir doença está documentada na literatura e sua incidência vem aumentando progressivamente, não só pelo fato do homem estar compartilhando o mesmo habitat, mas também pela melhora nos métodos de diagnóstico e identificação destes microrganismos [17].

As micobactérias *M. avium* e *M. intracellulare*, de difícil diferenciação entre si, formam o Complexo *Mycobacterium avium* (CMA). Atípica, as micobactérias do CMA são as mais relacionadas com doença humana, mais especificamente, são um patógeno pulmonar que afeta indivíduos imunocomprometidos (como HIV/AIDS, paciente em quimioterapia, entre outros). As micobactérias do CMA são a causa

mais comum de infecção por micobactéria não tuberculosa em pacientes com AIDS. *M. avium* é isolado em mais de 95% dos pacientes com AIDS que desenvolvem infecção por CMA. O acometimento pulmonar por esse complexo é raro em imunocompetentes [36].

Já o *M. smegmatis* é uma micobactéria de crescimento rápido e considerado não patogênico. Encontra-se associado a lesões dos tecidos moles após trauma ou cirurgia, tendo sido também descrito como possível fator na carcinogênese peniana. Como apresenta crescimento rápido, podendo multiplicar-se em gerações a cada 1-3 horas, fácil cultivo nos meios de cultura em laboratório, não patogênico e compartilha a mesma estrutura da parede celular de *M. tuberculosis* e das restantes micobactérias, *M. smegmatis* tem sido vastamente utilizado em trabalhos de investigação como microrganismo modelo para a manipulação laboratorial de espécies micobacterianas [35].

Devido aos problemas apresentados, tendo como principais fatores a dificuldade tanto no diagnóstico rápido quanto a diferenciação entre isolados geneticamente próximos, foram escolhidos para o teste da abordagem proposta os seguintes genomas [1]:

- *Mycobacterium tuberculosis* H37Rv, genoma alvo;
- *Mycobacterium smegmatis* str. MC2 155;
- *Mycobacterium canettii* CIPT 140010059;
- *Mycobacterium bovis* AF2122/97; e
- *Mycobacterium avium* 104.

[1]Todos os genomas podem ser obtidos no repositório do NCBI através do endereço: www.ncbi.nlm.nih.gov/genome/

Espécies do Gênero *Xanthomonas*

O gênero *Xanthomonas* inclui inúmeros patógenos de plantas [37]. Dentre as quais destacam-se diversas cepas patógenas de citrus [32]. Em 2002, foi publicado o genoma de *Xanthomonas citri* subsp. *citri* 306 [12], o agente causador do cancro cítrico, ou cancro A, em diversas espécies de citros. Cancro é um problema sério em plantações de laranja no Brasil e em outros locais do mundo [9].

Outros trabalhos em genômica se seguiram, incluindo a publicação dos genomas de duas cepas que causam fenótipos atenuados (cancros B e C) em diferentes espécies de citrus [31]. Mais recentemente, Jalan e colegas [27] publicaram o genoma da cepa Aw12879, que causa cancro somente em alguns tipos de limas, juntamente com uma análise de expressão gênica em diferentes meios de cultura. Esses trabalhos têm trazido uma compreensão cada vez maior das interações entre fitopatógenos e seus hospedeiros, em particular quanto ao repertório de proteínas efetoras que cada espécie ou cepa tem e que desempenha um papel na interação, em muitos casos restringindo os hospedeiros compatíveis [15].

O sequenciamento e a análise de um grupo substancialmente maior de espécies de *Xanthomonas* patogênicas em citros usando-se tecnologias de nova geração (NGS) vai permitir a obtenção de um panorama genômico mais detalhado da diversidade de patógenos de citros com muito maior amplitude do que se conhece até agora. Além da comparação genômica de *Xanthomonas patogênicas* a citros entre si, a comparação das mesmas com outras fitobactérias pode ser importante na deteção de segmentos exclusivos e de polimorfismos em regiões de interesse.

Assim, a comparação desses genomas entre si e com outros genomas existentes (*citri* AW, 306, por exemplo) permitirá melhor compreensão da patogenicidade e taxono-

mia desses patógenos, e ajudará a estabelecer hipóteses quanto aos fatores determinantes das suas gamas de hospedeiro, virulência e sintomatologia.

Neste trabalho, usamos como conjunto de teste os genomas das seguintes espécies do gênero *Xanthomonas*.

- *Xanthomonas axonopodis* pv. *citri* str. 306, genoma alvo;
- *Xanthomonas axonopodis* pv. *citrumelo* str. F1;
- *Xanthomonas axonopodis* Xac29-1;
- *Xanthomonas citri* subsp. *citri* Aw12879; e
- *Xanthomonas fuscans* subsp. *fuscans*.

Cepas de *Mycobacterium bovis*

A tuberculose bovina é uma importante enfermidade infecto-contagiosa, responsável por consideráveis perdas econômicas. Estima-se que os prejuízos anuais com a tuberculose bovina sejam da ordem de U$ 3 bilhões em todo mundo [20]. Além deste aspecto, a tuberculose bovina é um problema de saúde pública [13]. Embora a maioria dos casos humanos de tuberculose sejam causados por *Mycobacterium tuberculosis*, preocupações relacionadas a *M. bovis*, que causa a tuberculose bovina, têm sido expressas devido a ocorrência de casos em humanos [16].

Por ser uma doença de evolução crônica, na maioria das vezes os animais não demonstram alterações perceptíveis ao produtor, sendo o diagnóstico feito por meio de técnicas específicas, sob a responsabilidade de um médico veterinário.

No Brasil, métodos utilizados para o diagnóstico da tuberculose bovina podem levar de 72 horas até meses para a identificação bioquímica dos isolados [30]. Visando o desenvolvimento de métodos mais rápidos de diagnósticos, pesquisadores do Brasil e Argentina têm trabalhado no sequenciamento e anotação de genomas de alguns isolados de *M. bovis*, considerados de importância epidemiológica nos dois países.

Baseado na análise genômica e na identificação de SNPs e famílias de ortólogos comuns a todos os genomas investigados, a ideia é desenvolver um método de genotipagem de isolados de *M. bovis* da América do Sul que permita a discriminação dos isolados, além de uma metodologia automatizada por PCR em tempo real de alto rendimento [6].

O principal benefício esperado com o sequenciamento e a análise comparativa de diferentes cepas de *M. bovis* é o aumento da precisão do diagnóstico *post-mortem* da tuberculose bovina, em lesões encontradas em abatedouro, por meio de teste rápido com alta especificidade e sensibilidade. A partir da determinação das diferenças encontradas entre as cepas de *M. bovis*, por exemplo, a PCR em tempo real poderá fornecer resultados em 24 horas após a chegada do material ao laboratório, contrastando com o cultivo e caracterização bioquímica, que pode demandar até mais de três meses para o diagnóstico definitivo [30].

Para este trabalho, como é necessário que todo o genoma esteja sequenciado e as proteínas anotadas, foram escolhidas, aleatoriamente, três cepas de *Mycobacterium bovis* que possuem as características desejadas, são elas:

- *Mycobacterium bovis* AF2122/97, genoma alvo;
- *Mycobacterium bovis* AN5; e
- *Mycobacterium bovis* 04-303.

5.2 Metodologia de Avaliação

A metodologia de avaliação da abordagem proposta, utilizando os casos de teste apresentados em 5.1, foi realizada em duas fases distintas:

1. determinação dos oligonucleotídeos iniciadores nos trechos sugeridos pelo nosso método, usando um software especializado em projeto de oligonucleotídeos iniciadores; e
2. teste da especificidade dos oligonucleotídeos iniciadores encontrados.

Dessa forma, quanto mais específicos são os oligonucleotídeos iniciadores encontrados, melhor é a avaliação do método.

Determinação dos *oligonucleotídeos iniciadores* em trechos sugeridos pelo método proposto

O processo para determinação dos oligonucleotídeos iniciadores segue os passos descritos no Capítulo 4. Uma vez reunidos todos os genomas, inicialmente é realizada a restrição do campo de busca num dado genoma alvo através da geração dos *singletons* a partir da utilização da ferramenta orthoMCL (Seção 4.1).

O passo seguinte consiste em encontrar subcadeias com o mínimo de k diferenças do genoma alvo com relação aos demais. Tais buscas se concentram nas regiões formadas pelos genes *singletons*, encontrados no passo anterior, delimitadas em exatamente 800 bases, valor definido para que o tempo de processamento não excedesse um limite aceitável de 48 horas. Como não foi encontrado na literatura um número mínimo de diferenças que deve possuir um oligonucleotídeo iniciador para que não se

alinhe à uma região para o qual não foi projetado, nos experimentos deste trabalho o valor para k variou de 5 a 1, nesta ordem.

