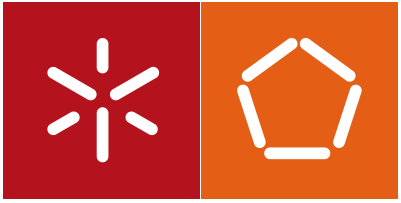
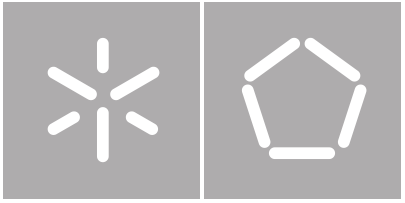


Universidade do Minho
Escola de Engenharia
Departamento de informática

José Carlos Ribeiro Pacheco

PGP: Prokaryote Gene Prediction software

outubro de 2013



Universidade do Minho
Escola de Engenharia
Departamento de informática

José Carlos Ribeiro Pacheco

PGP: Prokaryote Gene Prediction software

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de

Doutor Rui Mendes, Professor Auxiliar, Departamento de Informática, Universidade do Minho e Doutora Conceição Egas, Investigadora da Unidade de Sequenciação Avançada do Biocant, Associação de Transferência de Tecnologia

outubro de 2013

AGRADECIMENTOS

Gostaria de agradecer aos meus orientadores, Professor Rui Mendes e Doutora Conceição Egas, por toda a sua dedicação e empenho em fornecer-me rapidamente todo o auxílio, que várias vezes solicitei. A sua paciência, as suas palavras de incentivo foram importantes para etapa a etapa conseguir ser cada vez mais crítico no desenvolvimento deste trabalho.

Aos meus pais e irmã por estarem sempre ao meu lado com palavras de incentivo, sempre necessárias em momentos de aperto.

E um especial agradecimento aos técnicos de Bioinformática dos Serviços Avançados do Biocant Hugo e Felipe e à aluna de doutoramento Susana pela força e cooperação que me deram no decorrer do trabalho.

Por isso, o meu muito obrigado! A todos os que ajudaram neste trabalho.

RESUMO

A correta previsão e anotação de genes bacterianos é essencial para a aplicação da informação contida no ADN em muitos tópicos de pesquisa (bio)médica, como microbiologia, imunologia e doenças infeciosas. Embora existam vários softwares de previsão de genes bacterianos como GenemarkHMM, Glimmer e Prodigal e *pipelines* completos como ISGA, xBASE, Maker e Consensus Prediction, a previsão de genes pode ser melhorada. O principal objetivo deste trabalho foi o desenvolvimento de um *pipeline* de previsão de genes bacterianos, o Prokaryote Gene Prediction (PGP), que combina métodos de *ab initio* e de homologia. Uma vez que o *software ab initio* Prodigal mostrou um melhor desempenho relativamente a outros softwares estudados, foi usado como o passo inicial para o PGP. Considerando as proteínas previstas pelo Prodigal, o PGP a) analisa os alinhamentos obtidos, b) determina a necessidade de encurtar ou estender genes, c) introduz as correções necessárias, d) faz a previsão de ARNr e ARNt utilizando os programas RNAmmer e tRNA-scan2 e e) determina a existência de eventuais genes não identificados nas regiões intergénicas, através de um BLASTx. Quando comparados os resultados do PGP com os dados produzidos pelo Prodigal utilizando 4 genomas com conteúdo G+C% moderado e 3 com conteúdo em G+C% extremo, o PGP apresentou melhorias de 1% tanto na taxa de erro como na especificidade, exibindo a mesma sensibilidade. Foi observado que para genomas com conteúdos G+C% extremos, o PGP tem mais impacto e portanto realiza mais correções. Os resultados do PGP ainda foram comparados com os pipelines ISGA, xBASE e Consensus Prediction. O PGP melhorou a previsão de genes corretos em 4,4%, comparativamente com ISGA e xBASE e ainda 3,1% em relação à previsão do Consensus Prediction, mantendo uma sensibilidade idêntica entre previsões. No que respeita à deteção de genes na região intergénica verificou-se um acréscimo na ordem de 9 falsos positivos em 12 genomas modelo, necessitando esta vertente de um melhor desenvolvimento. Concluiu-se que o PGP melhora a correta previsão de genes, especialmente em genomas bacterianos com conteúdos G+C% extremos, contribuindo para a anotação automática de genomas bacterianos de elevada qualidade.

Palavras chave: *ab initio*, bactéria, genoma, homologia, previsão de genes, sequenciação

ABSTRACT

The correct bacterial gene prediction and annotation is essential for the application of the information contained in DNA in several areas of (bio)medicine, like microbiology, immunology and infection diseases. Although there are several softwares to perform bacterial gene prediction, like GenemarkHMM, Glimmer and Prodigal and also full pipelines as ISGA, xBASE, Maker and Consensus Prediction, gene prediction can be improved. The main objective of this work was the development of a bacterial gene prevision pipeline, the Prokaryote Gene Prediction (PGP) which combines *ab initio* and homology methods. Since the *ab initio* software Prodigal showed a better performance relatively to others studied softwares, it was used as the beginning step for the PGP. Taking into account the proteins predicted by Prodigal, the PGP a) analyses the results of the alignment, b) determines if it is necessary to shorten or extend or extension of genes, c) introduces the necessary corrections, d) predictsrRNA and tRNA using the RNAmmer and tRNA-scan2 programs and e) determines possible missing genes in intergenics regions through BLASTx. When comparing the results of PGP with data produced by Prodigal, the PGP showed improvements in both the error rate, and in the specificity, while displaying the same sensitivity. For genomes with extreme G+C% content, the PGP has higher impact and therefore performs more corrections. The results obtained with PGP were also compared with ISGA, xBASE and Consensus Prediction pipelines. The PGP improved the precision of correct genes in 4,4%, comparatively with ISGA and xBASE and 3,1% relative to the prediction of Consensus Prediction, keeping a similar sensibility among predictions. As regards the detection of genes in the intergenic region there was an increase in the range of 9 false positive in 12 model genomes, requiring this part a better development. It was concluded that PGP improves the correct prediction of genes, especially in bacterial genomes with extreme G+C% content, contributing to a high quality in automatic bacterial gene annotation.

Keys Words: *ab initio*, bacteria, genome, homology, gene prediction, sequencing

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS

ADN- Ácido desoxirribonucleico;

ARN- Ácido ribonucleico;

BLAST- Ferramenta de procura do alinhamento básico. Abreviatura adaptada da língua inglesa (*Basic Local Alignment Search Tool*);

Ch- Coordenadas do *Hit*;

CPUs- Unidade central de processamento. Abreviatura adaptada da língua inglesa (*central processing unit*);

Cq- Coordenadas da *Query*;

FN- Falsos negativos;

FP- Falsos positivos;

G+C%- Percentagem de conteúdo em guanina e citosina dos genomas ou *contigs***;**

Gff- Formato genérico da característica. Abreviatura adaptada da língua inglesa (*Generic Feature Format*);

Hit- Resultado do alinhamento efetuado pelo BLAST em relação à *Query*. Abreviatura adaptada da língua inglesa;

IN- Incorretos;

NGS- Nova geração de sequenciação. Abreviatura adaptada da língua inglesa (*New generation sequencing*);

ORF- Quadros de leitura abertos. Abreviatura adaptada da língua inglesa (*open reading frame*);

Query- Sequência de entrada no BLAST. Abreviatura adaptada da língua inglesa;

RBS- Local de ligação ribossomal. Abreviatura adaptada da língua inglesa (*Ribosomal local site*);

SC- Score de alinhamento;

SD- Regiões de Shine Dalgarno;

VP- Verdadeiros positivos;

XML- Linguagem de Marcação Extensível. Abreviatura adaptada da língua inglesa (*extensible markup language*).

Capítulo 1

1.1 MOTIVAÇÃO

Com o evoluir dos tempos a genética tem-se dotado de recursos em larga escala, quer ao nível da obtenção da informação contida nos genes e genomas, quer ao nível da interpretação dessa informação. Estes recursos genéticos são apoiados em estruturas computacionais, que constituem um precioso auxílio no processamento de dados e no armazenamento do conhecimento obtido. Mas, muito mais que isso, os recursos computacionais tornaram viáveis a construção de modelos, através do uso de um vasto leque de algoritmos, que permitem a manipulação, armazenamento e análise de toda a informação existente, com influência no progresso de áreas como a saúde ou a alimentação, entre muitas outras[1]. Tudo isto é Bioinformática, ou seja, integração da biologia em sistemas computacionais através de matérias como biologia de sistemas, estatística e informática[2].

O ponto de partida para o entendimento das sequências genéticas passa por obter a estrutura organizada dos seus nucleótidos. Com essa finalidade um dos campos em expansão da bioinformática é a sequenciação de genomas, no qual assistimos hoje a uma revolução dos seus métodos, caminhado para a 3ª geração deste tipo de tecnologia. Desde a sequenciação do primeiro genoma bacteriano (*Haemophilus influenzae* em 1995[3]) inúmeros avanços têm sido alcançados. Em 2003, existiam pouco mais de 100 genomas sequenciados. Este número duplicou em 2005 e decuplicou em 2010, isto considerando dados completamente sequenciados e disponíveis no Genbank. Este número cresce ainda mais quando se fala na base de dados GOLD que listou 5902 projetos de genomas em 2010[4].

Este volume de informação surge devido ao desenvolvimento de novas técnicas de sequenciação de nova geração (*Next Generation Sequencing* - NGS) de tal forma que hoje em dia existem diferentes plataformas como 454 Roche[5], ABI SOLiD[6] e Illumina[7] (segunda geração de sequenciadores) ou ainda mais recente o Pacific Biosciences[8] (terceira geração de sequenciadores) que tentam suplementar os custos e os processos associados à tecnologia de Sanger[9] (sequenciação de 1ª geração) por diferentes métodos (Tabela 1).

Tabela 1- Resumo comparativo das diferentes técnicas de sequenciação[10,11,12,13]**.** Gb-Gigabase, Mb-Megabase, pb-pares de base. * O Pacific Biosciences não utiliza métodos de amplificação.

Plataforma	Método de amplificação	Comprimento de Leitura	Preço	Rendimento por Leitura	Método	Precisão
Sanger	Eletroforese capilar	≈800/pb	≈$500/Mb	≈0,100 Gb	Sequenciação por síntese	99,99%
454	PCR de emulsão	≈400/pb	≈$20/Mb	≈0,250 a 4Gb	Pirosequenciação	99,00%
Illumina	PCR em fase sólida	≈100/pb	≈$0,500/Mb	≈3 a 6Gb	Sequenciação por síntese	>98,50%
SOLiD	PCR de emulsão	≈50/pb	≈$0,500/Mb	≈30 Gb	Sequenciação por ligação	99,94%
Pacific Biosciences	*	≈1300/pb	≈$2/Mb	≈0,100 Gb	Tempo real	99,90%

Estas plataformas determinam a sequência correta (embora este processo aconteça com um ligeiro erro associado) dos nucleótidos presentes no genoma. Para a sequenciação do genoma as moléculas de ácido desoxirribonucleico (ADN*)* são fragmentadas em segmentos de pequena dimensão e cada segmento é sequenciado de forma individual. Como resultado da sequenciação obtêm-se pequenas sequências – *reads* - que quando assembladas ou mapeadas (o mapeamento[a] ou a assemblagem[b] constituem o primeiro passo pós sequenciação) por grau de homologia ou *de novo,* geram sequências maiores – *contigs* - ou eventualmente a sequência completa do genoma para genomas microbianos mais pequenos[14,15].

Os baixos custos e rapidez com que os projetos de sequenciação são realizados, tornaram possível obter um conhecimento que num passado recente era difícil[16]. Em 2004 o National Institutes of Health (Estados Unidos) disponibilizou US$70 milhões para o desenvolvimento destas tecnologias de sequenciação, com vista à redução dos custos da sequenciação do genoma humano de US$3*$10^9$ para US$$10^3$ milhões. Em 2006, o X Prize Foundation (Santa Mónica, CA, USA) despendeu US$10 milhões como primeiro esforço privado para a sequenciação de 100 genomas humanos em 10 dias por menos de US$10 000 por cada genoma[17].

Na Tabela 1 verifica-se que as sequências obtidas pela tecnologia de Sanger possuem um custo associado de US$500 por Mb. Este preço de sequenciação tem vindo a diminuir com o aparecimento das novas tecnologias. Por exemplo, a utilização da tecnologia SOLiD custa US$5 por

a Mapeamento - Quando já existe um genoma sequenciado e que funciona como sequência de referência[14].
b Assemblagem - Processo pós-sequenciação para construção "de novo" de genomas, ou seja, sem existir informação prévia, pelo que é necessário utilizar ferramentas de assemblagem[15].

Mb[10].

A produção em larga escala de sequências criou um défice de compreensão da informação contida no ADN. Para se compreender esta informação é necessário identificar e caracterizar as regiões[18].

Um dos mecanismos mais eficiente seria a curadoria dos dados manualmente, isto é, análise manual da informação. Mas associado a este tipo de análise surge um outro problema, o tempo de análise. Apesar de ser bastante eficaz e assertivo, este tipo de investigação é um processo bastante moroso, pelo que não é suficiente para dar resposta ao atual volume de dados. Daí surgir a necessidade real de desenvolver ferramentas suficientemente precisas e autónomas, que realizem este processo de curadoria[18].

De certa forma a NGS tem vindo a acelerar a investigação em genomas procariotas e eucariotas e com este desenvolvimento muitos passos têm sido dados no sentido de apresentar soluções automatizadas para os processos de pós-sequenciação como: assemblagem ou mapeamento, previsão de genes e anotação funcional, que são processos necessários para interpretar os dados da sequenciação[19].

A anotação automática de genes é uma das questões essenciais na bioinformática, pelo que diversas aproximações têm vindo a ser propostas e ferramentas a ser desenvolvidas. Contudo, apesar desse desenvolvimento, muitas dessas técnicas não são exatas originando "saídas de anotação" com erros, criando a necessidade de melhoria dessas ferramentas[16].

1.2 OBJETIVOS

Após esta breve introdução compreende-se que com a tecnologia NGS tornou-se possível efetuar a descodificação genética de inúmeros organismos presentes no nosso planeta. Embora este processo tenha decrescido de complexidade, os passos posteriores ainda continuam a ser complexos e a requerer sistemas computacionais potentes, bem como, muita intervenção humana de forma a validar o processo. Deste modo, o presente trabalho visa a criação e manipulação de um sistema computacional que melhore e automatize a identificação de genes.

Em suma pretende-se a criação de um *pipeline* de previsão de genes existentes no genoma procariota, com recurso a métodos de *ab initio* e métodos por homologia (para validação ou correção dos resultados gerados pelos *ab initio*), processos essenciais na anotação de genes. Este objetivo pode-se dividir nos seguintes objetivos específicos:

Objetivos específicos:

1) Selecionar o previsor *ab initio* que atualmente oferece melhor capacidade de previsão;
2) Identificar os tipos e características de genes que são previstos de forma incorreta;
3) Desenvolver um *software* baseado em homologia de sequências para a correção de genes previstos incorretamente;
4) Desenvolver uma *pipeline* que englobe todas as ferramentas necessárias para a previsão de genes.