Por fim, uma vez encontradas as subcadeias com k diferenças, trechos de tamanho 500 [40] do genoma foram selecionados utilizando o Algoritmo 5. Tais trechos são selecionados desde que contenham pelo menos duas subcadeias e sejam maximais, ou seja, dado quaisquer dois trechos, com relação às subcadeias, um trecho não está contido em outro.

A saída do método proposto são os trechos formados a partir da busca por subcadeias com k diferenças e, para definir os oligonucleotídeos iniciadores a partir dos trechos, optou-se pela ferramenta *Primer-BLAST*[2], disponibilizada pelo Centro Nacional de Informação Biotecnológica dos Estados Unidos (*National Center for Biotechnology Information* - NCBI), mantido pela Biblioteca Nacional de Medicina dos Estados Unidos (*U.S. National Library of Medicine* - NLM), considerada uma das mais importantes fontes de informações e repositório de dados biológicos no mundo, além das diversas ferramentas disponibilizadas.

Para a determinação de oligonucleotídeos iniciadores, a utilização da ferramenta se dá a partir da seleção do trecho do genoma alvo, este então deverá ser inserido no quadro "*Enter accession, ig, or FASTA sequence*" da ferramenta, conforme exemplo da Figura 5.1.

Nos testes realizados, todos os parâmetros para a busca de oligonucleotídeos iniciadores não foram alterados, com exceção para a quantidade de oligonucleotídeos iniciadores que a ferramenta retorna que, como pode ser verificado na Figura 5.2, no quadro "*Primer Parameters*", o item "*# of primers to return*" foi estabelecido em 5.

[2]Ferramenta on-line disponível em: http://www.ncbi.nlm.nih.gov/tools/primer-blast/

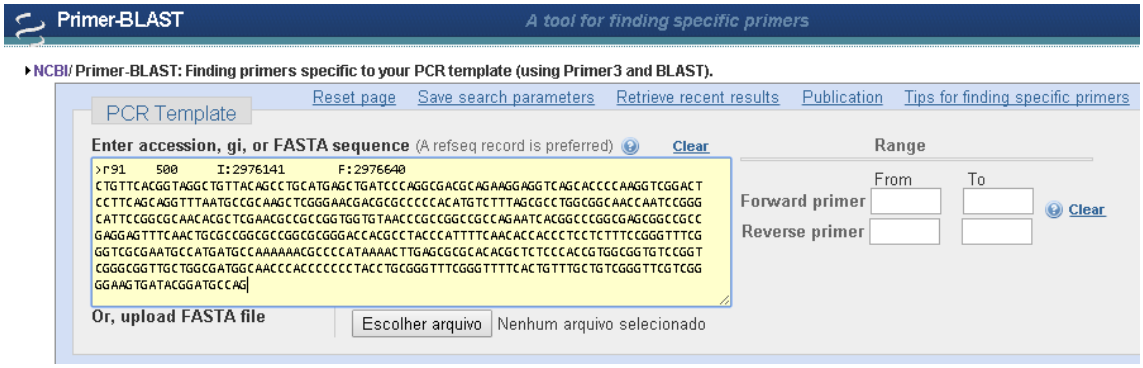


Figura 5.1: Imagem parcial da página inicial da ferramenta *Primer-BLAST* exemplificando a inserção de uma sequencia de bases.

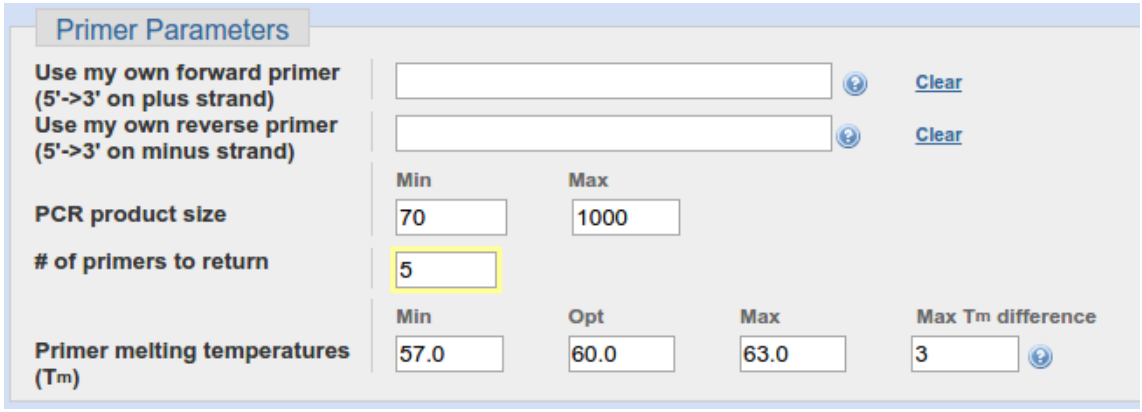


Figura 5.2: Detalhe dos parâmetros utilizados para a busca de oligonucleotídeos iniciadores pela ferramenta *Primer-BLAST*.

Como a finalidade é a busca por oligonucleotídeos iniciadores específicos para uma determinada sequência, no quadro "*Primer Pair Specificity Checking Parameters*" é selecionada a base de dados que contém todos os genomas para que os possíveis oligonucleotídeos iniciadores sejam comparados com a base de dados e sejam escolhidos aqueles mais específicos de acordo com o genoma alvo que é indicado em "*Organism*". Ver Figura 5.3. Feito isso, inicia-se o processo clicando em "*Get Primers*".

O resultado é então disponibilizado com um gráfico que apresenta as posições para cada um dos cinco melhores pares de oligonucleotídeos iniciadores encontrados, como pode ser visto na Figura 5.4, além de uma lista com informações das características

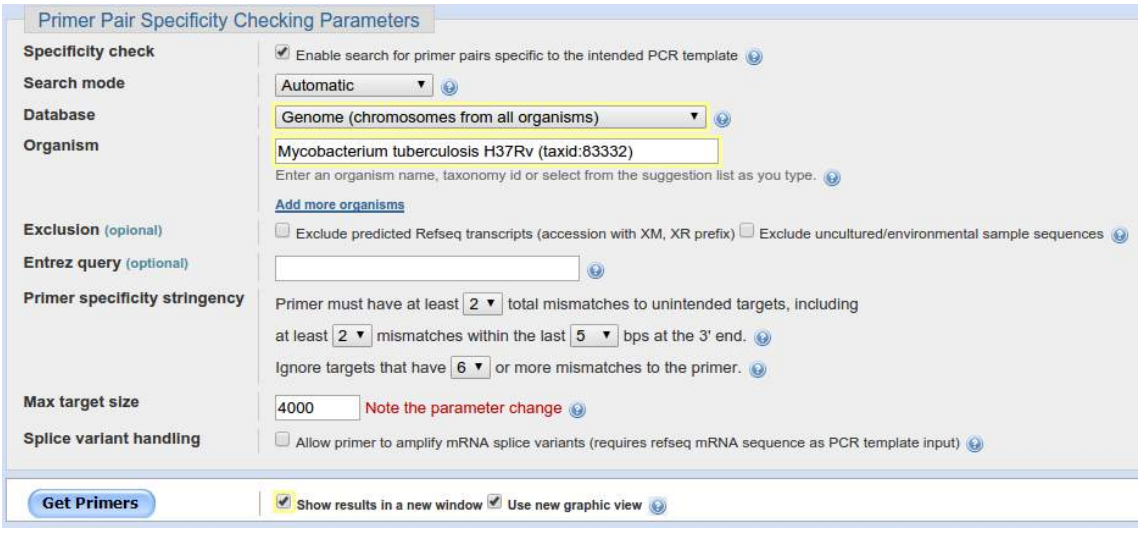


Figura 5.3: Exemplo de configuração de parâmetros com o intuito de aumentar a especificidade dos oligonucleotídeos iniciadores resultantes.

de cada um dos oligonucleotídeos iniciadores, como por exemplo a porcentagem de bases GC e a posição de alinhamento com o genoma alvo.

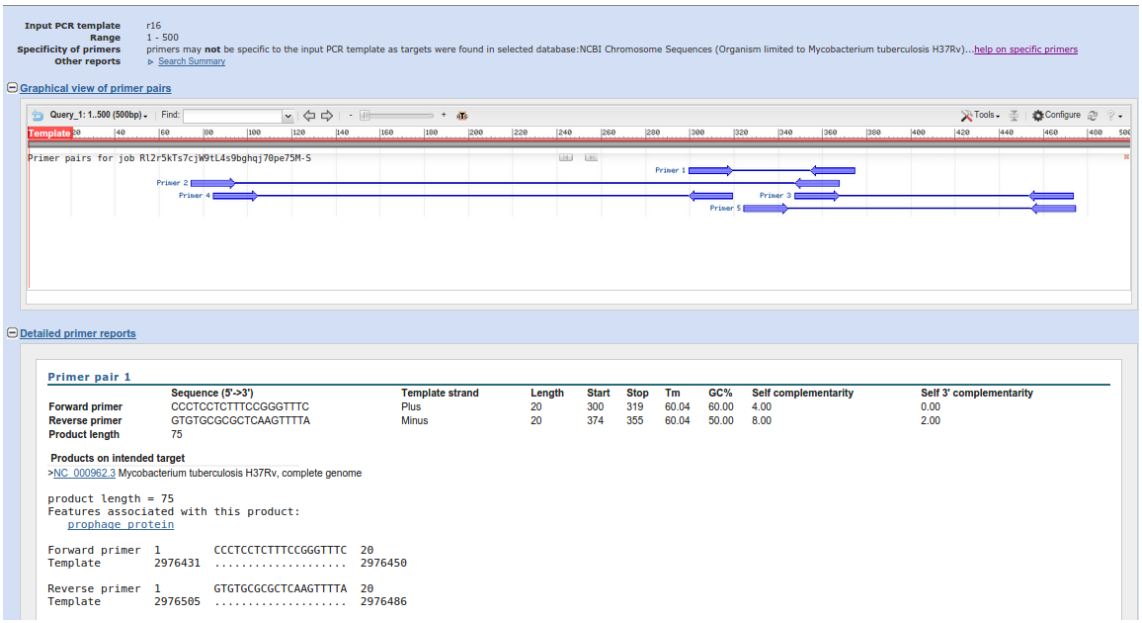


Figura 5.4: Exemplo de saída após execução da ferramenta *Primer-BLAST*.

Teste de especificidade dos *oligonucleotídeos iniciadores*

Para precisar a especificidade dos oligonucleotídeos iniciadores, a ferramenta *Primer-BLAST* é novamente utilizada. Dados os oligonucleotídeos iniciadores gerados como resultado a partir do trecho selecionado do genoma alvo, conforme apresentado na Figura 5.4, cada um destes é então submetido ao teste de checagem de especificidade.

Primeiramente, no quadro "*Primer Parameters*", é inserido o *primer forward* em "*Use my own forward primer (5' → 3' on plus strand)*" e, o *primer reverse*, em "*Use my own reverse primer (5' → 3' on minus strand)*", conforme ilustrado na Figura 5.5.

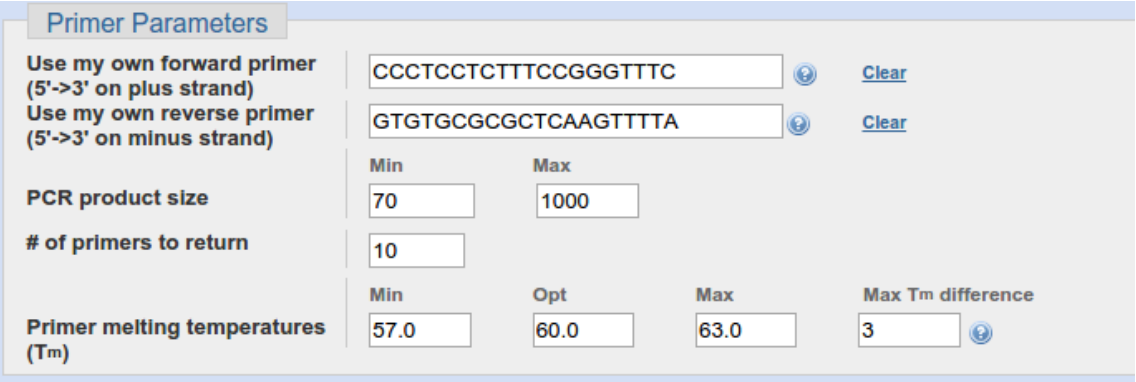


Figura 5.5: Especificação dos oligonucleotídeos iniciadores para determinação de especificidade.

Como agora o objetivo é checar se o oligonucleotídeo iniciador encontrado é específico para o genoma alvo, no quadro "*Primer Pair Specificity Checking Parameters*", são adicionados todos os genomas que fazem parte do conjunto não alvo, como exemplificado na Figura 5.6.

Como resultado da checagem, tem-se três diferentes situações possíveis:

- **Par de oligonucleotídeo iniciador específico**: não foi encontrado alinhamento entre o par de oligonucleotídeo iniciador e os genomas não alvo escolhi-

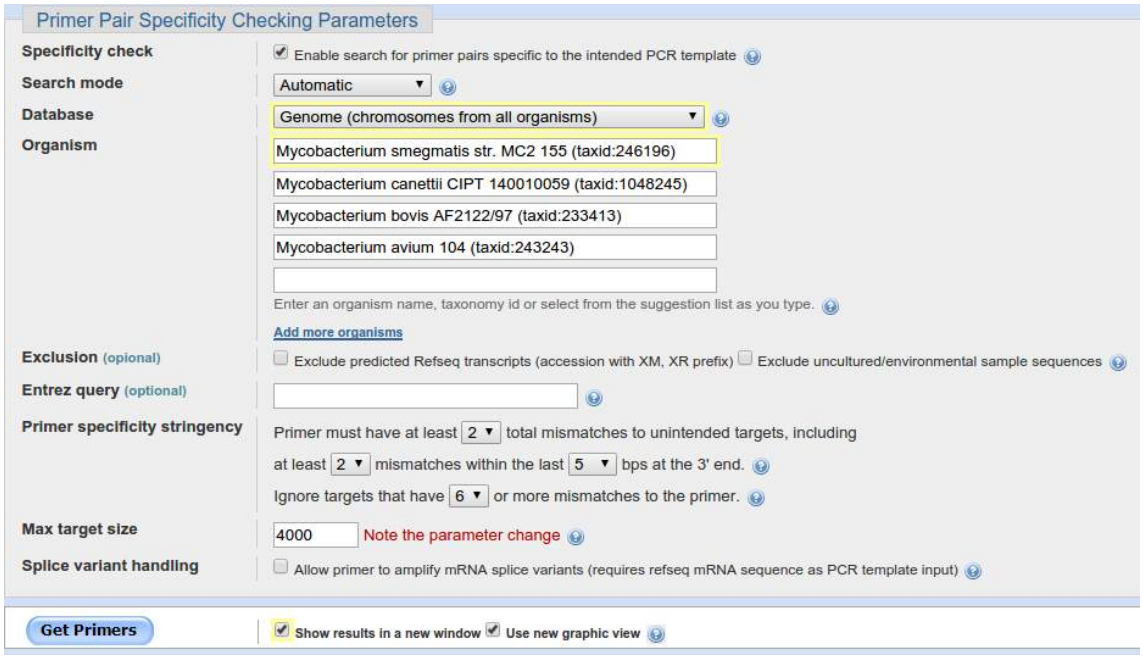


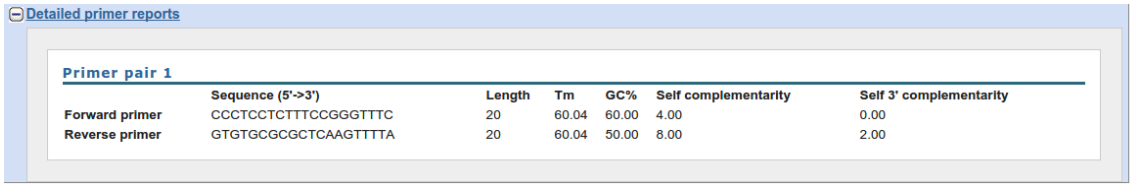
Figura 5.6: Especificação dos genomas do conjunto não alvo para determinação de especificidade do oligonucleotídeo iniciador.

dos e, portando, o oligonucleotídeo iniciador é considerado específico para o genoma alvo. Esta é a melhor situação, indicando que os trechos encontrados por nossa metodologia contêm bons candidatos a oligonucleotídeos iniciadores.

- **Par de oligonucleotídeo iniciador não específico com mismatches**: o oligonucleotídeo iniciador alinhou com algum dos genomas do conjunto não alvo. Porém, de acordo com as especificações padrão da ferramenta *Primer-BLAST*, o alinhamento apresenta entre 2 e 5 mismatches.

- **Par de oligonucleotídeo iniciador não específico sem mismatches**: o oligonucleotídeo iniciador alinhou com algum dos genomas do conjunto não alvo e não ocorreram mismatches. Este é o pior caso, indicando que nossa metodologia não foi capaz de identificar um bom trecho candidato a conter bons oligonucleotídeos iniciadores.

As imagens que ilustram as três diferentes saídas possíveis, como descritas anteriormente, podem ser vistas em 5.7, 5.8 e 5.9, respectivamente.