Estes objetivos serão concretizados através da seguinte abordagem:

1) Análise do estado de arte sobre os genomas procariotas, identificando quais os padrões essenciais na determinação das regiões codificantes, bem como, uma revisão sobre as atuais ferramentas de processamento de dados que se utilizam nos sistemas de anotação;

2) Realização de testes controlo com os *ab initio* e *pipelines*, por forma a compreender a sua manipulação e o tipo de resultados que produzem. Identificação de genes previstos de forma incorreta ou não previstos e respetivas causas.

3) Desenvolvimento do *software* em linguagem sql e perl. Este *pipeline* incorporará *softwares* de previsão, sobre os quais se procurará o desenvolvimento de novos métodos que permitam o melhoramento das previsões de genes.

4) Validação do *software* criado a partir de análises de genomas modelo para medir a precisão

com que poderá ser realizada uma previsão para genomas com pouca análise manual;

5) Comparação da ferramenta criada com outras existentes na literatura, por forma a medir o grau de confiança na estrutura construída.

1.3 ESTRUTURA DA TESE

Este documento encontra-se dividido em 4 capítulos. O presente capítulo (Capítulo 1) apresenta uma abordagem global do tema, assim como a motivação para a sua elaboração e os objetivos a cumprir.

O segundo capítulo apresenta a revisão da literatura, encontrando-se subdividido em 6 etapas. A 1ª etapa retrata os fundamentos da genética, ou seja, uma análise global sobre o ADN. Também se procurou, descrever um pouco os processos fundamentais da anotação e previsão, explicando os procedimentos gerais efetuados, bem como, uma descrição dos objetos de estudo, correspondendo estes conceitos à 2ª, 3ª e 4ª etapas. A 5ª etapa retrata algumas ferramentas (essencialmente para genomas procariotas) utilizadas nos processos de previsão. Por fim, a 6ª etapa explica os conceitos estatísticos que serão utilizados para o procedimento das análises do Capítulo 3.

O Capítulo 3 é referente ao trabalho prático elaborado, encontrando-se dividido em 3 etapas. A 1ª etapa consistiu em análises às previsões de genes de genomas procariontes, utilizando-se como *softwares* de previsão um *pipeline* existente (Maker) na Unidade de Serviços Avançados do Biocant e os previsores *ab inito* Prodigal e GeneMark. A 2ª etapa aborda um estudo aos previsores *ab initio* mais referenciados na literatura (Prodigal, GeneMark e Glimmer) por forma a compreender a precisão, de cada um, na determinação dos genes provenientes das sequências de ADN. A última etapa do capítulo 3 traduz a implementação do PGP e a comprovação de como este possui benefícios em relação a outros pipelines.

Por último, o Capítulo 4 é referente às conclusões do trabalhado elaborado fazendo-se um resumo global dos resultados obtidos. Neste capítulo são também incluídas possíveis tarefas futuras que contribuirão para o aperfeiçoamento do trabalho desenvolvido.

Capítulo 2

2.1 FUNDAMENTOS DA GENÉTICA

2.1.1 Revisão histórica

Parte do que nos define enquanto seres humanos está diretamente relacionado com a constituição génica que nos acompanha desde o início (herdada dos nossos progenitores) e com a passagem dessa informação de célula para célula durante a divisão celular, bem como, para os nossos descendentes. Contudo, à medida que as células se vão replicando e em resposta a estímulos externos, ocorrem modificações dessa informação genética, gerando diversidade[20].

Estes conceitos sobre genética e diversidade começaram a ter impacto com estudos realizados nos séculos XIX e XX através do trabalho de Charles Darwin, que em 1859 descreveu pela primeira vez a origem das espécies por seleção natural[21].

Em 1866, Mendel deu um impulso forte à comunidade científica demonstrando que certas características são passadas de geração em geração, propondo a existência de um par de unidades elementares de hereditariedade, atualmente conhecidas como genes. Em 1871, Friedrich Miescher isolou a primeira molécula de ADN descrevendo-a como provável portador da informação genética. Esta descoberta revolucionou a informação até então descrita, fazendo com que em 1953 Watson e Crick descrevessem a estrutura do ADN, revelando como este era replicado e codificava a informação genética[21].

Em 1972, Stanley realizou a primeira clonagem *in_vitro,* uma das mais revolucionárias metodologias que permitiu a multiplicação de segmentos selecionados de ADN, proporcionando a produção de diversas proteínas humanas de interesse. Em 2001 a sequência do genoma humano foi anunciada pela primeira vez, informação essa, que possibilitou o desenvolvimento da terapia genética, utilizadas no tratamento de doenças como o cancro, diabetes ou doenças neurodegenerativas (Alzheimer ou Parkinson)[22].

Atualmente a genética ainda é um campo em evolução e revolução, assistindo-se cada vez mais a avanços científicos e tecnológicos. Por isso mesmo, esta área tem conquistado um lugar de destaque entre as várias disciplinas científicas, tornando-se cada vez mais importante na sociedade atual. Esta disciplina origina melhorias a vários níveis e possibilita uma intervenção em várias áreas

como a saúde (tanto terapêutica como de diagnóstico), indústria agro-alimentar (produção em elevada escala de alimentos) ou áreas ambientais (redução dos fatores com elevado impacto ambiental)[23, 24, 25].

Como dizia James Watson "*We used to think that our fate was in our stars, but now we know, in large measures, our fate is in our genes* " [26].

2.1.2 Fundamentos genéricos do ADN

O ADN é uma molécula polimérica presente em todas as células vivas, com a capacidade de comandar o funcionamento das células, através do desencadeamento de ações. Na constituição da sua unidade básica, o nucleótido, identificam-se três constituintes fundamentais: um grupo fosfato (confere as características ácidas à molécula), uma pentose (desoxirribose) e uma base azotada, podendo estas serem: purinas (adenina e guanina) e ou pirimidinas (citosina e timina)[27].

Os nucleótidos ligam-se entre si pelo grupo fosfato ao carbono 3' da pentose no último nucleótido livre da cadeia. Este procedimento é repetido no sentido 5'- 3'. As ligações entre as pentoses e os grupos fosfato formam longas duplas cadeias de nucleótidos, que tendem a enrolar-se numa dupla hélice, através de, pontes de hidrogénio, o que lhe confere estabilidade. As duas cadeias crescem em sentidos opostos, mas mantêm a ligação por complementaridade de bases, a adenina liga-se a uma timina por duas ligações de hidrogénio e a guanina liga-se à citosina por três ligações de hidrogénio[27].

O ADN é responsável por transmitir informações vitais, uma vez que esta molécula contém a informação responsável pela formação de aminoácidos, que por sua vez terão um papel funcional na célula. O processo de formação das proteínas, exemplificado na Figura 1, inicia-se com o mecanismo da transcrição. A transcrição da informação genética nos sistemas vivos corresponde à formação do ácido ribonucleico (ARN*),*que são moléculas similares ao ADN*,* existindo em vez da desoxirribose uma ribose. Nesta etapa inicial, a ARN-polimerase fixa-se numa região promotora e desliza sobre a sequência de ADN, levando à formação da molécula pré-ARN mensageiro *(*pré-ARNm*),* ocorrendo deste modo a transcrição. O pré-ARNm apresenta na sua constituição exões e intrões, em que, após um pré-processamento, os intrões serão removidos, formando-se o ARN mensageiro (ARNm*)* funcional, que posteriormente migrará para fora do núcleo para se fixar nos ribossomas[27]. Este mecanismo ocorre nos seres eucariotas. No caso dos procariotas, como não existem intrões, o ARNm transcrito corresponde à versão final e não é necessário a sua saída para fora do núcleo, uma vez que esta estrutura não

existe[28].

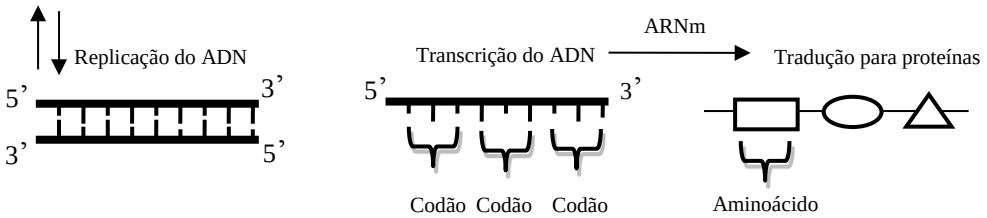


Figura 1- Resumo dos processos de manutenção celular. A replicação do ADN corresponde à formação de uma cadeia idêntica à existente. O processo de síntese de proteínas envolve duas fases: a transcrição do ADN, culminado na formação de ARNm, e a tradução no ARNm de forma a produzir a proteína .

A segunda etapa no processo de formação de proteínas consiste na tradução da informação genética. Com a tradução ocorre a descodificação da mensagem contida no ARNm, formando-se uma sequência de aminoácidos que constituem a cadeia polipeptídica[28]. Esta etapa envolve vários intervenientes como enzimas e ribossomas entre outros identificados no Tabela 2.

Tabela 2- Intervenientes no processo de tradução e respetiva função.

Intervenientes	Função
ARNm	Contém a "mensagem" a ser descodificada para a formação da proteína.
Aminoácidos	Monómeros das proteínas.
ARNt (ARN de transferência)	Seleciona e transfere os aminoácidos para a cadeia polipeptídica em formação.
Ribossomas	Complexo de construção de proteínas.
Enzimas	Catalisadores do processo.
ATP	Molécula que fornece energia às reações.

A tradução contempla três etapas (iniciação, alongamento e finalização). Na iniciação, o ARNm (proveniente da transcrição) liga-se a um complexo ribossómico na subunidade mais pequena, seguindo-se a ligação da subunidade maior com o conjunto formando uma estrutura funcional. A segunda etapa, o alongamento, reflete-se na tradução dos sucessivos codões (tripletos de nucleótidos) (Tabela 3) após o reconhecimento do codão de iniciação (AUG - o mais frequente) (Tabela 3). Os aminoácidos são sucessivamente transportados para o complexo ribossómico pelo ARNt, de acordo com a informação especificada no codão e aí se ligam através de uma ligação peptídica à cadeia de aminoácidos. Por fim a etapa de finalização ocorre quando no ARNm são encontrados codões de finalização (UAA, UAG, UGA) (Tabela 3) e a síntese proteica acaba, ocorrendo a libertação das subunidades ribossómicas[29].

Tabela 3- Codões utilizados nos processos de tradução[30]**.** Os números correspondem à probabilidade de escolha de determinado codão, neste caso referente ao organismo *Escherichia coli*.

	U			C			A			G			
U	UUU	Phe	19	UCU	Ser	9	UAU	Tyr	15	UGU	Cys	4	**U**
	UUC	Phe	17	UCC	Ser	10	UAC	Tyr	12	UGC	Cys	6	**C**
	UUA	Leu	11	UCA	Ser	6	UAA	*Stop*	2	UGA	*Stop*	0,8	**A**
	UUG	Leu	12	UCG	Ser	8	UAG	*Stop*	0,2	UGG	Trp	12	**G**
C	CUU	Leu	10	CCU	Pro	7	CAU	His	11	CGU	Arg	23	**U**
	CUC	Leu	10	CCC	Pro	5	CAC	His	10	CGC	Arg	23	**C**
	CUA	Leu	4	CCA	Pro	7	CAA	Gln	13	CGA	Arg	3	**A**
	CUG	Leu	55	COG	Pro	15	CAG	Gln	31	CGG	Arg	5	**G**
A	AUU	Ile	27	ACU	Thr	9	AAU	Asn	17	AGU	Ser	7	**U**
	AUC	Ile	27	ACC	Thr	25	AAC	Asn	24	AGC	Ser	16	**C**
	AUA	Ile	4	ACA	Thr	6	AAA	Lys	36	AGA	Arg	2	**A**
	AUG	Met	26	ACG	Thr	15	AAG	Lys	12	AGG	Arg	1	**G**
G	GUU	Val	17	GCU	Ala	16	GAU	Asp	33	GGU	Gly	24	**U**
	GUC	Val	16	GCC	Ala	25	GAC	Asp	22	GGC	Gly	33	**C**
	GUA	Val	12	GCA	Ala	16	GAA	Glu	43	GGA	Gly	6	**A**
	GUG	Val	26	GCG	Ala	37	GAG	Glu	20	GGG	Gly	10	**G**

2.2 Os genes: Objetos de estudo da anotação e previsão

Como se pode ver pelos fundamentos da genética, os genes são unidades moleculares de hereditariedade que codificam uma proteína, com função no organismo, sendo deste modo as estruturas que incorporam a informação essencial à manutenção celular[31].

Relativamente à sua estrutura é possível classificá-los em dois grandes grupos, genes eucariotas e genes procariotas.

2.2.1 Genes eucariotas

Os genes dos organismos eucariotas apresentam na sua constituição exões e intrões, bem como outras subestruturas que os permitem identificar[32]. Essas subestruturas estão representadas na Figura 2 nas letras c) a h) e nas etiquetas I e II. De modo sucinto, essas regiões podem ser descritas da seguinte forma: zona promotora ou promotor de um gene (c), sendo nesta região que se inicia a transcrição do gene[33]; local de tradução *Start* (d), região a partir da qual os fatores de tradução se ligam para dar início ao processo de tradução[30]; codão *Start (e)* que consiste em tripletos de

nucleótidos, sendo os de maior frequência AUG, GUG e UUG[29] (Tabela 3); doador (f) e recetor de *splicing (g)*, correspondendo aos limites do exão-intrão e vice-versa [34]; o codão *Stop* (h) indica o fim da tradução, podendo ser encontrado um dos seguintes três tripletos de nucleótidos: UAA, UAG e UGA[29] (Tabela 3); e as regiões 5' e 3' UTR (*Untranslated regions*, I e II), respetivamente, que se encontram antes do codão *Start* e depois do codão *Stop*. Estas regiões conferem estabilidade, definem a localização e regulam a eficiência dos mecanismos de tradução do gene codificante[35]

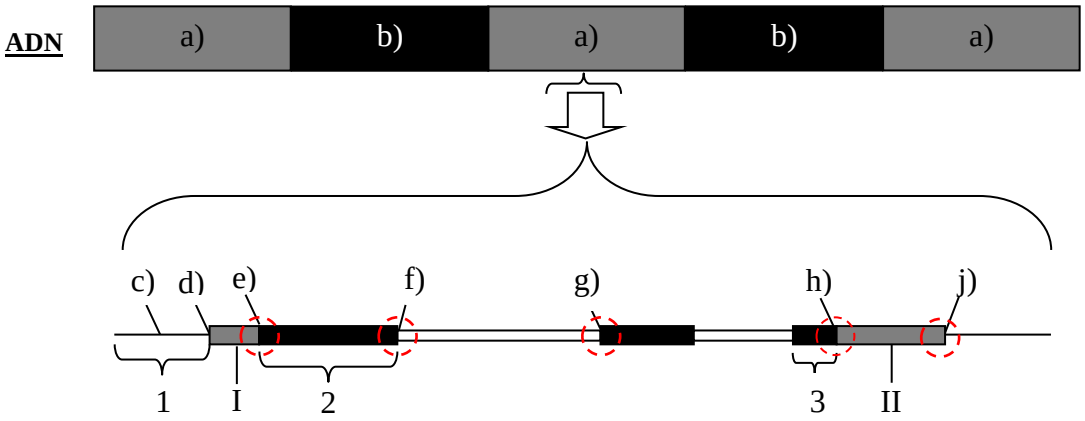


Figura 2- Estrutura do genoma eucariótico. O genoma eucariótico apresenta na sua estrutura porções codificantes (as regiões génicas - b)), e as porções não codificantes (as regiões intergénicas - a)). No interior das porções codificantes são encontradas outras subestruturas, que representam padrões conhecidos e que tem características específicas importantes para a codificação da proteína. As posições 1, 2 e 3 representam respetivamente: ilhas GpG normalmente maiores que 200 bp, exão inicial e exão final. As letras de c) a j) representam: c) região promotora, d) local inicial de tradução, e) codão *Start*, f) doador *splicing*, g) aceitador *splicing* e h) codão *Stop*. Finalmente, as porções I e II correspondem às regiões UTR, 5' UTR em I e 3' UTR em II. Figura adaptada[32].

2.2.2 Genes procariotas

O grupo dos genes procariotas compreende estruturas mais simples. Desde logo não possuem intrões como os genes eucariotas. Como pode ser visualizado na Figura 3 as estruturas internas do gene procariota, estão divididas em 3 regiões designadas como: região reguladora (1), região codificante (2) e região terminal (3)[36].

A porção reguladora de um gene procariota corresponde à porção de nucleótidos antecedentes ao codão *Start*, ou seja, à porção codificante. É através desta região que é permitida a ligação ribossomal (*Ribosomal local site*-RBS) com o ARNm e consequente início da tradução da proteína. É frequente distinguirem-se 3 regiões essenciais na porção reguladora: as regiões Shine Dalgarno (SD) (5) e as regiões TTGACA e Pribnow box (4). As regiões TTGACA são identificáveis nos 35 nucleótidos antecedentes ao codão *Start*, e tal como os Pribnow box[37,38], porções de nucleótidos a 10 nucleótidos

antes do codão *Start*, representam os sinais primários para o começo dos processos de transcrição. Consequentemente a porção mais próxima do codão *Start*, sensivelmente a 8 nucleótidos, é representada pela região de SD [39]. Ao contrário das regiões precedentes já inumeradas, a região SD apresenta mais especificidade, ou seja, esta porção conhecida de nucleótidos corresponde à posição inicial do ARNm no ribossoma, estando por isso envolvido no reconhecimento das sequências ricas em purinas, localizadas a montante ao codão *Start* no ARNm[39]. Sucintamente as regiões SD são os locais de RBS no ARNm.

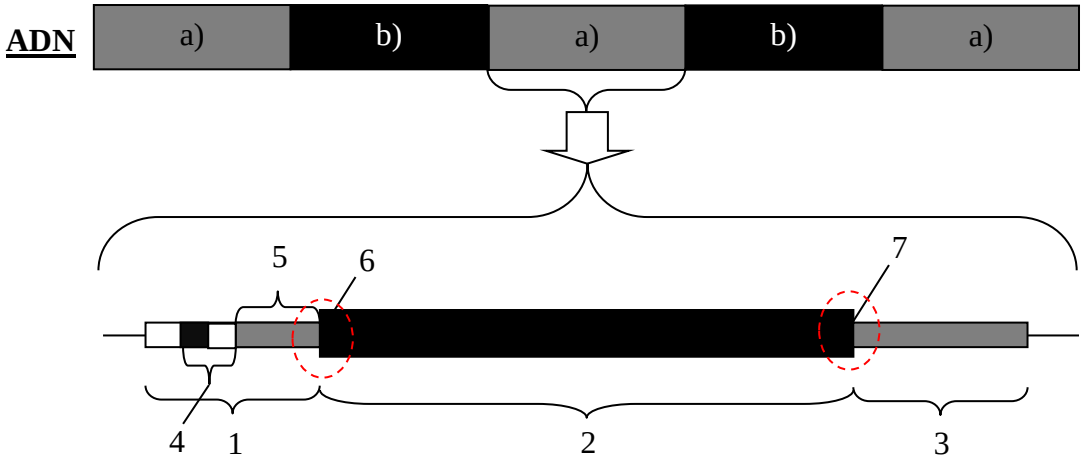


Figura 3- Estrutura do genoma procariota. À semelhança da estrutura do genoma dos seres eucariotas, também o genoma dos procariotas se encontra dividido em a) porções génicas e b) porções intergénicas. As regiões génicas apresentam as seguintes estruturas: 1 - região reguladora, 2 - região codificante, 3 - região terminal, 4 - regiões de TTGACA (preto) e Pribnow box (branco), 5 - região Shine-Dalgarno, 6 - codão *Start* e 7 - codão *Stop*. Figura adaptada[32].

A região entre o codão *Start* (ponto 6 - Figura 3) e o codão *Stop* (ponto 7 - Figura 3) – a região codificante (ponto 2 - Figura 3) - contém a informação genética essencial à formação da proteína. Na literatura, esta porção pode ser designada como *Open Reading Frame* (ORF) ou sequência codificante.

Por fim, a região terminal prolonga-se para além do codão *Stop*. As sequências posicionadas nessas regiões proporcionam à sequência transcrita uma estrutura secundária, que facilita a terminação da tradução. Por exemplo, um dos motivos estruturais bastante explorados na atualidade para a determinação dos locais terminadores são as sequências ricas em G+C%. Compreende-se conteúdos em G+C% como a determinação da frequência que estes dois nucleótidos se apresentam em determinada sequência[40].

2.3 INTRODUÇÃO AOS PROCESSOS DE ANOTAÇÃO DE GENES

Os processos de anotação podem ser sucintamente descritos como sistemas geralmente estatísticos e de procura por homologias, que compreendem por norma à seguinte métrica: padrões geralmente conhecidos A, numa dada sequência S devem ser suficientemente capazes para calcular a probabilidade condicionada dada por P (A|S), ou seja, para todos valores de A encontrar S, que traduzam a anotação da sequência mais provável de ADN[40]. De uma outra forma, esta métrica estabelece a procura de padrões numa dada sequência de ADN e através desses padrões consegue-se fazer a identificação da sequência sendo possível a atribuição de dada função aos genes previstos. Os padrões geralmente procurados para identificação dos genes são evidenciados nas Figuras 2 e 3. Segundo Lincoln Stein *et al.*, em 2001[41] o objetivo principal de anotação é a identificação das características principais do genoma, em particular do gene e dos seus produtos, sendo portanto um conjunto de passos múltiplos de procura que visam a extração de informação de ADN. Algo semelhante redige Natalie Castellana *et al.,* em 2010[36] dizendo que o principal objetivo da anotação remete-se para a identificação das coordenadas dos exões, as porções codificantes do genoma, utilizando múltiplas fontes para a identificação dessas parcelas. Deste modo, a anotação de genomas consiste na atribuição, identificação e interpretação das sequências brutas de ADN, processo este descrito sucintamente na Figura 4.

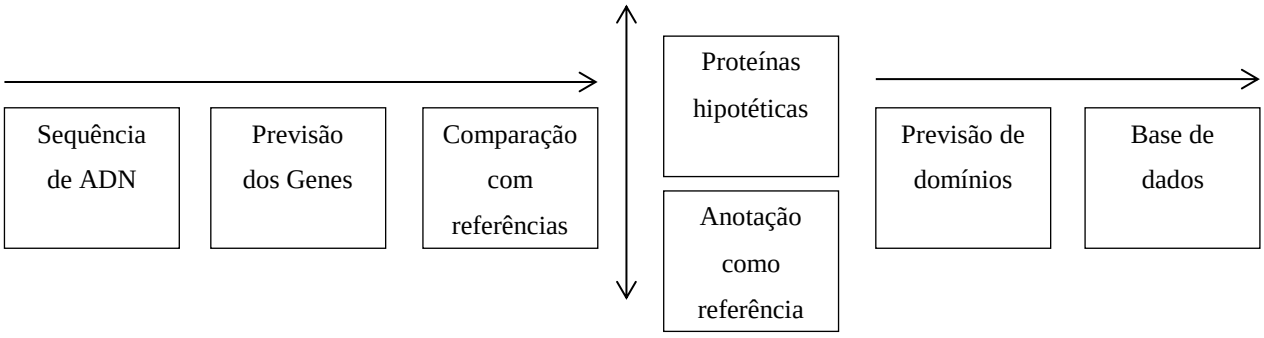


Figura 4- Processo de anotação. O primeiro passo nos processos de anotação por norma tende a ser um registo de padrões que permitam a previsão de genes provenientes na sequência fasta. Seguidamente as previsões devem ser comparadas com o que existe já anotado, por forma a ter-se a certeza no processo de anotação. Caso não exista certeza os genes previstos devem ser marcados como proteínas hipotéticas. Seguidamente deve-se prever os domínios das proteínas, submetendo-se o resultado final nas bases de dados de anotação[42].

O processo de anotação encontra-se ainda dividido em 3 grupos: i) previsão de genes (este grupo procura dar resposta à organização dos nucleótidos na sequência de ADN, isto é, através de programas conhecidos da literatura como previsores de genes, procura determinadas parcelas no

genoma e através das mesmas estabelece ligações que permitam a delineação de estratégias que distinguem porções codificantes das não codificantes), ii) anotação ao nível das proteínas (após a identificação das parcelas codificantes de determinado genoma, ou seja, após a determinação dos genes, é comum procurar-se a proteína que determinado gene codifica) e iii) anotação ao nível dos processos (tenta identificar o processo celular onde a proteína participa, ou seja, a pesquisa de vias metabólicas)[43].

2.4 INTRODUÇÃO AOS PROCESSOS DE PREVISÃO DE GENES

A previsão é um dos passos mais críticos em todo o processo de anotação, pois é nesta fase em que os genes são definidos (Figura 4). Se for previsto um gene de forma incorreta a sua anotação será também incorreta, traduzindo uma má interpretação das sequências de ADN.

Geralmente começa-se por executar um *software* de previsão genética no ficheiro fasta resultante da assemblagem/mapeamento para cada genoma que se pretende anotar. Seguidamente é comum fazer-se um processo de comparação de referências, ou seja, aceder às bases de dados e através delas procurar homologias. Se este processo for bem-sucedido, isto é, se se verificar a existência de homologias, utilizam-se referências para a previsão dos domínios, caso contrário, deve-se etiquetar a proteína como hipotética antes de se fazer o processo de previsão de domínios. Por último, adiciona-se o resultado às bases de dados para servirem de modelo a posteriores anotações, em conjunto com outras características que possam descrever a anotação [42].

Assim, um dos passos mais marcantes na previsão e ao mesmo tempo com maiores carências são os motores de previsão de genes acima introduzidos. Como alicerce principal, estas estruturas tentam a compreensão do porquê de determinados nucleótidos aparecerem em determinadas posições, bem como, compreender quais as verdadeiras regiões codificantes. Com estes resultados é então possível a atribuição de determinado padrão a dada função. Posteriormente determina-se as proteínas implícitas nos genes evidenciados e consequentemente os campos de influência dessa proteína [41]. Então como são construídos estes motores? Como são compreendidas as suas estruturas internas? Como conseguem estabelecer métricas objetivas para a identificação destas porções?

Convém salientar que atualmente o processo de previsão é feito de forma automática, mas comprovado de forma manual. Em suma, este processo procura prever os genes da forma mais precisa no menor tempo possível[44].

A previsão manual é um processo bastante moroso, com grande carga laboratorial ou de comparação com informação já disponível em base de dados, mas com elevada exatidão quanto à avaliação dos resultados. Já a previsão automática (*in silico)* é um processo realizado por estruturas computacionais, que incorporam a previsão de genes através de grupos intrínsecos ou grupos extrínsecos de modo a tirar o máximo proveito na anotação[44]. Por isso, o estudo potenciado *in silico*, onde os previsores de genes são incorporados, acelera todo o processo de validação, fazendo com que os investigadores possuam informações relevantes na construção de modelos genómicos antes de qualquer execução laboratorial, diminuindo exponencialmente o tempo dos processos de anotação manual mas mantendo, ou mesmo aumentando a sua performance[45].

Respondendo às questões acima abordadas os previsores de genes estão agrupados em dois grupos: I-grupo extrínseco (homologias) e II-intrínseco (*ab initio*), isto quanto ao nível da previsão de genes que realizam.

2.4.1 Previsão por métodos extrínsecos

As evidências extrínsecas de estruturas genéticas podem ser recolhidas de várias fontes de informação alocadas em bases de dados. Muitas dessas informações compreendem a comparação genoma-genoma, *Expressed Sequence Tags* (ESTs), alinhamentos proteicos, entre outras. Estima-se que 60-80% dos genes sequenciados possuem homologias conhecidas com outras espécies e com os avanços nos processos de anotação este número tenderá a aumentar, fazendo com que a incorporação do grupo extrínseco nos processos de anotação se torne um campo de extrema importância[46].

Para o grupo extrínseco é muito frequente o uso do algoritmo *Basic Local Alignment Search Tool* (BLAST). Como o próprio nome indica este tipo de ferramenta procura um alinhamento local contra outras sequências existentes nas bases de dados, fazendo uma procura por similaridades biológicas. O alinhamento das sequências normalmente providência a primeira conexão entre uma sequência de entrada (*Query*), com sequências já anotadas e alocadas nas bases de dados (*Hit*), fornecendo como saída não só esse alinhamento como também um conjunto de estatísticas que ajudam a determinar o grau de proximidade entre a *Query* e os diferentes *Hits*. O BLAST requer como entrada uma sequência de aminoácidos ou nucleótidos, realizando posteriormente uma procura por sequências homólogas armazenadas nas bases de dados, que é escolhida pelo utilizador no início da procura[47]. Esta ferramenta é bastante útil existindo atualmente diferentes tipos da mesma (Tabela 4),

como por exemplo o BLASTp, que auxilia na determinação das porções codificantes de um genoma realizando a procura e alinhamento de aminoácidos contra dados anotados de proteínas[48].

Tabela 4- Diferentes variações da ferramenta bioinformática BLAST[48].

Programa	Tipo de sequência	Alvo da sequência	Descrição da ferramenta
BLASTp	Proteína	Proteína	Compara uma sequência de aminoácidos contra a base de dados de proteínas.
BLASTn	Nucleótido	Nucleótido	Compara a sequência de nucleótidos contra a base de dados de nucleótidos.
BLASTx	Nucleótido (tradução)	Proteína	Compara a sequência de nucleótidos traduzidos em todas as *reading frames*, contra as bases de dados de proteínas.
tBLASTn	Proteína	Nucleótido (tradução)	Compara a sequência de proteínas contra a sequência das bases de dados dinamicamente traduzidas em todas as *reading frame*s.
tBLASTx	Nucleótido (tradução)	Nucleótido (tradução)	Compara as seis possíveis ORFs traduzidas da sequência de nucleótidos contra as seis possíveis traduções dos nucleótidos nas bases de dados.

Contudo, essas evidências ainda não são completas, muito por falta de genomas completamente anotados até à data, o que traz problemas de incoerência quando apenas este tipo de anotação é realizado. Desta forma, as evidências extrínsecas não são normalmente suficientes para cobrir a estrutura completa de todos genes, são antes marcas que evidenciam as anotações realizadas pelo grupo intrínseco, caracterizando todas as evidências demarcadas por este grupo com a literatura[49].

2.4.2 Previsão por métodos intrínsecos

Os métodos intrínsecos são normalmente considerados como aqueles que olham para dentro da sequência e através dela tentam identificar determinados padrões que são conhecidos da literatura como codificadores de determinadas regiões importantes à formação das proteínas[49].