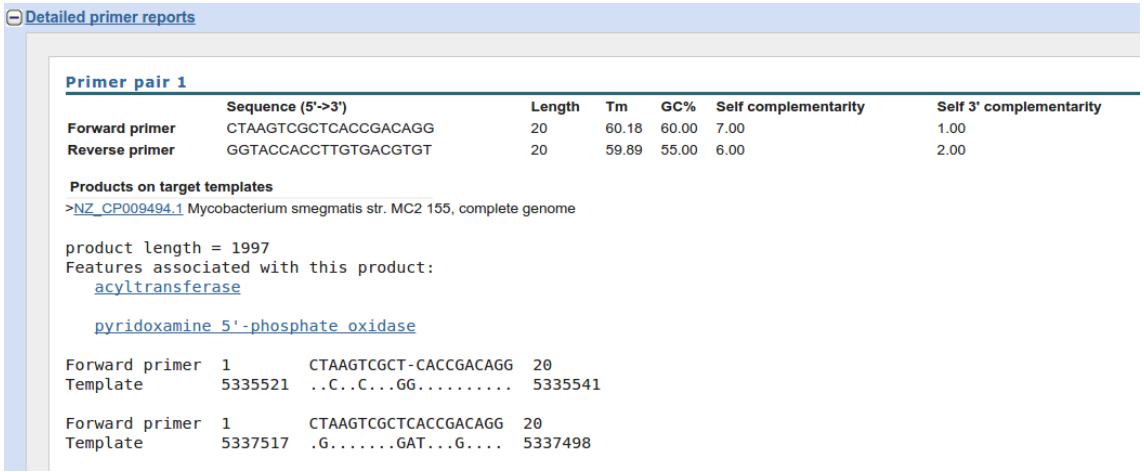


Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CCCTCCTCTTTCCGGGTTTC	20	60.04	60.00	4.00	0.00
Reverse primer	GTGTGCGCGCTCAAGTTTTA	20	60.04	50.00	8.00	2.00

Figura 5.7: Exemplo de checagem onde, de acordo com o padrão da ferramenta *Primer-BLAST*, é constatado que o oligonucleotídeo iniciador é específico para o genoma alvo.



Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTAAGTCGCTCACCGACAGG	20	60.18	60.00	7.00	1.00
Reverse primer	GGTACCACCTTGTGACGTGT	20	59.89	55.00	6.00	2.00

Products on target templates

>NZ_CP009494.1 Mycobacterium smegmatis str. MC2 155, complete genome

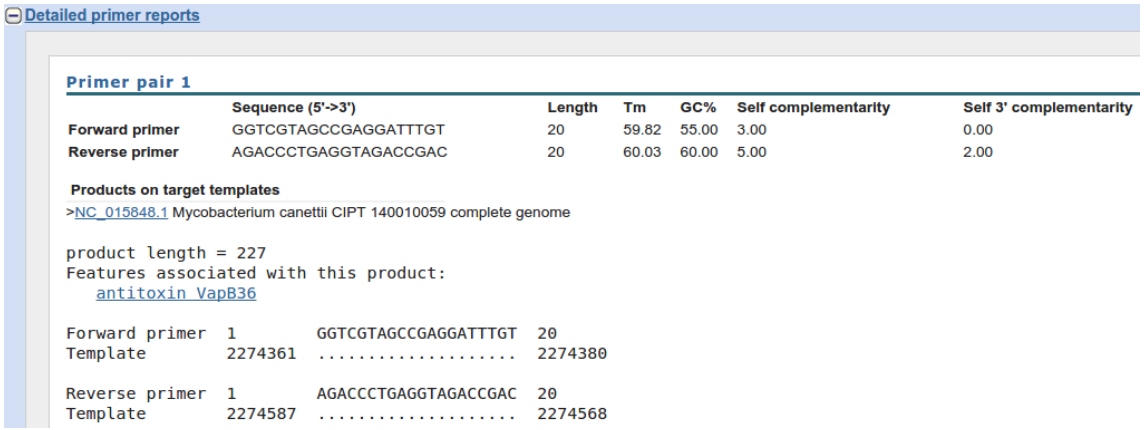
product length = 1997
Features associated with this product:
acyltransferase

pyridoxamine 5'-phosphate oxidase

```
Forward primer  1        CTAAGTCGCT-CACCGACAGG  20
Template        5335521  ..C..C...GG.........  5335541

Forward primer  1        CTAAGTCGCTCACCGACAGG  20
Template        5337517  .G.......GAT...G....  5337498
```

Figura 5.8: Caso que ilustra a ocorrência do oligonucleotídeo iniciador em outro genoma, porém, com *mismatches*.



Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GGTCGTAGCCGAGGATTTGT	20	59.82	55.00	3.00	0.00
Reverse primer	AGACCCTGAGGTAGACCGAC	20	60.03	60.00	5.00	2.00

Products on target templates

>NC_015848.1 Mycobacterium canettii CIPT 140010059 complete genome

product length = 227
Features associated with this product:
antitoxin VapB36

```
Forward primer  1        GGTCGTAGCCGAGGATTTGT  20
Template        2274361  ....................  2274380

Reverse primer  1        AGACCCTGAGGTAGACCGAC  20
Template        2274587  ....................  2274568
```

Figura 5.9: Exemplo de um par de oligonucleotídeo iniciador não específico que ocorre sem mismatches num dado genoma do conjunto não alvo.

5.3 Resultados

Para cada um dos casos de testes especificados na Seção 5.1, cinco diferentes avaliações foram realizadas, duas a partir do fluxo normal da metodologia apresentada e as demais removendo alguma etapa ou sendo executada de uma forma distinta da especificada. Tais avaliações foram executadas com o intuito de verificar se todos os passos são relevantes na busca por bons oligonucleotídeos iniciadores para identificar um genoma alvo.

Os testes foram realizados em equipamento com a seguinte configuração: Processador Dual Intel® Xeon® E5-2620 com frequência de 2.0GHz, 12 núcleos e 24 threads, 64GB de memória RAM e 15MB de memória cache, sendo todos os programas desenvolvidos em linguagem C.

Os resultados apresentados nas Tabelas 5.1 e 5.2 são referentes ao teste de especificidade dos oligonucleotídeos iniciadores a partir da execução completa da abordagem proposta. A diferença entre elas se dá que na Tabela 5.1 foram utilizados, para cada caso de teste, os 10 trechos que mais contém blocos com k diferenças. Ao passo que, na Tabela 5.2, foram utilizados os 10 trechos que menos contém blocos, lembrando que cada trecho deve apresentar no mínimo dois blocos. Importante destacar que no teste com as cepas de *M. bovis*, somente 7 trechos foram encontrados.

Tabela 5.1: Resultados encontrados após a execução da metodologia proposta, utilizando os trechos que contém a maior quantidade de blocos

	Total de Trechos com 2 ou mais Blocos	Trechos com Maior Quantidade de Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
				Total	%	Total	%	Total	%
Xanthomonas	328	10	50	46	92,00 %	4	8,00 %	0	0,00 %
Mycobacterium	47	10	50	38	76,00 %	12	24,00 %	0	0,00 %
cepas *M. bovis*	7	7	35	0	0,00 %	5	14,29 %	30	85,71 %

Tabela 5.2: Resultados encontrados após a execução da metodologia proposta, utilizando os trechos que contém a menor quantidade de blocos

	Total de Trechos com 2 ou mais Blocos	Trechos com Menor Quantidade de Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
				Total	%	Total	%	Total	%
Xanthomonas	328	10	50	41	82,00 %	9	18,00 %	0	0,00 %
Mycobacterium	47	10	50	15	30,00 %	31	62,00 %	4	8,00 %
cepas *M. bovis*	7	7	35	0	0,00 %	5	14,29 %	30	85,71 %

Para os testes com *Mycobacterium* e *Xanthomonas*, foi utilizado um valor de k igual a 4, resultando num total de 652 blocos encontrados no primeiro e 2823 no segundo. Chegou-se neste valor, pois, com um valor de k igual a 5, para o teste com espécies de *Mycobacterium* foi encontrado apenas 1 bloco e para o teste com as espécies de *Xanthomonas* foram encontrados um total de 13 blocos com 4 diferenças. Já para os testes com as cepas de *Mycobacterium bovis*, o valor de k foi estabelecido em 1, sendo identificado um conjunto com 121 blocos, o valor de k foi o mínimo possível, uma vez que, como os genomas apresentam elevada similaridade, para os valores de k variando de 2 a 5 um total de zero blocos foram encontrados.

Analisando os resultados das Tabelas 5.1 e 5.2, é possível verificar que a quantidade de oligonucleotídeos iniciadores específicos é relativamente superior na Tabela 5.1 com relação aos dados apresentados na Tabela 5.2, ainda, observando as duas Tabelas, tem-se que a quantidade de oligonucleotídeos iniciadores não específicos é menor na Tabela 5.1. Tais constatações estão diretamente relacionadas com a quantidade de blocos com diferenças presentes nos trechos utilizados para os testes, permitindo assim encontrar trechos melhores para a determinação de oligonucleotídeos iniciadores específicos.