De um modo geral, todos os previsores de genes começam com a procura de ORFs. Esta porção de ADN inclui a parte codificante da proteína, existindo 6 possíveis combinações para cada ORF, 3 no sentido *forward* ou *sense* (s*trand* positiva) e 3 no sentido r*everse* ou *antisense* (*strand* negativa). Estas possíveis combinações estão relacionadas com a posição inicial do codão de dada ORF. Por exemplo, dada a sequência ATGGGGGGC as possíveis ORFs são ATGGGGGGC/ TGGGGG/ GGGGGG em sentido *Upstream* e CGGGGGGTA/ GGGGGG/ GGGGGT em sentido *downstream*. Como se pode verificar começando numa das três possíveis posições que determinam a possível ORFs obtém-se conjuntos distintos de codões, que por sua vez podem codificar diferentes proteínas. Convém salientar que entre estes conjuntos de ORFs somente uma das possíveis *frames* contém a porção

codificante correta[28,29].

Normalmente, um genoma inclui inúmeros nucleótidos e com eles inúmeras possíveis ORFs, desta forma como primeiro método de previsão os previsores de genes, implementam modelos escondidos de Markov ou interpolações deste. Estes modelos correspondem a algoritmos que representam o sistema como conjuntos de estados discretos e transições entre estados. Cada transição é associada a uma probabilidade e com o conjunto das probabilidades conseguimos determinar as possíveis ORFs[49].

Após o cálculo das possíveis ORFs, convém fazer uma filtragem, de maneira a eliminar o máximo de falsos positivos que os modelos de Markov produzem. Uma dessas filtragens consiste na procura de conteúdos G+C%. As ORFs com maior conteúdo em G+C% são providas de poucos codões *Stop* e tem mais hipótese de serem verdadeiras ORFs. Apesar de se utilizar esta abordagem de forma a reduzir os falsos positivos na escolha das ORFs e essencialmente nos codões *Stop*, os falsos positivos para codões *Start* aumentam. Ou seja, dentro de uma mesma ORF podemos encontrar vários codões *Start* (estes apresentam dupla funcionalidade, pois podem demarcar o início da tradução ou apenas codificar um aminoácido na proteína) pelo que a escolha de um deles pode implicar uma maior ou menor extensão da possível ORF[46]. Por isso, a escolha assertiva destes codões é de extrema importância na previsão genética e corresponde a outro nível de filtragem das possíveis ORFs. Na procura dos verdadeiros codões *Start* é utilizada a procura de regiões (acima já referidas Figuras 2 e 3) Pribnow Box, Shine Dalgarno, TTGACA, ilhas GpG e local inicial de tardução[42].

Normalmente estas filtragens são acompanhadas de sistemas de pontuação, pelo que no final de todo processo somente as regiões com maior pontuação é que são selecionadas para os estudos seguintes[40,46,50].

2.5 Ferramentas de Previsão de genes

Atualmente e com o aumento de genomas sequenciados pelas plataformas de sequenciação massiva, as ferramentas de anotação estão em constante evolução, tornando-se cada vez mais eficazes na previsão de genes e estruturas implícitas nas sequências[51]. Exemplos dessas ferramentas são os *softwares* de previsão de genes como o Glimmer[52], o GeneMark[44], o Prodigal[46] e o Critica[50] sendo estes mais adequadas a genomas procariotas. Existem outros com funções similares como o Augustus[53] ou o

SNAP[45] para estruturas eucariotas. Estes não serão abordados uma vez que o trabalho passará apenas por genomas procariotas.

2.5.1 Softwares ab initio

2.5.1.1 Glimmer

Este *software*, incorpora-se na classe intrínseca de previsão genética, ou seja, na procura de padrões internos evidenciados na sequência gerada pela sequenciação e posterior assemblagem. Sendo um *ab initio* o Glimmer tem como principal propósito encontrar potenciais genes a partir de uma sequência genética. O principal campo de estudo foca-se em organismos como bactérias, *archaea* e vírus, geralmente com genomas muito densos, onde regiões que codificam as proteínas compreendem cerca de 90% ou mais da sequência de ADN. Deste modo, exatidão da avaliação de genes procariotas depende em primeiro lugar da identificação do verdadeiro gene, existente numa das seis *frames* de leitura possíveis[52].

O sistema pode de forma rápida treinar um modelo, realizando a previsão e usando somente a sequência de interesse como fonte de entrada. Para a previsão do modelo, faz interpolação de modelos de Markov, ou seja, calcula a probabilidade de dado intervalo determina a sequência de ADN. O resultado é gerado por um modelo codificante, contra ADN não codificante, isto é, através deste sistema consegue-se a probabilidade de determinado nucleótido receber um subconjunto de posições adjacentes. Quanto ao algoritmo por trás deste previsor genético estão incorporados essencialmente quatro componentes: i) *score* reverso das possíveis ORF, ii) locais RBS, iii) redução de sobreposições e iv) melhoramento do método através de treino de ORFs longas[52].

2.5.1.2 GeneMark

Este modelo computacional de previsão genética foi modelado de forma a conseguir boas soluções para os limites dos genes. Ao contrário do Glimmer, este *software* utiliza modelos escondidos de Markov, com o qual realiza a previsão de genes modelando as transições entre estados. O GeneMark utiliza ainda a procura de padrões de RBS com intuito de melhorar a previsão dos codões *Start* [44]. O algoritmo implementado no GeneMark começa por demarcar todas as possíveis *ORFs* que

compõem uma sequência. Seguidamente, cada possível ORF será pontuada. Esta tarefa de pontuar e classificar as possíveis ORF sofrerá sucessivas filtragens, ou melhoramentos compreendendo as seguintes etapas: i) modelos de estruturas de sequências para procariotas (modelos escondidos de Markov), ii) pós-processamento com procura de RBS, iii) derivação dos modelos probabilísticos das RBS[40].

2.5.1.3 Prodigal

O Prodigal foca-se essencialmente em três objetivos: melhorar a previsão da estrutura do gene, melhorar o reconhecimento de locais iniciais de tradução e reduzir os falsos positivos obtidos durante a previsão. À semelhança do GeneMark, faz uso dos modelos escondidos de Markov para a previsão de padrões internos que fazem distinção entre possíveis regiões codificantes e não codificantes[46].

O primeiro passo do algoritmo do qual o Prodigal se rege é a procura de ORFs, em regiões com elevada frequência de conteúdos G+C%. Para isso começa por localizar todos os possíveis codões *Start* e *Stop* e constrói um quadro com as tendências de ligação das posições *Start* e*Stop*, isto de acordo com o tamanho das possíveis ORFs. Através deste quadro são obtidos *scores* de forma a se encontrar as posições *Start* e escolher as verdadeiras regiões codificantes.

Para melhorar a previsão da posição *Start* o Prodigal insere no seu código a procura de regiões RBS. Deste modo, é construído um *background* de RBS e ATG/GTG/TTG para todas as ligações das posições *Start* e *Stop*. Dependendo das distâncias obtidas da posição *Start* e RBS é escolhido a melhor posição *Start*.

Por fim, através do algoritmo, o Prodigal faz uma filtragem das evidências diminuindo a criação de falsos positivos[46].

2.5.1.4 Critica

O *software* Critica faz a identificação de regiões codificantes prováveis e identificáveis nas sequências de estudo. Para o efeito, realiza análises de comparação nas sequências de ADN armazenadas nas bases de dados biológicas que forneceram um conjunto de dados padrão para a formação de uma base de dados interna, utilizada para treinar em conjunto com os métodos não comparativos de ADN. Assim, o Critica procura alinhamentos com sequências armazenadas na base

de dados interna. Se a tradução resultante desses alinhamentos identifica um aminoácido potencial de codificação, então essa região é considerada como uma evidência de codificação[50].

Sinteticamente, este programa utiliza 4 passos essenciais para fazer a análise de determinada sequência. No primeiro passo há uma quebra uma dada sequência em tripletos, calculando posteriormente um *score* (baseado nas semelhanças que esse codão tem numa sequência codificante, em vez de numa sequência não codificante). O segundo passo faz a identificação das regiões que tem um *score* elevado aleatório para codificação. Seguidamente, o terceiro passo faz a extensão do candidato, na região codificante, no sentido do último codão da *Query* da sequência. O passo quarto examina os efeitos da escolha de cada codão *Start* disponível, por incorporação de um *score* de preferências de codão *Start* e um *score* para qualquer potencial sequência SD hélice[50].

2.5.2 Pipelines de previsão

As ferramentas de anotação nem sempre são fáceis de manusear, requerendo um conhecimento prévio das suas estruturas internas. De forma a ultrapassar esta questão existem alguns modelos simples que integram na sua algoritmia muitas das estruturas computacionais acima mencionadas bem como outras (por exemplo tRNAscan-SE[54] ou RNAmmer[55], que procuram o ARNt e ARNm, respetivamente), possibilitando aos utilizadores formas cómodas e rápidas de anotação.

Estas ferramentas são *pipelines* que têm como função principal identificar regiões codificantes, de modo a que se consiga atribuir o seu significado biológico, ou seja, atribuir a determinada região codificante a função desempenhada no genoma, ou por exemplo, atribuir as vias metabólicas que estas possuem influência, bem como, outras características implícitas nos processos de anotação[56]. Posteriormente, os resultados gerados por estes *pipelines* são produzidos em ficheiros do tipo *Generic Feature Format* (gff3), GenBank, entre outros, que poderão ser automaticamente carregados para visualizadores/editores (como por exemplo o Artemis[57]).

Atualmente, existem alguns p*ipelines* como o XNOMEX [56], o Maker[58], o ISGA [59], o RAST[60], o xBASE[61], o BASys[62]. Contudo para o trabalho elaborado apenas se fará manipulação do *software* Maker.

2.5.2.1- Maker

O Maker é um *pipeline* de anotação de genomas eucarióticos e procariotas (mas mais adaptado a genomas eucariotas). Esta ferramenta identifica, repete e alinha ESTs e proteínas, produzindo diagnósticos sobre determinado genoma, construindo automaticamente os dados para a anotação dos genes englobados no genoma de estudo[58].

O Maker possui uma arquitetura modular que se estende a classes encriptadas com Bioperl, GenericHot e GenericHSP. A estrutura modular faz com que o processo de anotação seja realizado em 5 etapas distintas:

1ª Etapa- Fase computacional: Nesta etapa, a sequência de entrada é analisada por um conjunto de 4 programas: o RepeatMasker é utilizado na análise de genomas para mascarar as repetições de baixa complexidade, permitindo a exclusão das regiões mascaradas nos alinhamentos efetuados pelo programa BLAST. O Exonerate é um algoritmo de alinhamento capaz de alinhar as proteínas e sequências de nucleótidos do genoma. E por fim, o SNAP é o programa que permite a realização de treino de previsão adicional[58].

2ª Etapa- Filtros: Esta fase consiste em filtrar os resultados, identificando e removendo previsões erradas[58].

3ª Etapa- Polimento: Neste passo os *Hits* de maior sucesso obtidos no BLAST serão alinhados segundo um algoritmo de alinhamento, procurando assim mais precisão nos limites dos exões[58].

4ª Etapa- Síntese: Na 4ª etapa, o Maker sintetiza informações sobre ESTs, *clusters* e alinhamentos de proteínas para produzir provas consistentes na anotação. Por último estas provas de anotação são introduzidas no programa SNAP, que com base nesta informação realiza a tarefa acima descrita[58].

5ª Etapa-Anotação: Os resultados da análise que o programa SNAP realiza, na etapa anterior, serão posteriormente recombinados com evidências, por forma a gerar anotações completas. Desta forma, o Maker faz uma previsão dos dados gerados pelo SNAP, contra EST, ARNm e regiões 5' e 3' não traduzidas identificando as coordenadas possíveis dos exões codificantes[58].

O Maker não é uma ferramenta exaustiva, isto é, não procura uma solução abrangente para todos os problemas de anotação genómica. Em vez disso, este sistema é projetado para dar inicio a anotações de modelos emergentes.

2.5.3 Software de correção de ORFs previstas

Apesar dos *ab initio* apresentaram uma elevada *performance* nos resultados de previsão, apresentam ainda um erro associado. Por esse facto existem *softwares* que tentam alertar os utilizadores desse erro, facultando possíveis melhoramentos às saídas providenciadas pelos *ab initio*. Um exemplo desse tipo *software* é o GenePRIMP[63]:

2.5.3.1 GenePRIMP

O GenePRIMP é um *software* que avalia as evidências baseadas em modelos de genes procariotas e reporta anomalias que poderão auxiliar o procedimento de anotação ou de curadoria manual das regiões codificantes ou génicas. Basicamente este *software* é um auxílio para a anotação manual, pois reduz as inconstâncias geradas pelos previsores genéticos, fazendo um relatório demonstrativo das irregularidades geradas no decorrer dos previsores de genes pela utilização da ferramenta BLAST. Como demonstra a Figura 5, na análise manual, com os dados gerados do GenePRIMP possuiu-se um relatório de: a) genes curtos, b) genes longos, c) genes partidos, d) genes interrompidos, e) genes únicos, f) possíveis genes perdidos e g) genes duvidosos.

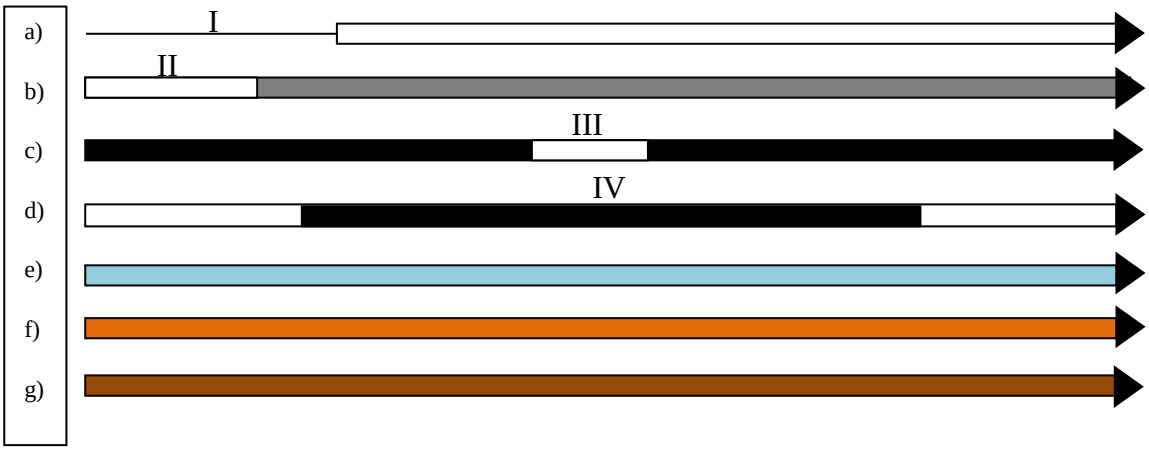


Figura 5- Ilustração gráfica do relatório gerado pelo GenePRIMP. a) gene curto devido à ausência do fragmento I verificado nos seus genes homólogos resultantes da pesquisa BLAST efetuada, b) gene longo, por aparecimento de uma porção II em excesso quando comparado com os genes homólogos, c) gene partido por falta do fragmento III encontrado em genes homólogos, d) gene interrompido devido ao desaparecimento de longas cadeias de aminoácidos como em IV, e) gene único sem resultados de *Hit*s no BLAST, f) possível gene perdido pelos previsores de genes, que faz *match* com gene homólogos durante o BLAST g) gene duvidoso devido a possuir um *Hit* duvidoso no resultado do BLAST.

Para a deteção de genes considerados como longos ou curtos, o GenePrimp usa um critério chamado *score* de alinhamento (SC), que é calculado através dos resultados obtidos no BLAST, sendo descrito por:

Fórmula 1: $$SC = \frac{Cq - Ch}{Cq + Ch}$$

onde Cq corresponde às coordenadas do local *Start* da *Query* prevista pelo previsor, e o Ch das coordenadas do local *Start* da *Query* homóloga. Realizado isto, o valor alcançado é comparado com um valor padrão (que foi determinado pelos criadores do GenePRIMP) e assim o gene previsto é designado como curto, longo ou correto. Para a deteção de genes partidos ou interrompidos o GenePRIMP realiza um BLAST por forma a encontrar 2 genes (identificados no previsor *ab initio*) chamados adjacentes a outro (encontrado por homologias no BLAST), no qual se verifica a ausência de nucleótidos que ligariam os dois genes previstos formando o homólogo.

Outra das aplicações deste *software* incluí o melhoramento da qualidade final da deteção dos *frameshifts* (mutações genéticas causadas por inserções ou deleções de nucleótidos na sequência de ADN[64]) na sequência gerada, através de várias tecnologias de análise a genomas de fungos e eucariotas superiores com menor probabilidades em se associar heurísticas (métodos online que não serão trabalhados nesta tese)[63].

2.6 Termos Estatísticos

Ao longo deste documento vão ser utilizados alguns termos estatísticos de forma a comparar as previsões efetuadas por diferentes *softwares*. Deste modo, este capítulo insere os conceitos estatísticos gerais utilizados.

As comparações efetuadas terão sempre em conta as posições *Start* e *Stop* com que os genes são previstos de acordo com os genomas de referência. Assim, dependendo das posições, os genes poderão ser considerados como falsos negativos (FN), falsos positivos (FP), verdadeiros positivos (VP) e de incorretos (IN) como assinalado na Figura 6.

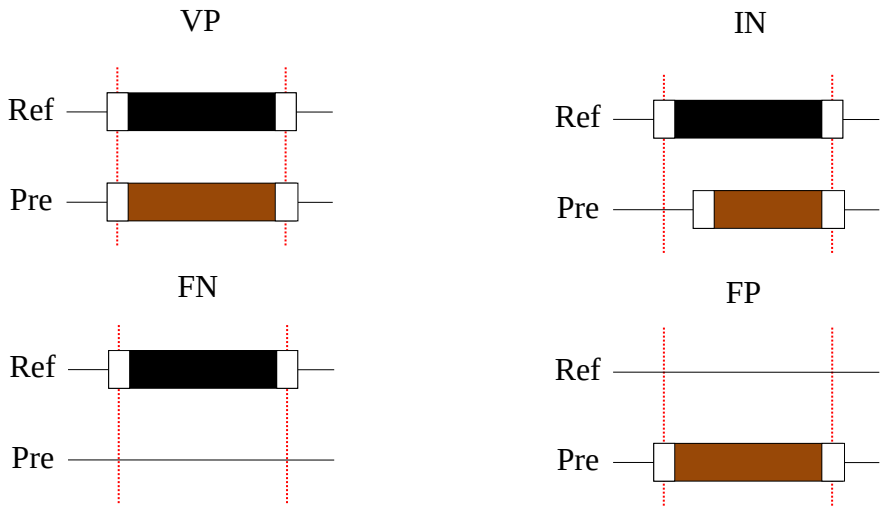


Figura 6- Determinação de genes. Ref representa um gene anotado de um genoma de referência. Pre representa um gene previsto por um *software*. Dependendo das posições *Start* e *Stop*, os genes neste trabalho foram assinalados como: verdadeiros positivos (VP) quando ambas as posições são coincidentes, incorrectos (IN) quando a posição *Start* do gene previsto não coincide com o gene do genoma anotado no genoma de referência, falsos negativos (FN) quando um gene anotado no genoma de referência não é previsto pelo *software* e falso positivo (FP) quando um gene é previsto pelo software mas não existe num gene anotado no genoma de referência.

A determinação dos genes considerados como VP, IN, FN e FP permite a determinação das seguintes formas de cálculo: taxa de erro (consiste no quociente entre o somatório da adição das previsões como IN e FP com o somatório da adição das previsões assinaladas como VP, IN e FP), especificidade (consiste no inverso da taxa de erro, ou seja na diferença de 1 pela taxa de erro), sensibilidade (é quociente entre os somatório da adição dos VP e IN com o número de genes existentes no genoma de referência) e impacto (consiste no produto da taxa de erro com o quociente da especificidade pelo *rank*; correspondendo o *rank* à pontuação atribuída a um conjunto de parâmetros, ou seja, num caso de avaliação de 3 parâmetros o melhor classificado toma o valor de um, o segundo classificado toma o valor de 2 e o terceiro classificado o valor de 3) dado pelas seguintes fórmulas[65]:

Fórmula 2: $$Taxa\ de\ erro = \frac{\sum(Previsões\ Incorretas + Falsos\ Positivos)}{\sum(Verdadeiros\ Positivos + Previsões\ Incorrectas + Falsos\ Positivos)}$$

Fórmula 3: $$Especificidade = (1 - \text{Taxa de erro})$$

Fórmula 4: $$Sensibilidade = \frac{\sum(Verdadeiros\ Positivos + Incorretas)}{Número\ de\ genes\ no\ genoma\ de\ referencia}$$

Fórmula 5: $$Impacto = (1 - Especificidade) * \left(\frac{Sensibilidade}{Rank}\right)$$

Capítulo 3

3.1 Primeira estratégia – O Bom (Prodigal), o Mau (GeneMark) e o Vilão (Maker)

Como primeira estratégia para implementar um *pipeline* para previsão de genes de genoma procariontes, utilizou-se como modelo um *pipeline* já existente na Unidade de Serviços Avançados do Biocant. Este utilizava o GeneMark como previsor inicial incorporado no Maker por forma a melhorar os resultados previstos.

O Prodigal é um previsor de genes mais recente e mais assertivo quando comparado com o GeneMark[18,66]. Desta forma, o primeiro passo deste trabalho consistiu em averiguar a possibilidade da integração do Prodigal no *pipeline* já implementado.

3.1.1 Métodos

3.1.1.1 Previsão das posições *Start* e *Stop* - Prodigal vs GeneMark

Utilizaram-se 3 *contigs* com aproximadamente 50 mil pares de bases de três bactérias para testar preliminarmente os *softwares* Prodigal e GeneMark: *Rubrobacter radiotolerans* (com ≈70% em conteúdo G+C%), *Escherichia coli* (com ≈50% em conteúdo G+C%) e *Acinetobacter baumannii* (com ≈40% em conteúdo G+C%). Os *contigs* de *A. baumannii* [67] e *E. coli* [70] foram obtidos de forma aleatória da base de dados de genomas do NCBI. Já o *contig* de *R. radiotolerans* foi retirado da sequência do genoma da bactéria sequenciada na Unidade de Serviços Avançados do Biocant.

As sequências dos *contigs* das diferentes bactérias serviram de entrada tanto ao Prodigal como ao GeneMark. O Prodigal foi executado usando a tabela de código genético 11 e treino das regiões Shine-Dalgarno, produzindo saídas em formato gff3. Para o GeneMark inicialmente é necessário o cálculo do conteúdo em G+C%, etapa que foi realizada com o recurso a um *script* de nome probuild. Este cálculo do conteúdo em G+C% é essencial pois permite escolher a heurística necessária à execução do GeneMark. No GeneMark também se utilizou a determinação das RBSs para a determinação das previsões, sendo o resultado final providenciado em formato gff2. Os ficheiros em formato gff contêm as posições de *Start* e *Stop*, sendo deste modo os ficheiros utilizados para as

análises posteriores. Nesta análise as posições *Start* e *Stop* foram comparadas de forma a perceber quais as diferenças entre *softwares*.

3.1.1.2 Previsão de genes - Prodigal vs GeneMark vs Maker

Para realizar *benchmarking* entre Prodigal, GeneMark e Maker utilizaram-se 2 genomas de diferentes estirpes da bactéria *R. radiotolerans* contendo aproximadamente 2850 genes previstos em cada conjunto. Um dos genomas foi anotado pelo Biocant, utilizando a ferramenta Maker com o previsor GeneMark. O outro genoma foi anotado pelo Joint Genome Institute (JGI) recorrendo ao *software* Prodigal e com um *pipeline* por eles construído.

Foram usados também 2 genomas modelos com conteúdo em G+C% similar (aproximadamente 50%). O genoma da *E. coli* (NC_000913) e o genoma do *Bacillus subtilis* (NC_000964), obtidos na base de dados do NCBI, dos quais se retiraram as sequências em formato fasta e a anotação final em formato Genbank.

Os ficheiros fasta relativos aos genomas descritos anteriormente foram usados como ficheiros de entrada nos programas Prodigal e GeneMark. Os parâmetros e o modo utilizado para executar o GeneMark e Prodigal encontram-se descritos no capítulo 3 secção 3.1.1.1. Para a utilização do Maker utilizaram-se parâmetros diferentes. Quando se utilizou o GeneMark como previsor inicial no Maker, uma vez que este já se encontra implementado, foram utilizados os seguintes parâmetros: como entrada foi dada unicamente a sequência fasta, sobre a qual é necessário a determinação do conteúdo em G+C%; utilizou-se o repeat masking por forma a mascarar as regiões repetitivas em genomas procariotas; foram utilizados de 16 cpus (*central processing unit*) para executar o Maker, e os restantes parâmetros foram os existentes por defeito. Para a execução do conjunto Prodigal e Maker, visto que o Prodigal não se encontra implementado no Maker, foi necessário executá-lo externamente e dar o resultado das previsões efetuadas em formato gff, com os resultados das proteínas previstas e o resultado das sequências de nucleótidos referente às proteínas. À exceção da introdução do conteúdo em G+C%, os restantes parâmetros foram semelhantes ao conjunto GeneMark e Maker.

Para a elaboração desta análise apenas se considerou nos genomas de referência os genes assinalados como regiões codificantes, por forma a se poder determinar os FN, FP, VP e IN e consequentemente calcular a taxa de erro, especificidade e sensibilidade dadas pelas fórmulas 2, 3 e 4 indicadas no capítulo 2 secção 2.6. A determinação destes parâmetros estatísticos permitiu a

verificação da precisão com que as previsões foram efetuadas pelos diferentes previsores em relação às respetivas referências.

3.1.2 Resultados e discussão

3.1.2.1 Previsão das posições *Start* e *Stop* - Prodigal vs GeneMark

Em conjunto, os previsores Prodigal e GeneMark identificaram 218 genes idênticos de um total de 293 (soma de todos os genes previstos pelos *ab initio* para as três bactérias), originando uma correspondência de ≈74%. Verificou-se também que tanto o Prodigal como o GeneMark previram aproximadamente o mesmo número de genes, com diferença de ≈1 gene para cada *contig* analisado (Tabela 5). Os restantes ≈16% corresponderam a diferenças obtidas devido ao uso de métricas distintas internas de pontuação que os previsores Prodigal e GeneMark usam para a validação dos candidatos a genes verdadeiros e a genes falsos.

Tabela 5- Diferenças entre as previsões dos *softwares* Prodigal e GeneMark. Genes GeneMark e genes Prodigal correspondem ao total de genes previstos utilizando o GeneMark e Prodigal como previsores, respetivamente. Posições *Start* e *Stop* iguais correspondem às posições iguais entre previsores, neste caso Prodigal e GeneMark. Posições *Start* e *Stop* diferentes correspondem ao número de genes previstos por um previsor (por exemplo o Prodigal) que não possui previsão idêntica quando utilizado outro previsor (por exemplo o GeneMark).

Contig	Genes GeneMark	Genes Prodigal	Posições *Start* iguais	Posições *Stop* iguais	*Posições Start* e *Stop* diferentes
Rubrobacter radiotolerans	48	47	30	46	1
Escherichia coli	40	41	26	38	3
Acinetobacter baumannii	58	59	51	58	1

Analisando somente o número de diferenças para os vários *contigs* é possível verificar que o número de igualdades de posições *Start* é menor relativamente ao número de igualdades de posições *Stop.*

No *contig* de *R. radiotolerans* observou-se o maior número de diferenças entre posições. Neste caso o conteúdo em G+C% é elevado e por isso é permissível a conclusão que altos conteúdos em G+C% facilitam a determinação das posições *Stop* mas dificultam a determinação das posições *Start* (neste caso verificou-se uma igualdade de 46 posições *Stop* entre 47 possíveis e 30 posições *Start* iguais de 48 possíveis) indo deste modo ao encontro com o que é descrito na literatura.

Em contrapartida, quando os conteúdos em G+C% são mais baixos, a determinação das posições *Start* afigura-se mais assertiva. Exemplo disso são as previsões dos *ab initio* Prodigal e GeneMark para o *contig* de *A.baumannii* (neste caso foram assinalados 51 genes de 58 possíveis com a mesma posição *Start* e todos os 58 possíveis com a mesma posição *Stop*).

Na análise efetuada para a *E. coli*, que representa um *contig* com conteúdo em G+C% intermédio, verificou-se que o número de igualdades entre os previsores Prodigal e GeneMark para as posições *Start* aumentou em relação ao *contig* da *R. radiotolerans*, contudo este aumento não se verificou tão significativo como no caso da *A.baumannii.*

Conclui-se assim que a maior diferença entre estes dois previsores estará na correta previsão da posição *Start*, sendo dependente do conteúdo em G+C% do *contig* procariota[69, 70].

3.1.2.2 Previsão de genes - Prodigal vs GeneMark vs Maker

O genoma anotado pelo Biocant, quando foi utilizado o GeneMark como previsor, apresentou melhores resultados comparativamente aos resultados gerados pelo Prodigal. Neste caso, averiguou-se a existência de 52 FN e 60 FP, uma percentagem de especificidade de ≈89% com sensibilidade correspondente de ≈98%, como demonstra a Tabela 6. Para o mesmo genoma, com o Prodigal, verificou-se um aumento de 27 FN e de 262 FP (relativamente à previsão efetuada com o GeneMark) e um decréscimo na especificidade (foram apresentados valores de especificidade de ≈73%). Contudo, a sensibilidade com que o genoma foi previsto pelo Prodigal foi semelhante à do GeneMark, apresentando um valor de ≈97%.

Para os dados anotados pelo JGI, como demonstra a Tabela 6, para o genoma de *R. radiotolerans* observou-se que o Prodigal conseguiu prever os genes com uma especificidade de ≈90% e uma sensibilidade de ≈98%. Estes valores apresentaram-se superiores aos apresentados pelo GeneMark (o GeneMark realizou a previsão para o genoma da *R. radiotolerans* anotado no JGI com valores de especificidade de ≈77% e sensibilidade de ≈98%).