Conforme descrito na Seção 4.3, a abordagem proposta seleciona os trechos de forma que o mesmo contenha a maior quantidade de blocos possíveis, de tal forma que um trecho contenha no mínimo dois blocos e o mesmo seja maximal com relação à quantidade de blocos. Por conta disso, os resultados descritos na Tabela 5.3 descrevem o teste no qual os trechos foram selecionados não em função dos blocos, mas

com relação à região gênica dos *singletons*. A comparação da Tabela 5.3 com a Tabela 5.1 certifica novamente que seguir todos os passos propostos na abordagem ajuda na busca por bons oligonucleotídeos iniciadores, apesar da diferença na qualidade dos oligonucleotídeos iniciadores ser muito próxima, a metodologia proposta para a seleção dos trechos apresenta melhores resultados.

Tabela 5.3: Resultados encontrados após a execução da metodologia proposta utilizando a região gênica dos *singletons* com maior quantidade de blocos para a seleção dos trechos

	Total Regiões Gênicas de *Singleton* com 2 ou mais Blocos	Trechos de Região Gênica de *Singleton* com Maior Quantidade de Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
				Total	%	Total	%	Total	%
Xanthomonas	24	10	50	40	80,00 %	10	20,00 %	0	0,00 %
Mycobacterium	15	10	50	32	64,00 %	18	36,00 %	0	0,00 %
cepas *M. Bovis*	5	5	25	0	0,00 %	0	0,00 %	25	100,00 %

Testes de forma a validar a busca por blocos com k diferenças, etapa descrita na Seção 4.2, também foram realizados. Para isso, foram selecionados aleatoriamente trechos do genoma alvo a partir de regiões gênicas *singletons* encontradas, conforme descrito na Seção 4.1, porém, nas regiões selecionadas para este teste não foram encontrados subcadeias com k diferenças. Ao examinar os resultados apresentados na Tabela 5.4, fica evidente a importância da busca por blocos com diferenças, tendo em vista que a totalidade dos oligonucleotídeos iniciadores encontrados se alinharam com algum outro genoma do conjunto não alvo.

Tabela 5.4: Resultados encontrados após a execução da metodologia proposta utilizando a região gênica dos *singletons* que não contém blocos para a seleção dos trechos

	Total Regiões Gênicas de *Singleton* sem Blocos	Trechos de Região Gênica de *Singleton* sem Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
				Total	%	Total	%	Total	%
Xanthomonas	61	10	50	0	0,00 %	0	0,00 %	50	100,00 %
Mycobacterium	68	10	50	0	0,00 %	0	0,00 %	50	100,00 %
cepas *M. bovis*	192	10	50	0	0,00 %	0	0,00 %	50	100,00 %

Com o objetivo de avaliar que a restrição do campo de busca a partir de regiões *singletons*, de acordo com os passos descritos na Seção 4.1, contribui para uma melhor performance na busca por bons oligonucleotídeos iniciadores específicos, foram selecionados os dez trechos com a maior quantidade de blocos do genoma alvo a

partir de regiões gênicas que não foram identificadas como *singletons* e os resultados são apresentados na Tabela 5.5. Analisando os resultados da Tabela, é possível inferir que a escolha por regiões *singletons*, além de restringir o campo de busca do genoma alvo, resulta num aprimoramento considerável, tendo em vista que, a quase totalidade dos oligonucleotídeos iniciadores são não específicos.

Tabela 5.5: Resultados encontrados após a execução da metodologia proposta utilizando a região gênica dos não *singletons* para a seleção dos trechos

	Trechos de Região Gênica Não *Singleton*	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
			Total	%	Total	%	Total	%
Xanthomonas	10	50	0	0,00 %	0	0,00 %	50	100,00 %
Mycobacterium	10	48	3	6,25 %	7	14,58 %	38	79,17 %
cepas *M. bovis*	10	50	0	0,00 %	0	0,00 %	50	100,00 %

5.4 Discussões

Os resultados apresentados na Seção 5.3, indicam que os blocos k diferentes contribuem na busca por oligonucleotídeos iniciadores específicos, uma vez que, comparando os resultados, tanto para espécies de *Mycobacterium* quanto para espécies de *Xanthomonas*, a porcentagem de oligonucleotídeos iniciadores específicos apresentados na Tabela 5.1 é maior do que aquela apresentada na Tabela 5.2, sendo a quantidade de blocos contidos nos trechos o diferencial na realização dos experimentos.

A diferença na especificidade do oligonucleotídeo iniciador é ainda mais evidente quando os resultados na Tabela 5.3 são comparados com a Tabela 5.4, sendo que nesta última em nenhum dos experimentos foram encontrados oligonucleotídeos iniciadores específicos ou com mismatches, sendo que em todos os casos, oligonucleotídeos iniciadores foram encontrados com 100% de similaridade nos genomas do conjunto não alvo e, novamente, a diferença entre os experimentos é relativa aos

blocos, na Tabela 5.3 tem-se trechos com o maior número de blocos ao passo que na Tabela 5.4 aqueles com ausência de blocos.

Considerações podem ser feitas com relação à utilização de regiões *singletons* na busca por blocos com k diferenças. Em um primeiro momento, quando confrontados os resultados retratados na Tabela 5.4 e Tabela 5.5, tem-se a impressão que pode não ter eficácia alguma a limitação do genoma alvo às regiões *singletons*, pois, a Tabela 5.5, para o teste com espécies de *Mycobacterium*, apresentou melhores resultados do que aqueles expressos na Tabela 5.4.

Contudo, tomando como exemplo os testes com espécies de *Mycobacterium*, o genoma alvo, *Mycobacterium tuberculosis* H37Rv, contém 4.411.709 bases e realizar uma busca por blocos com k diferenças em todo o genoma alvo com relação a todos do conjunto não alvo com quantidades de bases relativamente próximos seria muito custoso computacionalmente. Por isso, como já foi mencionado no Capítulo 4, faz-se necessário identificar áreas no genoma alvo com potencial para encontrar regiões que possam conter oligonucleotídeos iniciadores específicos. Este mesmo genoma contém um total de 4.036 proteínas. Logo, comparar proteínas se mostra menos dispendioso e, utilizando a metodologia adotada neste trabalho (ver Subseção 2.3.1), após esta comparação, tem-se que um total de 84 *singletons* para o genoma alvo, ou seja, um total de 67.200 bases do genoma alvo para comparação, uma redução considerável se comparadas com as 4.411.709 bases e, apesar da restrição, bons resultados podem ser encontrados de acordo com a Tabela 5.1.

É evidente, pelos números apresentados nas Tabelas 5.4 e 5.5, que somente a restrição por regiões *singletons* não resulta em trechos com bons candidatos a oligonucleotídeos iniciadores específicos, sendo esta, simplesmente uma técnica empregada

para reduzir o espaço de busca no genoma alvo preservando possíveis regiões diferentes neste genoma com relação aos do conjunto não alvo.

Ainda com relação a estratégia de reduzir a busca por oligonucleotídeos iniciadores em regiões gênicas de *singletons*, é importante observar que na Tabela 5.5 foram encontrados 3 oligonucleotídeos iniciadores específicos e 7 oligonucleotídeos iniciadores não específicos com mismatches para o teste com espécies de *Mycobacterium*, uma possível razão para estes dados se dá pelo fato de uma mesma proteína poder ser formada a partir de uma sequência diferente de nucleotídeos, como já foi dito na Subseção 2.1.3 e, com isso, mesmo que proteínas pertençam a uma mesma família, não necessariamente a sequência de nucleotídeos será a mesma.

Se para os testes com espécies de *Mycobacterium* e *Xanthomonas* os resultados foram satisfatórios, o mesmo não foi constatado no teste com cepas de *M. bovis*, onde, em nenhum dos experimentos foi encontrado oligonucleotídeo iniciador específico. Como justificativa para o desempenho ruim, duas razões podem ser apontadas.

A primeira delas diz respeito à elevada similaridade entre os genomas. Tal similaridade[3] pode ser observada nas Figuras 5.10, 5.11 e 5.12. Encontrar blocos com k diferenças em genomas tão semelhantes não foi possível para um valor de k maior que 1, comprometendo assim a qualidade dos trechos e por consequência a especificidade dos oligonucleotídeos iniciadores gerados.