Tabela 6- Análises a 2 genomas não modelo de *R. radiotolerans*. As métricas de análise são os falsos negativos (FN), falsos positivos (FP), verdadeiros positivos (VP) e incorretos (IN). A taxa de erro é dada pela fórmula 1, a precisão é dada pela fórmula 2 e a sensibilidade pela fórmula 3 do capítulo 2 secção 2.6. *Nos genomas de referência apenas se consideraram as posições marcadas como regiões codificantes por forma a realizar esta análise.

Genoma	Teste	FN	FP	VP	IN	Taxa de erro	Especificidade	Sensibilidade
R. radiotolerans (Biocant)	GeneMark	52	60	2611	236	0,102	0,898	0,982
	GeneMark Maker	20	82	2848	38	0,040	0,960	0,993
	Prodigal	79	322	2111	444	0,266	0,734	0,970
	Maker Prodigal	91	545	1894	424	0,338	0,662	0,962
R. radiotolerans (JGI)	GeneMark	64	64	2203	602	0,232	0,768	0,978
	GeneMark Maker	67	137	2013	784	0,314	0,686	0,977
	Prodigal	54	38	2590	239	0,097	0,903	0,981
	Prodigal Maker	57	119	2571	269	0,131	0,869	0,980

Com as análises levadas a cabo para o Maker com o genoma de *R. radiotolerans* anotado pelo Biocant, quando o previsor utilizado foi o GeneMark, foram obtidos 20 FN, 82 FP, uma percentagem de especificidade de ≈96 % e uma sensibilidade de ≈99%. Quando os dados foram providenciados pelo Prodigal ao Maker, estes valores decresceram apresentando 91 FN, 545 FP, ≈66% de especificidade e uma sensibilidade de ≈96%. Para esta análise a escolha do conjunto GeneMark Maker apresentou-se mais assertiva que o conjunto Prodigal Maker.

Para o genoma anotado pelo JGI averiguou-se a tendência contrária, ou seja, o conjunto Maker Prodigal apresentou melhores resultados, obtendo-se neste caso 57 FN e 119 FP, um valor de precisão de ≈87%, sendo a sensibilidade de ≈98%. Os resultados do conjunto GeneMark Maker apresentaram-se inferiores: 67 FN, 137 FP, um valor de 77% para a precisão e uma sensibilidade de aproximadamente ≈97%.

Assim, analisando somente os genomas da *R. radiotolerans* parece verificar-se uma tendência para o previsor utilizado, ou seja, os dados revelaram-se melhores para o GeneMark no genoma anotado pelo Biocant (revelando-se ainda mais favoráveis quando o GeneMark é incorporado com o Maker) e mais favoráveis para o genoma anotado pelo JGI quando o *ab initio* foi o Prodigal. Como redigido anteriormente, para o genoma de *R. radiotolerans* anotado pelo Biocant utilizou-se o *ab initio* GeneMark, já com o genoma anotado pelo JGI o *ab initio* utilizado foi o Prodigal. No entanto, não se pode mencionar que os resultados gerados pelos *ab initio* e pelos conjuntos Prodigal com Maker e GeneMark com Maker aqui estudados estejam errados. Embora os *softwares* de previsão procurem a maior fidelidade dos dados gerados eles não procuram a perfeição. A sua maior "missão" é a produção de uma hipótese inicial mais próxima da real, com a qual será mais fácil os trabalhos de análise

manual na validação dos modelos genómicos. Eventualmente este processo de trabalho manual mudará os resultados levados a cabo pelos *ab initio* quanto aos genes previstos, mas o processo inicial facilitará exponencialmente esta tarefa[51].

Paralelamente a esta análise procedeu-se a outro estudo com genomas modelo de *E. coli* e *B. subtillis*, por forma a compreender o benefício dos *ab initio* e a integração destes no Maker. Pelo facto de serem genomas modelos, a sua anotação encontra-se mais precisa relativamente aos genomas pouco anotados, permitindo assim uma maior confiança aquando da comparação com os resultados obtidos nestas análises.

Para os genomas modelo utilizados nesta abordagem conseguiu-se perceber que o Maker possui uma precisão mais baixa que os *ab initio*. Para o genoma modelo *E. coli* os resultados de taxa de erro, especificidade e sensibilidade foram superiores quando somente o previsor *ab initio* foi utilizado (tanto para o Prodigal como para o GeneMark), gerando previsões com menos FT, FN, IN e mais VP (Tabela 7). Para o genoma *B. subtilis* apurou-se o mesmo.

Tabela 7- Análises a 2 genomas modelo, um de *B. subtilis* e outro de *E. coli*. As métricas de análise são os falsos negativos (FN), falsos positivos (FP), verdadeiros positivos (VP) e incorretos (IN). A taxa de erro é dada pela fórmula 1, a precisão é dada pela fórmula 2 e a sensibilidade dada pela fórmula 3 do capítulo 2 secção 2.6. *Nos genomas de referência apenas se considerou as posições marcadas como regiões codificantes por forma a realizar esta análise.

Genoma	Teste	FN	FP	VP	IN	Taxa de erro	Especificidade	Sensibilidade
B. subtilis	GeneMark	115	195	2820	1241	0,338	0,662	0,970
	GeneMark Maker	133	254	2732	1313	0,365	0,635	0,969
	Prodigal	63	108	3753	360	0,111	0,889	0,985
	Prodigal Maker	68	129	3479	629	0,179	0,821	0,984
E. coli	GeneMark	258	243	3089	974	0,280	0,720	0,940
	GeneMark Maker	265	393	3358	1059	0,302	0,698	0,930
	Prodigal	176	165	3826	319	0,111	0,889	0,960
	Prodigal Maker	174	299	3570	577	0,197	0,803	0,960

Deste modo conclui-se que o Maker, tende a piorar as previsões efetuadas pelos *ab initio*, ou seja, quando se executam os *ab initio* individualmente e sem intervenção do Maker estes conseguem ser mais precisos. Estes dados vêm de encontro com um estudo levado a cabo por Angelova *et al.*, em 2010[18], onde os programas Maker, GeneMark e Prodigal foram testados para *Pseudomonas aeruginosa* LESB58 (P.a. LESB58), com conteúdos de G+C% aproximado de ≈66%. Nesse estudo verificou-se que os previsores Prodigal e GeneMark revelaram melhores resultados de previsão,

apresentando melhores percentagens de genes corretos em relação aos genomas de referência para o Prodigal (≈87%) e para o GeneMark (≈82 %) relativamente ao Maker (≈74%).

3.2 A NOVA ESTRATÉGIA. ESCOLHA DO MELHOR PREVISOR

Com a análise efetuada anteriormente conseguiu-se perceber que os *ab initio* possuem diferentes formas de realizarem as suas previsões, e com isso geram resultados diferentes. Concluiu-se também que para genomas pouco anotados, ou seja, que ainda possuem falta de análise manual, parece existir uma tendência para os *softwares* utilizados. Para além disso, foi também possível observar que a utilização do *software* Maker não é a mais indicada para a tarefa de previsão de genes procariotas. Desta forma, a estratégia apresentada nos capítulos anteriores foi remodelada no sentido de criar um *pipeline* de previsão de genes mais eficiente.

Para o desenvolvimento deste *pipeline* aplicou-se o conhecimento adquirido anteriormente, sendo necessário utilizar ainda genomas com conteúdos G+C% variável e serem genomas modelo (Tabela 8). Neste capítulo pretende-se perceber realmente qual o previsor dos mais referenciados na literatura (Prodigal, GeneMark e Glimmer) com melhor precisão na determinação dos genes provenientes das sequências de ADN.

Tabela 8- Genomas modelo em estudo.

	Genoma	Número de acesso no NCBI	G+C%	Gram(+/-)
I	*Mycoplasma mobile*	NC_006908.1	≈25	-
II	*Lactococcus lactis*	NC_013656.1	≈35	+
III	*Haemophilus influenzae*	NC_000907.1	≈38	-
IV	*Streptococcus pneumoniae*	NC_003028.3	≈40	+
V	*Lactobacillus plantarum*	NC_004567.1	≈44	+
VI	*Bacillus subtilis*	NC_000964.3	≈45	+
VII	*Escherichia coli*	NC_000913.2	≈50	-
VIII	*Neisseria meningitis*	NC_003112.2	≈52	-
IX	*Salmonela enterica*	NC_004631.1	≈52	-
X	*Pseudomonas putida*	NC_002947.3	≈61	-
XI	*Mycobacterium tuberculosis*	NC_000962.2	≈66	-
XII	*Streptomyces coelicolor*	NC_003888.3	≈72	+

3.2.1 Métodos

Usaram-se as sequências fastas dos genomas apresentados na Tabela 8, obtidas da base de dados NCBI, como ficheiro de entrada nos programas Prodigal, GeneMark e Glimmer. Os parâmetros de execução do Prodigal e GeneMark encontram-se descritos no capítulo 3 secção 3.1.1.1. No caso do Glimmer este foi executado via *Webserver*

(*www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi,* acedido em junho de 2013) com a tabela de código genético 11 e topologia circular.

Os resultados de cada previsor foram comparados com a anotação depositada na base de dados (ficheiro GenBank) considerando apenas os genes codificadores. Desta forma foi possível a determinação dos FP, FN, TP e IN. Consequentemente procedeu-se a uma análise que ilustra a classificação para as métricas estatísticas: sensibilidade, especificidade e taxa de erro, onde se calculou as médias para essas métricas.

3.2.2 Resultados e discussão

Os resultados obtidos, apresentados na Tabela 9, comprovaram que o Prodigal tende a ser aquele que prevê mais ORFs corretas.

Tabela 9- Resultados da comparação entre Prodigal, GeneMark e Glimmer. Falsos negativos (FN), falsos positivos (FP), Verdadeiros positivos (VP) e incorretos (IN). Verde corresponde ao melhor resultado, branco ao resultado intermédio e cinzento ao pior resultado. Nos genomas de referência providenciados pelo ficheiro GenBank apenas se consideraram as posições marcadas com a *Tag* das regiões codificantes (*coding sequence*-CDS) por forma a realizar esta análise.

	I-*Mycoplasma mobile*				II-*Lactococcus lactis*				III-*Haemophilus influenzae*			
Software	FN	FP	VP	IN	FN	FP	VP	IN	FN	FP	VP	IN
Prodigal	259	926	136	238	47	119	2235	162	20	109	1474	163
GeneMark	263	988	134	236	33	177	2044	367	25	154	1435	196
Glimmer	265	946	127	241	1177	1338	1102	165	786	946	710	160
	IV-*Streptococcus pneumoniae*				V-*Lactobacillus plantarum*				VI-*Bacillus subtilis*			
Software	FN	FP	VP	IN	FN	FP	VP	IN	FN	FP	VP	IN
Prodigal	186	197	1670	249	40	74	2770	247	63	108	3753	360
GeneMark	167	269	1603	335	91	106	2203	763	115	195	2820	1241
Glimmer	1035	1165	856	214	51	174	2710	296	2183	2436	1696	297
	VII-*Escherichia coli*				VIII-*Neisseria meningitis*				IX-*Salmonela enterica*			
Software	FN	FP	VP	IN	FN	FP	VP	IN	FN	FP	VP	IN
Prodigal	176	166	3827	318	177	234	1623	263	149	417	3845	376
GeneMark	237	226	3091	990	223	354	1483	357	226	482	3162	982
Glimmer	2142	2307	1738	438	1054	1598	720	289	2005	2580	1855	510
	X-*Pseudomanas putida*				XI-*Mycobacterium tuberculosis*				XII-*Streptomyces coelicolor*			
Software	FN	FP	VP	IN	FN	FP	VP	IN	FN	FP	VP	IN
Prodigal	175	409	4107	1068	134	215	2973	896	213	186	5995	1560
GeneMark	254	562	3790	1306	205	221	2448	1350	212	309	5500	2056
Glimmer	186	471	3857	1307	139	479	2558	1306	235	680	5418	2115

A avaliação dos IN não deve ser interpretada como um fator muito influente na previsão, pois apesar de algumas das ORF previstas por um *ab initio* terem demonstrado um maior número de IN relativamente a outras, este número não invalida que determinado *ab initio* tenha errado completamente as suas previsões. Por exemplo, o Glimmer tendeu a apresentar mais FN, valor que o Prodigal diminuiu apresentando mais VP e mais IN (relembro que os IN são considerados como uma ORF onde uma das posições é erradamente prevista relativamente ao genoma de referência).

O GeneMark foi o segundo melhor previsor. De uma forma geral previu mais VP e IN e menos FN e FP quando comparado com o Glimmer.

A Tabela 10 retrata os resultados de uma forma mais precisa pela implementação das fórmulas de cálculo de sensibilidade, especificidade e taxa de erro. Assim, analisando os resultados verifica-se que o Prodigal para todas as fórmulas aplicadas apresentou melhores valores, tendo sido o *ab initio* com maior sensibilidade (nesta métrica ocorreram 3 exceções, mas os valores foram semelhantes ao que apresentou o melhor valor) mais especificidade e menor taxa de erro associado.

Tabela 10- Resultado das comparações para os 12 genomas modelo. A taxa de erro, especificidade e sensibilidade dados pelas fórmulas 4, 5 e 6 respetivamente. Genomas: I- *M. mobile*, II- *L. lactis*, III-*H.influenzae*, IV- *S. pneumoniae*, V- *L. plantarum*, VI- *B. subtilis*, VII- *E. coli*, VIII- *N. meningites*, IX- *S. entérica*, X- *P. putida*, XI- *M. tuberculosis* e XII- *S. coelicolor*. Nos genomas de referência providenciados pelo ficheiro GenBank apenas se consideraram as posições marcadas com a *Tag* das regiões codificantes (*coding sequence*-CDS) por forma a realizar esta análise.

	Taxa de erro			Especificidade			Sensibilidade		
Genoma	Prodigal	GeneMark	Glimmer	Prodigal	GeneMark	Glimmer	Prodigal	GeneMark	Glimmer
I	0,895	0,901	0,903	0,105	0,099	0,097	0,591	0,585	0,581
II	0,112	0,223	0,615	0,888	0,777	0,385	0,981	0,986	0,518
III	0,156	0,211	0,668	0,844	0,789	0,332	0,988	0,985	0,525
IV	0,211	0,287	0,655	0,789	0,713	0,345	0,912	0,921	0,508
V	0,104	0,283	0,148	0,896	0,717	0,852	0,987	0,97	0,983
VI	0,112	0,344	0,654	0,888	0,656	0,346	0,985	0,972	0,477
VII	0,112	0,282	0,636	0,888	0,718	0,364	0,959	0,944	0,504
VIII	0,234	0,345	0,915	0,766	0,655	0,085	0,914	0,892	0,489
IX	0,171	0,335	0,707	0,829	0,665	0,293	0,966	0,948	0,541
X	0,265	0,330	0,316	0,735	0,67	0,684	0,967	0,953	0,965
XI	0,272	0,391	0,411	0,728	0,609	0,589	0,967	0,949	0,965
XII	0,226	0,301	0,340	0,774	0,699	0,66	0,973	0,973	0,97
Média	**0,239**	0,353	0,581	**0,761**	0,647	0,419	**0,933**	0,923	0,669

O GeneMark apresentou-se como o segundo melhor previsor para as análises efetuadas, contudo ao contrário do Prodigal este não conseguiu ser superior em todos os campos estatísticos em relação ao Glimmer. Por exemplo para os genomas V e X os valores de sensibilidade, especificidade e

taxa de erro foram mais favoráveis ao Glimmer, assim como no genoma XI os valores de sensibilidade também foram favoráveis ao Glimmer.