Outro motivo que pode justificar o fraco desempenho da nossa metodologia com cepas de *M. bovis* é ralacionado à ferramenta utilizada para a geração de oligonucleotídeos iniciadores. Na fase de determinação dos oligonucleotídeos iniciadores a partir do trecho selecionado, a ferramenta compara os oligonucleotídeos iniciadores encontrados naquele trecho e retorna os melhores, tendo como critério o alinhamento

[3]Para o cálculo da similaridade entre os genomas foi utilizada a ferramenta "*nucleotide blast*" disponível em "http://www.blast.ncbi.nlm.nih.gov/Blast.cgi"

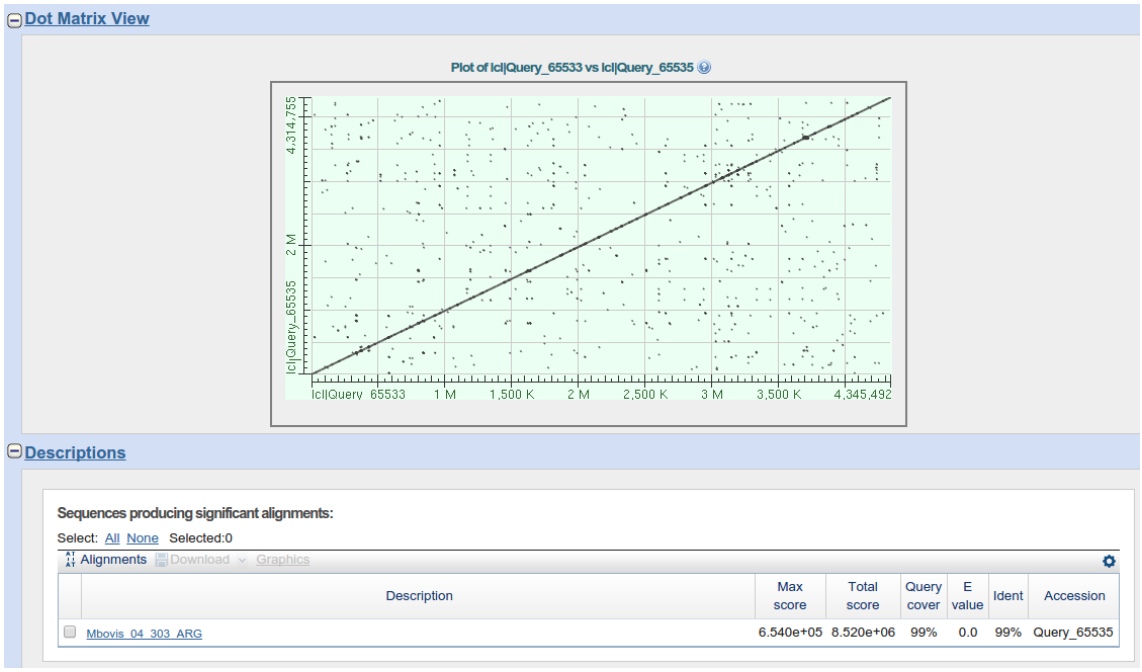


Figura 5.10: Alinhamento entre o genoma *Mycobacterium bovis* AF2122/97, Query_65533 e o genoma *Mycobacterium bovis* 04-303, Query_65535.

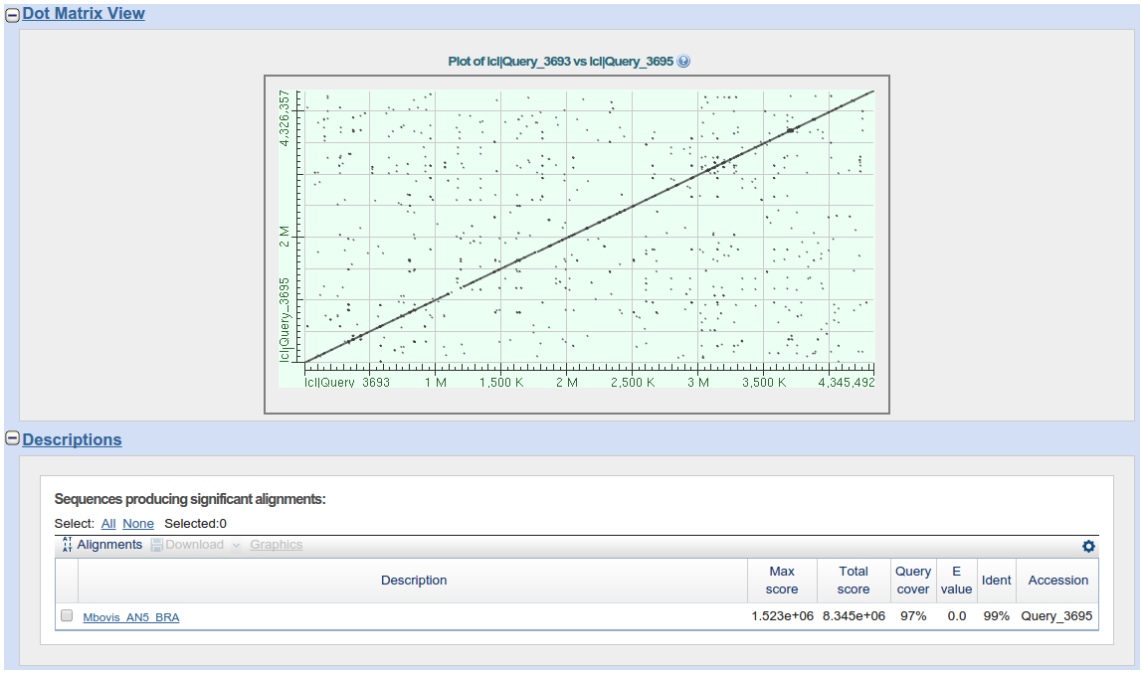


Figura 5.11: Cálculo da similaridade entre os genomas *Mycobacterium bovis* AF2122/97, Query_3693 e *Mycobacterium bovis* AN5, Query_3695.

com relação a base dados selecionada como ilustrado na Firura 5.3, ou seja, retorna os oligonucleotídeos iniciadores gerados a partir do trecho do genoma alvo que são

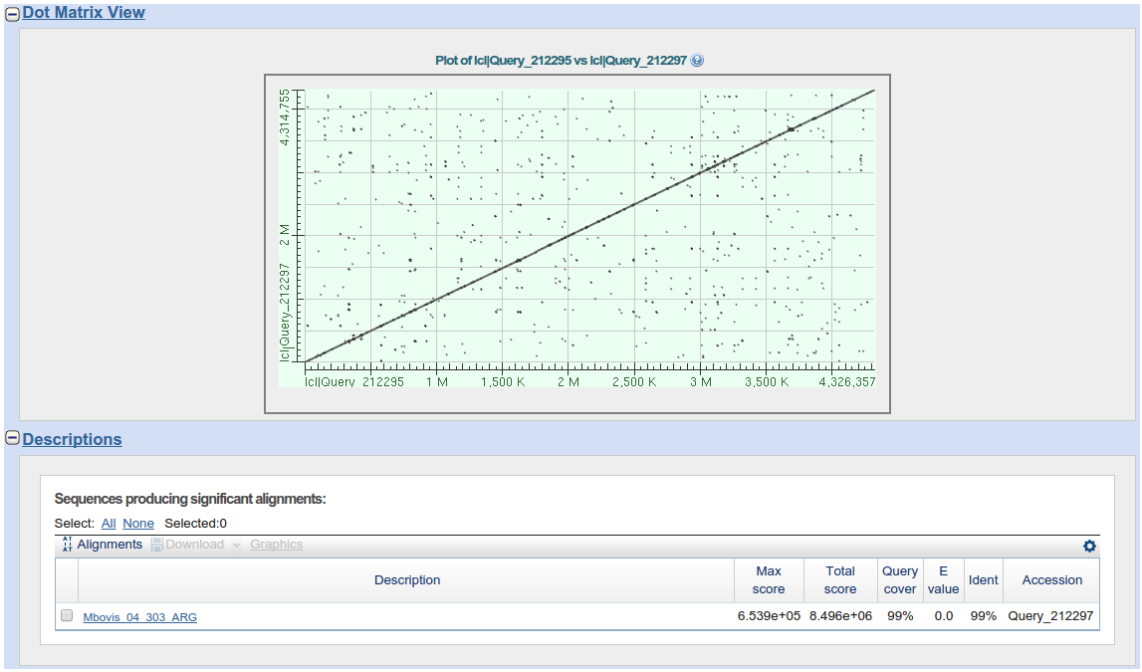


Figura 5.12: Identidade entre os genomas *Mycobacterium bovis* AN5, Query_212295 e *Mycobacterium bovis* 04-303, Query_212297 pertencentes ao conjunto não alvo.

menos similares aos genomas presentes na base de dados, visando garantir assim a especificidade do oligonucleotídeo iniciador. O problema do teste em questão para as cepas de *M. bovis* é que os genomas do conjunto não alvo não se encontram nas bases de dados do NCBI, com isso, a qualidade dos oligonucleotídeos iniciadores específicos que são retornados pode ser comprometida, influenciando diretamente no fraco desempenho da metodologia.

Na tentativa de encontrar melhores resultados para as cepas de *Mycobacterium bovis*, dois novos testes foram realizados. No primeiro, foi verificada a saída da ferramenta OrthoMCL e constatou-se a ocorrência de genes parálogos no genoma alvo, por conseguinte, tais genes não são retornados como *singletons*, pois, não são encontrados uma única vez no genoma. A partir desta observação e, selecionando genes parálogos que somente ocorrem no genoma alvo, um novo teste foi realizado, onde o resultado do mesmo pode ser visto na Tabela 5.6.