O Prodigal demonstrou ser o mais eficaz na tarefa de previsão quando comparando com os previsores GeneMark e Glimmer. Em média o Prodigal evidenciou ORFs com uma sensibilidade ≈93% uma especificidade de ≈76% e com uma taxa de erro associada de ≈23%. Por sua vez, o GeneMark manifestou prever genes com uma sensibilidade de ≈93 %, uma especificidade de ≈64% e uma taxa de erro associada de ≈39%. O Glimmer de entre os três *ab initio* testados, foi o que apresentou piores resultados possuindo por norma uma sensibilidade de ≈66%, uma especificidade de ≈41% e uma taxa de erro de ≈58%.

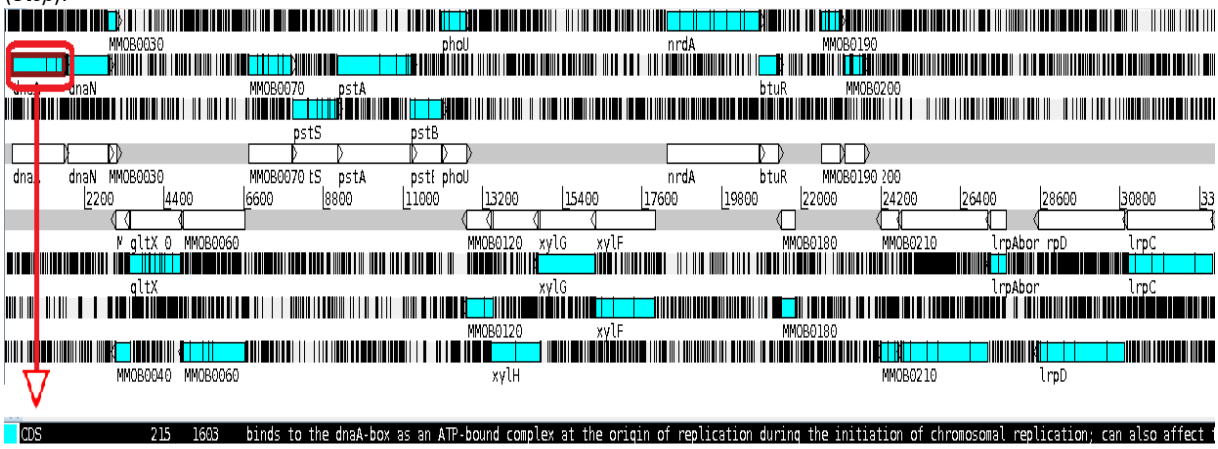
É de salientar que estes valores não se verificaram em todos os genomas, pois por exemplo para o genoma I (*M. mobile*) parecem ter ocorrido falhas graves durante as previsões dos genes, sendo que para este caso os valores de sensibilidade e de especificidade diminuíram para todos os previsores, contrapondo um aumento da taxa de erro.

Esta discrepância foi analisada com o *software* Artemis[c] (procedeu-se à introdução do ficheiro em formato GenBank), por forma a compreender a disparidade de valores obtidos. A Figura 7 demonstra que no genoma anotado existem vários codões *Stop* no interior de genes, revelando que a anotação deste parece incorreta. Por exemplo, na região codificante assinalada no genoma de referência com as posições 215 e 1603, existem 3 codões *Stop* na mesma *frame* e s*trand*, que na prática não pode existir. Quando se utilizaram os *softwares ab initio* este gene foi assinalado, mas com posição *Stop* diferente (para o Prodigal a posição *Stop* foi 1141 e sem codões *Stop* no interior do gene).

Outro indício que permitiu esta conclusão depara-se na existência de muitos codões stop ao longo da sequência de ADN, ou seja, pela análise da Figura 7 verifica-se que o genoma é extremamente preenchido por codões *Stop* o que leva a supor que o erro possa estar associado à sequência de ADN (possivelmente numa inserção ou deleção de algum nucleótido aquando a sua submissão) resultando numa incongruência entre a sequência de ADN e as posições indicadas no ficheiro gff como ORFs e a respetiva anotação. Contudo, como o principal foco deste trabalho não se enquadrava neste âmbito de descobrir os possíveis erros aqui detetados não foram pesquisados indícios mais fortes.

[c] *Software* que permite a edição e visualização dos resultados da previsão-anotação[57].

Figura 7- Ficheiro GenBank da *M. mobile* visualizado com o *software* Artemis. Verifica-se que o genoma encontra-se extremamente preenchido por codões *Stop* representados por traços negros nas ORFs (campos azuis). As 3 primeiras linhas correspondem às regiões codificantes a montante para as 3 diferentes *frames*. As linhas do meio correspondem aos genes nas duas *strands*. As três linhas finais correspondem às regiões codificantes a jusante nas 3 diferentes *frames*. A vermelho está assinalado uma região codificante descrita corresponde às posições 215 (*Start*) e 1603 (*Stop*).



3.3 PGP: PROKARYOTE GENOME PREDICTION *SOFTWARE*

Como se tem vindo a verificar, os *ab initio* são ferramentas de extrema importância para os processos de anotação, que sem recursos a dados anotados, conseguem prever as porções codificantes dos *contigs* ou genomas. Contudo, estes *softwares* ainda acarretam debilidades, como a previsão de FP e IN e a não previsão de FN gerando incoerências nos processos de anotação. Assim, para combater os erros gerados pelos *ab initio* procedeu-se à elaboração do PGP, um *pipeline* de previsão de ORFs, que automaticamente corrige possíveis erros gerados pelos *ab intio* através de métodos de procura de homologias. Em suma, um sistema híbrido que valida e corrige os dados gerados pelos *ab initio*.

Sucintamente o PGP (Figura 8) compreende três etapas, com múltiplas tarefas entre elas. A primeira etapa do PGP consiste no início da previsão de genes, com recurso ao *ab initio* Prodigal. Como consequência dessa previsão serão geradas proteínas que formam as sequências de entradas no BLASTp. O BLASTp gera a informação necessária para a execução da segunda etapa, ou seja, a execução dos filtros que validam ou corrigem as ORFs previstas. Por fim, como terceira e última etapa, o PGP procura nas regiões intergénicas possíveis genes que não foram detetados no decurso do algoritmo implementado, pela realização de um BLASTx e pela implementação de um filtro que interpreta os resultados, fazendo a identificação dos novos genes.

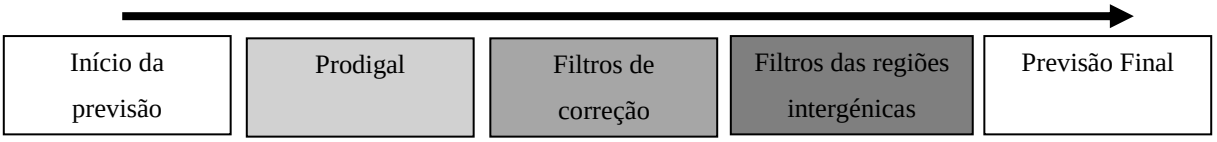


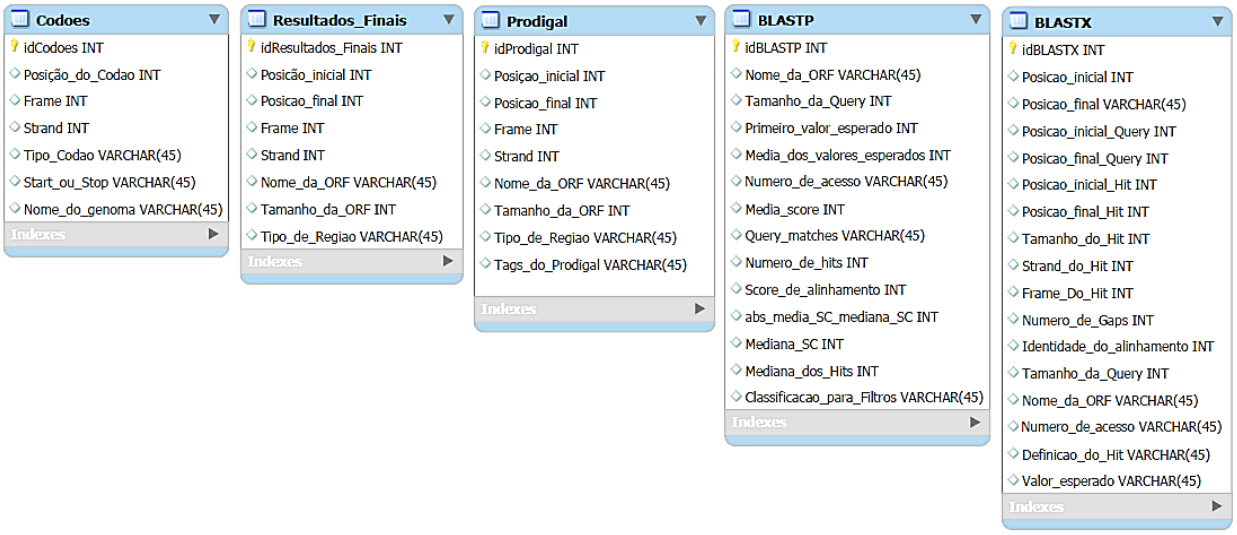
Figura 8- Resumo dos processos decorridos pelo PGP. O PGP começa por correr o *software* Prodigal sobre o qual são efetuados alinhamentos contra a base de dados do NCBI. Posteriormente, os dados resultantes do BLASTp são utilizados em filtros implementados para catalogar e corrigir caso se verifique a essa necessidade como nas previsões longas ou curtas. Nas regiões intergénicas é corrido um BLASTx por forma a encontrar ORFs não encontradas pelo *ab initio.*

3.3.1 Algoritmo e implementação do PGP

O PGP, como a Figura 8 demonstra, é um sistema que manipula diferentes fontes de informação, resultantes das várias ferramentas utilizadas. Deste modo para uma melhor manipulação de todos os dados gerados a primeira tarefa do PGP é a construção de uma base de dados na linguagem sqlite. Esta base de dados (Tabela 11) é constituída por 5 tabelas: Prodigal ORF (contém a

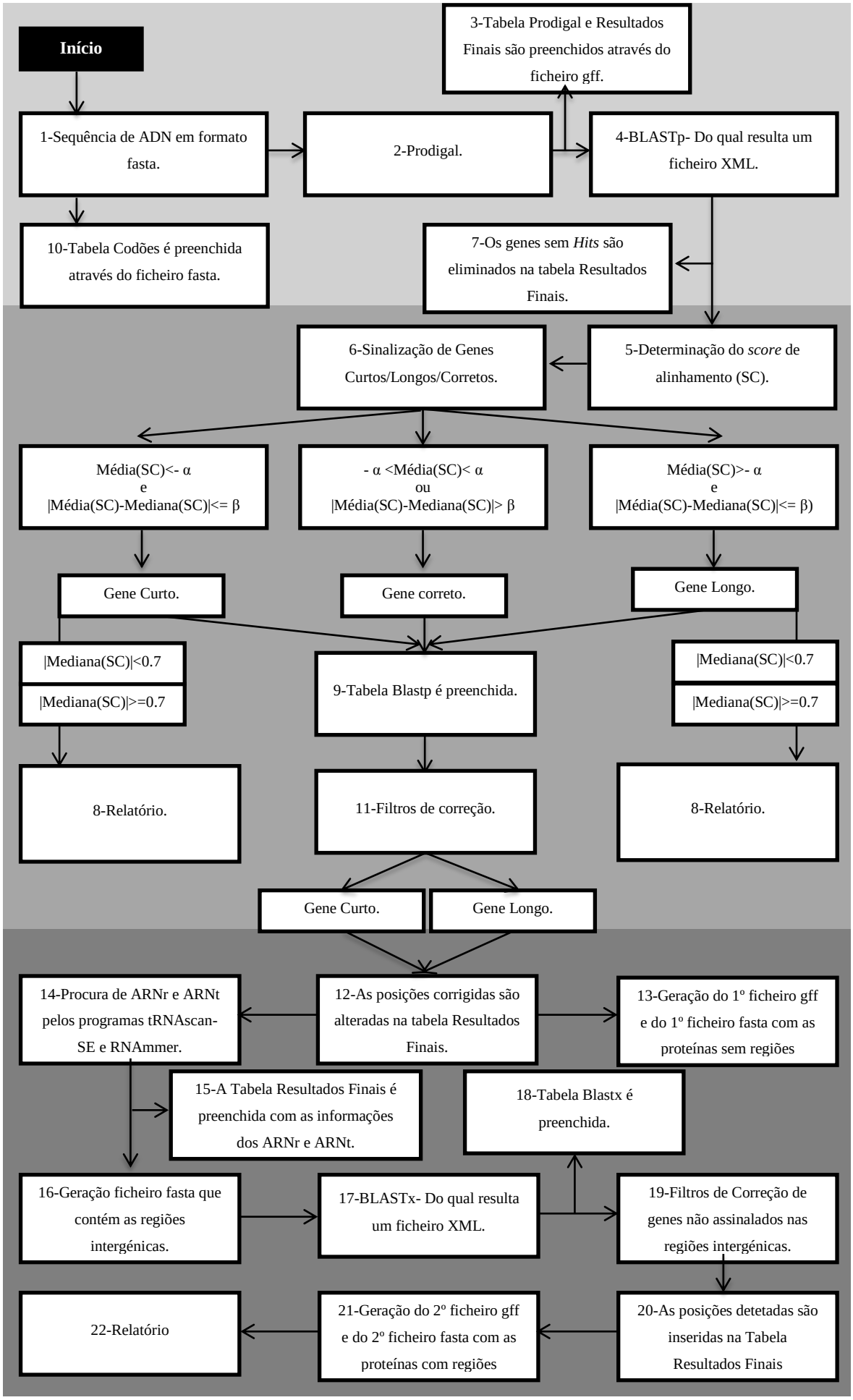
informação gerada pelo Prodigal proveniente do ficheiro gff), Resultados finais (detém a informação das previsões finais sendo atualizada no decurso do PGP), Codões (possui a informação dos locais *Start* e *Stop* de toda a sequência fasta que é dada como entrada), BLASTp (contém a informação resultante do BLASTp) e BLASTx (tem a informação resultante do BLASTx).

Tabela 11- Tabelas da base de dados criado pelo PGP. A informação gerada e necessária à implementação do *software* desenvolvido encontra-se postulada na base de dados de sqlite. Esta base de dados é gerada de cada vez que o programa é executado.



Após a construção da base de dados, o PGP começa por executar o *ab initio* Prodigal (Figura 9 ponto 2). A primeira tarefa deste previsor é a realização de um treino inicial de modo a melhorar a previsão final. Muitos previsores apenas escolhem ORFs acima de um determinado tamanho de pares de bases considerado como real. Embora esse processo seja eficaz para baixos conteúdos de G+C%, assume-se perigoso para genomas com conteúdos elevados em G+C%. Devido a falta de A (adenina) e T (timina) em genomas com altos conteúdos em G+C% existem muitos codões *Stop*, dificultando a previsão de ORFs longas e consequentemente a previsão exata de codões *Start*, gerando muitas das vezes previsões que não correspondem verdadeiramente a ORFs reais. Assim o Prodigal ultrapassa este problema com um treino preliminar, onde examina todas as ORFs no genoma olhando para o *bias* (propensão) de G ou C na 1ª, 2ª e 3ª posição de cada codão[71]. ´

Figura 9- Resumo de todos os passos dados pelo PGP, em cada uma das 3 etapas. As etapas encontram-se representadas por cores diferentes.