Tabela 5.6: Experimento realizado utilizando como região gênica candidata parálogos que somente ocorrem no genoma alvo

	Total Trechos com 2 ou mais Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
			Total	%	Total	%	Total	%
cepas *M. bovis*	1	5	0	0,00 %	0	0,00 %	5	100,00 %

No segundo teste, ao invés das regiões *singletons*, regiões gênicas da família de proteínas PPE foram utilizadas, uma vez que foram constatadas variações sequenciais nos genes destas famílias em genomas do gênero *Mycobacterium* [23, 34, 43]. Foram encontrados no genoma alvo um total de 57 genes da família PPE, onde 4 destes genes já haviam sido identificados como *singletons*.

Foi realizada a busca por blocos nas regiões gênicas formadas pelos 53 genes restantes da família PPE com o valor de k igual a um. Como resultado, somente quarenta e quatro blocos com uma diferença foram encontrados, dando origem a três trechos que, após submetidos à ferramenta *Primer-BLAST* e realizada análise de especificidade dos oligonucleotídeos iniciadores, não foram encontrados oligonucleotídeos iniciadores específicos. Contudo, os resultados foram discretamente superiores aqueles encontrados na Tabela 5.1, com relação a porcentagem de oligonucleotídeos iniciadores não específicos, conforme ilustrado na Tabela 5.7.

Tabela 5.7: Experimento realizado utilizando como região gênica candidata parálogos que somente ocorrem no genoma alvo

	Total Trechos com 2 ou mais Blocos	Quantidade de pares de oligonucleotídeos iniciadores	oligonucleotídeos iniciadores Específicos		oligonucleotídeos iniciadores Não Específicos com Mismatches		oligonucleotídeos iniciadores Não Específicos sem Mismatches	
			Total	%	Total	%	Total	%
cepas *M. bovis*	3	15	0	0,00 %	3	20,00 %	12	80,00 %

Após a execução dos testes adicionais para cepas de *Mycobacterium bovis*, a abordagem empregada neste trabalho não se mostrou eficiente na busca por oligonucleotídeos iniciadores específicos quando o genoma alvo e os genomas do conjunto não alvo apresentam elevada similaridade.

Capítulo 6

Conclusão

Com o surgimento de novas técnicas de baixo custo para o sequenciamento de genomas, um maior número de organismos passou a ser sequenciado. Com isso, houve um crescimento na utilização dessas sequências em diversas aplicações, dentre elas, na diferenciação de genomas, com aplicabilidade significativa na área médica no diagnóstico molecular.

Este trabalho propôs uma abordagem computacional para encontrar trechos no genoma para resolver o problema da determinação de oligonucleotídeos iniciadores específicos que possam identificar um determinado genoma com relação a outros. Resumidamente, a metodologia proposta consiste na execução das seguintes etapas:

1. Procurar *singletons* no genoma alvo;
2. Buscar por blocos com k diferenças; e
3. Determinar trechos candidatos do genoma alvo a conter oligonucleotídeos iniciadores específicos.

São dois os diferenciais da abordagem proposta. O primeiro diz respeito à restrição da busca no genoma alvo em regiões específicas. Tal restrição é definida a partir dos *singletons* encontrados no genoma alvo. Em outras palavras, sequências de bases no genoma alvo que compreendem proteínas que não possuem similaridade com qualquer proteína dos genomas do conjunto não alvo são consideradas regiões em potencial na busca por oligonucleotídeos iniciadores específicos. Já o segundo é relativo a busca por blocos do genoma alvo que somente ocorrem com k diferenças nos demais genomas. Uma vez que o objetivo é encontrar por oligonucleotídeos iniciadores específicos para um determinado genoma, esta estratégia visa garantir um mínimo de diferença entre o genoma alvo e os demais.

A partir de regiões potencialmente diferenciadas do genoma alvo com relação aos demais e da busca por blocos com k diferenças, como resultante, a metodologia proposta retorna trechos do genoma alvo que, uma vez tomados como entrada por ferramentas de determinação de oligonucleotídeos iniciadores, oligonucleotídeos iniciadores específicos possam ser encontrados.

Para os testes, foram selecionados três conjuntos de genomas, dois deles contendo genomas de espécies diferentes de um mesmo gênero, o caso de espécies do gênero *Xanthomonas* e espécies do gênero *Mycobacterium* e um terceiro contendo cepas de organismos de uma mesma espécie, *M. bovis*.

Como descrito no Capítulo 5, a metodologia proposta se mostrou eficaz nos testes realizados para diferentes espécies de um mesmo gênero, onde naqueles trechos que foram selecionados de acordo com a maior quantidade de blocos, em pelo menos 76% dos casos, os oligonucleotídeos iniciadores encontrados foram específicos para o genoma alvo. Já para o caso de teste com cepas de uma mesma espécie, a metodologia não se mostrou eficiente, não encontrando oligonucleotídeos iniciadores específicos

em nenhum dos testes realizados. Portanto, a abordagem se mostrou útil quando se trata de espécies diferentes e, não apresentou resultados mínimos satisfatórios na diferenciação de cepas de uma mesma espécie. Resultados preliminares foram publicados em [10].

Trabalhos Futuros

Dois aspectos importantes podem ser melhorados na metodologia proposta: o pré-processamento realizado para encontrar regiões específicas do genoma alvo e o tempo de processamento necessário para a execução de todas as fases da metodologia.

Como já foi discutido, a metodologia não se mostrou eficaz na busca por oligonucleotídeos iniciadores específicos para cepas da espécie *M. bovis*. Estudos podem ser realizados neste aspecto para buscar regiões com outras características de potencial diferença do genoma alvo com relação aos demais, uma vez que os genes *singletons* não se mostraram eficazes neste caso.

O tempo de processamento para buscar por blocos com k diferenças depende diretamente da quantidade de processadores disponíveis para processamento, uma vez que esta busca é realizada em paralelo. Nos testes realizados, o tempo de processamento variou de 33 horas para o teste com genomas do gênero *Mycobacterium* até 44 horas para o teste com cepas de *M. bovis*. Pesquisas nesta área podem ser realizadas na busca por arquiteturas diferentes de processamento, como por exemplo, a plataforma de computação paralela CUDA (*Compute Unified Device Architecture*) desenvolvida pela NVIDIA®.

Referências Bibliográficas

[1] K. A. Abd-Elsalam. Bioinformatic tools and guideline for pcr primer design. *African Journal of Biotechnology*, 02(5):91–95, 2003.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, EUA, 2002.

[3] K. A. Alexander, P. N. Laver, A. L. Michel, M. Williams, P. D. Helden, R. M. Warren, and C. G. Pittius. Novel *mycobacterium tuberculosis* complex pathogen, *m. mungi*. *Emerging Infectious Disease*, 16(8):1296–1299, 2010.

[4] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped Blast and Psi-Blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[5] A. Aranaz, D. Cousins, A. Mateos, and L. Domínguez. Elevation of *mycobacterium tuberculosis* subsp. caprae aranaz et al. 1999 to species rank as mycobacterium caprae comb. nov., sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 53(6):1785–1789, 2003.

[6] C. P. Araújo, A. L. A. R. Osório, K. S. G. Jorge, C. A. N. Ramos, A. F. S. Filho, C. E. S. Vidal, E. Roxo, C. Nishibe, N. F. Almeida, A. A. Fonseca-Júnior, M. R. Silva, J. D. B. Neto, V. D. Cerqueira, M. J. Zumárraga, and

F. R. Araújo. Detection of *Mycobacterium bovis* in Bovine and Bubaline Tissues Using Nested-PCR for TbD1. *PLoS One*, 9(3):e91023, 2014.

[7] R. Brosch, S. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L.M. Parsons, A. Pym, S. Samper, D. van Soolingen, and S. Cole. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *PNAS*, 99(6):3684–3689, 2012.

[8] T. A. Brown. *Genomes*. Garland Science, Department of Biomolecular Sciences, UMIST, Manchester, UK. Oxford: Wiley-Liss, 2002.

[9] A. M. Brunings and D. W. Gabriel. *Xanthomonas citri*: breaking the surface. *Molecular plant pathology*, 4(3):141–157, 2003.

[10] R. A. Cardoso, N. F. Almeida, and F. R. Araújo. Finding specific primers from syntenic blocks and characteristic strings. Submitted to International Society for Computational Biology Latin America 2014.

[11] D. V. Cousins, R. Bastida, A. Cataldi, V. Quse, S. Redrobe, S. Dow, P. Duignan, A. Murray, C. Dupont, N. Ahmed, D. M. Collins, W. R. Butler, D. Dawson, D. Rodríguez, J. Loureiro, M. I. Romano, A. Alito, M. Zumarraga, and A. Bernardelli. Tuberculosis in seals caused by a novel member of the *mycobacterium tuberculosis* complex: Mycobacterium pinnipedii sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 53(5):1305–1314, 2003.

[12] A. C. da Silva, J. A. Ferro, F. C. Reinach, C. S. Farah, L. R. Furlan, R. B. Quaggio, C. B. Monteiro-Vitorello, M. A. van Sluys, N. F. Almeida, L. M. Alves, A. M. do Amaral, M. C. Bertolini, L. E. Camargo, G. Camarotte, F. Cannavan, J. Cardozo, F. Chambergo, L. P. Ciapina, R. M. Cicarelli, L. L. Coutinho, J. R. Cursino-Santos, H. El-Dorry, J. B. Faria, A. J. Ferreira, R. C. Ferreira, M. I. Ferro, E. F. Formighieri, M. C. Franco, C. C. Greggio, A. Gruber, A. M.

Katsuyama, L. T. Kishi, R. P. Leite, E. G. Lemos, M. V. Lemos, E. C. Locali, M. A. Machado, A. M. Madeira, N. M. Martinez-Rossi, E. C. Martins, J. Meidanis, C. F. Menck, C. Y. Miyaki, D. H. Moon, L. M. Moreira, M. T. Novo, V. K. Okura, M. C. Oliveira, V. R. Oliveira, H. A. Pereira, A. Rossi, J. A. Sena, C. Silva, R. F. de Souza, L. A. Spinola, M. A. Takita, R. E. Tamura, E. C. Teixeira, R. I. Tezza, M. Trindade dos Santos, D. Truffi, S. M. Tsai, F. F. White, J. C. Setubal, and J. P. Kitajima. Comparison of the genomes of two xanthomonas pathogens with differing host specificities. *Nature*, 417(6887):459–463, 2002.

[13] P. D. Davies. Tuberculosis in humans and animals: are we a threat to each other? *Journal of the Royal Society of Medicine*, 99(10):539–540, 2006.

[14] A. J. Enright, S. van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[15] A. Escalon, S. Javegny, C. Vernière, L. D. Noël, K. Vital, S. Poussier, A. Hajri, T. Boureau, O. Pruvost, M. Arlat, and L. Gagnevin. Variations in type iii effector repertoires, pathological phenotypes and host range of *Xanthomonas citri pv. citri* pathotypes. *Molecular plant pathology*, 14(5):483–496, 2013.

[16] J. Evans, E. Smith, and et al. Cluster of human tuberculosis caused by Mycobacterium bovis: evidence for person-to-person transmission in the UK. *Lancet*, 369:1270–1276, 2007.

[17] J. O. Falkinham. Surrounded by mycobacteria: nontuberculous mycobacteria in the human environment. *J Appl Microbiol*, 2(107):356–367, 2009.

[18] N. C. Farias. Projeto de primers via árvore de sufixos., 2010. Monografia (Bacharelado em Ciência da Computação), UFMS (Universidade Federal de Mato Grosso do Sul), Campo Grande, Brazil.

[19] N. C. Farias. Orthologsorter: inferindo genotipagem e funcionalidade a partir de famílias de proteínas ortólogas. Master's thesis, UFMS (Universidade Federal de Mato Grosso do Sul), Campo Grande, Brazil, 2013.

[20] T. Garnier, K. Eigneier, and et al. The complete genome sequence of Mycobacterium bovis. *PNAS*, 100(13):7787–7782, 2003.

[21] R. Ghai, T. Hain, and T. Chakraborty. Genomeviz: visualizing microbial genomes. *BMC Bioinformatics*, 5(198), 2004.

[22] A. J. F. Griffiths, S. R. Wessler, R. C. Lewontin, and S. B. Carroll. *An Introduction to Genetic Analysis*. Freeman and Company, New York, EUA, 2000.

[23] P. W. M. Hermans, D. Van Soolingen, and J. D. A. Van Embden. Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *Journal of Bacteriology*, 174(12):4157–4165, 1992.

[24] I. Hershkovitz, H. D. Donoghue, D. E. Minnikin, H. May, O. Y. Lee, M. Feldman, E. Galili, M. Spigelman, B. M. Rothschild, and G. K. Bar-Gal. Tuberculosis origin: The neolithic scenario. *Tuberculosis*, pages 122–126, 1995.

[25] J. Ingen, Z. Rahum, A. Mulder, M. J. Boeree, R. Simeone, R. Brosch, and D. Soolingen. Characterization of *mycobacterium orygis* as *m. tuberculosis* complex subspecies. *Emerging Infectious Disease*, 18(4):653–655, 2012.

[26] M. Ito, K. Shimizu, M. Nakanishi, and A. Hashimoto. A polynomial-time algorithm for computing characteristic strings under a set of strings. *Systems and Computers in Japan*, 26(3):30–38, 1995.

[27] N. Jalan, D. Kumar, M. O. Andrade, F. Yu, J. B. Jones, J. H. Graham, F. F. White, J. C. Setubal, and N. Wang. Comparative genomic and transcriptome analyses of pathotypes of xanthomonas citri subsp. citri provide insights into mechanisms of bacterial virulence and host range. *BMC Genomics*, 14(551), 2013.

[28] L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–89, 2003.

[29] M. G. Maaß. A fast algorithm for the inexact characteristic string problem. Technical Report TUM-I0312, Fakultät für Informatik, TU München, 2003.

[30] M. L. Monaghan, M. L. Dohertya, J. D. Collinsa, J. F. Kazdab, and P. J. Quinna. The tuberculin test. *Veterinary Microbiology*, 40(1-2):111–124, 1994.

[31] L. M. Moreira, N. F. Almeida, N. Potnis, L. A. Digiampietri, S. S. Adi, J. C. Bortolossi, A. C. da Silva, A. M. da Silva, F. E. de Moraes, J. C. de Oliveira, R. F. de Souza, A. P. Facincani, A. L. Ferraz, M. I. Ferro, L. R. Furlan, D. F. Gimenez, J. B. Jones, E. W. Kitajima, M. L. Laia, R. P. Leite, M. Y. Nishiyama, J. Rodrigues Neto, L. A. Nociti, D. J. Norman, E. H. Ostroski, H. A. Pereira, B. J. Staskawicz, R. I. Tezza, J. A. Ferro, B. A. Vinatzer, and J. C. Setubal. Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of *Xanthomonas fuscans* subsp. *aurantifolii*. *BMC Genomics*, 11:238–262, 2010.

[32] L. M. Moreira, R. F. de Souza, N. F. Almeida, J. C. Setubal, J. C. Oliveira, L. R. Furlan, J. A. Ferro, and A. C. da Silva. Comparative genomics analyses of citrus-associated bacteria. *Annual review of phytopathology*, 42:163–184, 2004.

[33] J. Pevsner. *Bioinformatics and Functional Genomics*. John Willey and Sons Inc., 2009.

[34] S. Poulet and S. T. Cole. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Archives of Microbiology*, 163(2):87–95, 1995.

[35] J. A. S. Ramos. Transporte de brometo de etídio através da parede celular de *mycobacterium smegmatis*: Desenvolvimento e aplicação de metodologias de quantificação do transporte e correlação com a resistência aos antibióticos. Master's thesis, Universidade Nova de Lisboa, 05 2010.

[36] L. Rindi and C. Garzelli. Genetic diversity and phylogeny of *mycobacterium avium*. *Infect Genet Evol*, 21:375–383, 2014.

[37] R. P. Ryan, F. J. Vorhölter, N. Potnis, J. B. Jones, M. A. Van Sluys, A. J. Bogdanove, and J. M. Dow. Pathogenomics of *Xanthomonas*: understanding bacterium-plant interactions. *Nature reviews Microbiology*, 9(5):344–355, 2011.

[38] J. C. Setubal and N. F. Almeida. *Introduction to Bioinformatics using bacterial genomes*. Springer-Verlag, 2015. To be published.

[39] C. da Silva Jr, S. Sasson, and N. Caldini Jr. *Biologia*, volume Único. Saraiva, 2015.

[40] D. F. Talkington. *Real-Time PCR in Food Science: Current Technology and Applications*, chapter Introduction to the Real-time PCR. Caister Academic Press, 2013.

[41] S. Y. M. Ueki, M. C. Martins, M. A. S. Telles, M. C. Virgilio, C. M. S. Giampaglia, E. Chimara, and L. Ferrazoli. Micobactérias não tuberculosas: diversidade das espécies no estado de São Paulo. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, 41:1–8, 02 2005.

[42] A. Untergrasser, I. Cutcutache, T. Koressaar, J Ye, B.C. Faircloth, M. Remm, and S.G. Rozen. Primer3 - new capabilities and interfaces. *Nucleic Acids Research*, 40(15), 2012.

[43] R. Warren, M. Richardson, S. Sampson, J. H. Hauman, N. Beyers, P. R. Donald, and P. D. Van Helden. Genotyping of *Mycobacterium tuberculosis* with additional markers enhances accuracy in epidemiological studies. *Journal of Clinical Microbiology*, 34(9):2219–2224, 1996.

[44] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. Madden. Primer-blast: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13(134), 2012.

[45] A. Zaha. *Biologia Molecular Básica*. Mercado Aberto, Rio de Janeiro, Brasil, 2003.