Seguidamente o Prodigal é executado da mesma forma abordada no capítulo 3 secção 3.1.1.1 com a inclusão do treino anteriormente realizado. O Prodigal gera 4 ficheiros de saída, 2 em formato fasta, 1 em formato de texto e outro em formato gff. É a partir do ficheiro gff que as tabela Prodigal ORF e Resultados Finais são populadas (Figura 9 ponto 3), sendo que a segunda tabela (Resultados Finais) sofrerá alterações no decorrer do PGP. Um dos ficheiros de saída em formato fasta possui as ORFs traduzidas, contendo por isso as sequências de aminoácidos previstas e é através deste ficheiro que se realiza a próxima tarefa do PGP, o BLASTp (Figura 9 ponto 4).

O BLASTp é utilizado com intuito de determinar se uma ORF prevista pelo *ab initio* é real ou não, isto com recurso às evidências existentes nas bases de dados do NCBI. Assim, este BLASTp é concretizado para os primeiros 15 Hits com valores esperados superiores a 1e-5. Este BLAST é realizado sem que ocorra mascaramento de zonas repetitivas, por forma a se obter os resultados exatos do alinhamento. No final, os resultados são reportados para um ficheiro XML (*extensible markup language*), a partir do qual vai ser extraída a informação que permite popular a tabela BLASTp e a aplicação de um algoritmo que assinala as previsões como corretas ou incorretas (este algoritmo preenche uma das colunas da tabela BLASTp) (Figura 9 ponto 9).

O algoritmo que assinala as previsões como corretas e incorretas é fundamentado no *software* GenePrimp. Este *software*, como assinalado no capítulo 2 secção 2.5.3, alerta apenas o utilizador para os genes assinalados como IN, de maneira a que este proceda manualmente à sua correção. O PGP engloba assim, parte deste algoritmo que consegue detetar genes considerados IN, mas procede seguidamente à sua correção por implementação de novos métodos mais à frente explicados.

Na Figura 10 estão expostos três casos de alinhamentos que ocorrem no BLASTp, nos quais posteriormente se procedeu ao cálculo do *score* de alinhamento (SC) (Figura 9 ponto 5) dado pela fórmula 1 do capítulo 2 secção 2.4.2, que sucintamente corresponde ao "erro" que a *Query* tem em relação aos seus *Hits*[63]. No caso I, a *Query* alinhada contra o *Hit* é demasiado curta, sendo que através do SC consegue-se verificar que a evidência gerada pelo *ab initio* pode ser alongada. O mesmo se averigua no caso III, porém nesta situação a *Query* apresenta-se demasiado longa em relação ao *Hit*, logo pelo SC verifica-se que a *Query* pode ser reduzida. Em contra partida, no caso II a *Query* faz um *match* idêntico ao *Hit*, por isso essa previsão deve ser mantida da mesma forma que é prevista[63].

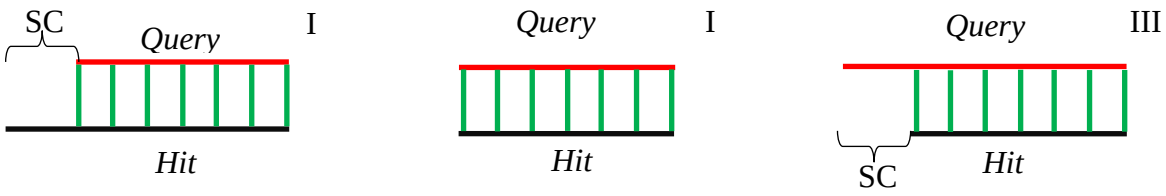


Figura 10- Aplicação do *score* de alinhamento. I- ORF prevista é curta em relação ao seu *Hit*, II- ORF com *match* idêntico ao seu *Hit*, III- ORF demasiado longa em relação ao seu *Hit*. SC corresponde ao score de alinhamento dado pela fórmula 1 do capítulo 2 secção 2.5.3.

O PGP em vez de tomar decisões baseadas só na simples diferença do local *Start* com a métrica Cq-Ch, considera o resultado dos melhores 15 primeiros *Hits* (relacionados à mesma *Query*) que possuem um valor esperado inferior a 1E-5, sobre os quais é a construída uma árvore de decisão de acordo com o cálculo da média e da mediana do SC. Com esta árvore de decisão as ORFS são sinalizadas como longas, curtas e corretas (Figura 11). Os valores da média e da mediana quanto mais próximas de 0 mais assertivamente sinalizam os genes como corretos. Contudo nem todos os genes possuem esse valor, pelo que é necessário um limite, neste caso α para a média de SC e β para o caso da diferença da média com a mediana do SC (Figura 9 ponto 6).

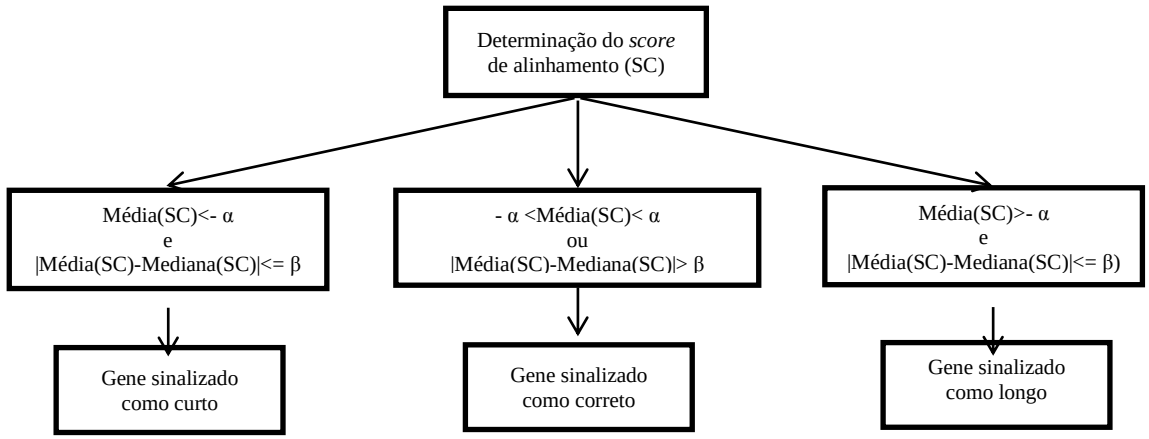


Figura 11- Árvore de decisão para a sinalização dos genes longos, curtos e corretos. O valor de α e β correspondem a limites que permitem sinalizar os genes longos, curtos e corretos com maior exatidão.

O próximo passo PGP consiste em determinar todos os possíveis codões *Start* e *Stop* da sequência fasta que é dada como entrada para a previsão das ORFs no *ab initio* (Figura 9 ponto 10). Este passo efetua a procura de 12 tripletos de nucleótidos, 6 para os codões *Start* em que 3 estão em sentido *forward* (ATG, GTG, TTG) determinando a *strand* positiva e 3 em sentido *reverse* (CAT, CAC, CAA) determinando a *strand* negativa. Na procura dos codões na *strand* negativa é preciso ter em consideração o complementar reverso, por exemplo o reverso de ATG é GTA mas o complementar reverso é CAT. Para os codões *Stop* o mesmo processo acontece, sendo necessário a procura de 3

tripletos (TAG, TAA, TGA) em *strand* positiva e 3 tripletos na *strand* negativa (TCA, TTA, CTA). Outro dos pontos essenciais na procura dos codões é a *frame* em que eles se encontram. Para o cálculo da *frame* é utilizada a seguinte fórmula:

Fórmula 6: $$Frame = Resto\ da\ divisão\ inteira(\frac{Posição\ do\ codão}{3})$$

Após este cálculo a tabela Codões é populada, e juntamente com a informação gerada até ao momento é possível a aplicação dos filtros de correção, estando estes divididos em 3 tipos:

I-Filtros de correção de ORFs longa

Este tipo de filtro faz a correção das ORFs sinalizadas como longas. Contudo, nem todas as ORFs que entram neste filtro são corrigidas, uma vez que é necessário possuírem determinadas características (Figura 9 ponto 11).

O primeiro passo deste filtro é selecionar a informação alocada na tabela Prodigal (seleção da posição inicial, posição final, *frame*, *strand*, nome do *contig* ou genoma) com correspondência na tabela BLASTp (seleção da mediana das posições *Start* dos Hits e dos genes sinalizados como longos em relação aos seus *Hits*). Uma vez realizada esta pesquisa a próxima tarefa é aceder à tabela Codões e procurar uma nova posição *Start* que satisfaça as seguintes condições:

1- Para *strands* positivas, a nova posição *Start* tem de estar compreendida entre as posições *Start* e *Stop* em que a ORF é prevista pelo Prodigal. Para *strand*s negativas, a nova posição *Start* tem de estar compreendida entre posição *Stop* e *Start* em que a ORF é prevista pelo Prodigal (Figura 12);

2- A *frame* da nova posição da tabela Codões tem de ser igual à *frame* da ORF prevista pelo Prodigal;

3- A *strand* da nova posição da tabela Codões tem se ser igual à *strand* da ORF prevista pelo Prodigal;

4- O nome do *contig*/ genoma da nova posição da tabela Codões tem de ser igual ao nome do *contig*/ genoma da ORF prevista pelo Prodigal;

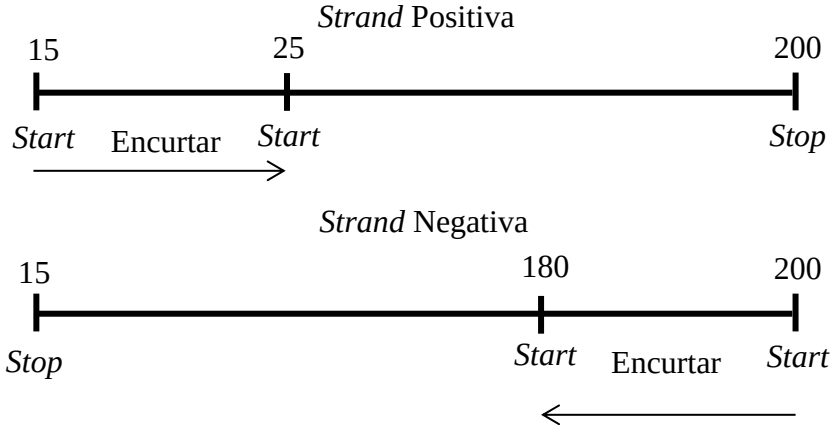


Figura 12- Determinação da nova posição *Start* por filtros de correção de ORFs curtas. Considerando dois genes com as mesmas posições mas com *strands* diferentes, procede-se à procura da nova posição com sentidos opostos. O gene inicialmente previsto em *strand* positiva compreende as posições 15 e 200, enquanto que com a correção passa a deter as posições 25 e 200. O gene inicialmente previsto em *strand* negativa compreende as posições 15 e 200, enquanto que com a correção passa a deter as posições 15 e 180.

5- A diferença entre a mediana das posições *Start* dos *Hits* com uma média local tem de ser menor ou igual a um limite (por padrão considerou-se 20 como o número máximo de aminoácidos permitidos na redução da ORF considerada longa, contudo este parâmetro é aberto ao utilizador, de maneira a que este possa escolher o parâmetro que achar mais conveniente). O valor da média local é dado pelo valor absoluto da subtração da posição *Start* com que a ORF é prevista com o Prodigal pela posição do novo codão *Start* a dividir por 3:

Fórmula 7: $$Média\ local = \frac{(|\text{Posição Inicial Prodigal} - \text{Posição do novo codão } Start|)}{3}$$

De todos os codões que possuírem as características acima demarcadas, somente o primeiro a aparecer em ordem crescente é escolhido para corrigir a ORF. Futuramente neste passo deverão ser implementados novas métricas, de modo a que se escolha entre uma gama de valores ao invés de se escolher somente o primeiro codão.

II-Filtros de correção de ORFs curtas

Este filtro é muito similar ao anterior, uma vez que a determinação da nova posição *Start* também deve obedecer ao conjunto de condições já assinaladas (Figura 9 ponto 11). No entanto, a procura não é realizada no interior da ORF prevista mas sim para além da posição *Start* da ORF inicial

sinalizada como curta. Assim, para a correção como no filtro anterior é selecionada a mesma informação da tabela Prodigal e BLASTp e para a correção a nova posição *Start,* que se encontra na tabela Codões, deve de obedecer as seguintes regras:

1- A nova posição *Start* deve de estar compreendida entre o primeiro codão *Stop* da tabela Codões (que se encontra além da ORF sinalizada como curta) e a posição *Start* (da ORF sinalizada como curta (Figura 13).

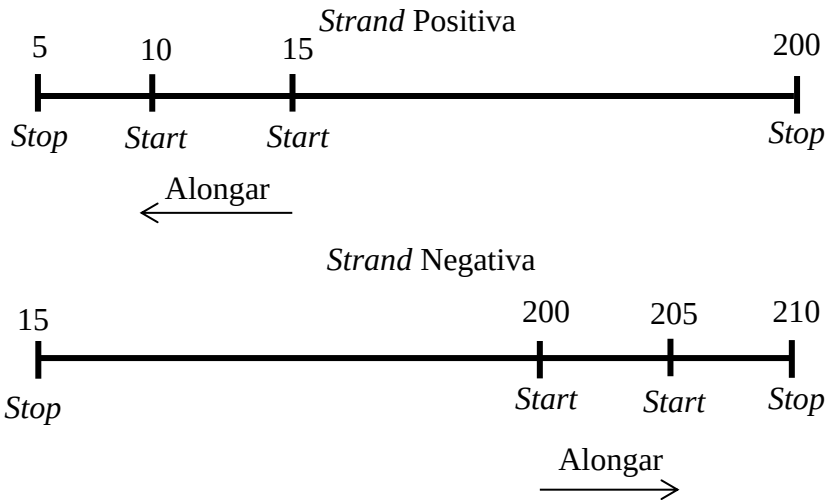


Figura 13 Determinação da nova posição *Start* por filtros de correção de ORFs longas. Considerando dois genes com as mesmas posições mas com *strands* diferentes, procede-se à procura da nova posição com sentidos opostos. O gene inicialmente previsto em *strand* positiva compreende as posições 15 e 200, enquanto que, enquanto que com a correção passa a deter as posições 10 e 200. O gene inicialmente previsto em *strand* negativa compreende as posições 15 e 200, enquanto que, enquanto que com a correção passa a deter as posições 15 e 205.

2- A *frame* da nova posição da tabela Codões tem de ser igual à *frame* da ORF prevista pelo Prodigal;

3- A *strand* da nova posição da tabela Codões tem se ser igual à *strand* da ORF prevista pelo Prodigal;

4- O nome do *contig/* genoma da nova posição da tabela Codões tem de ser igual ao nome do *contig/* genoma da ORF prevista pelo Prodigal;

Encontrados as novas posições que obedecem a estas regras, é atribuída prioridade à posição referente a um ATG, se houver mais que um posição deste tipo a ORF é alongada até à primeira posição que corresponde a um ATG. Caso não se encontre nenhum posição que referente a um ATG a posição *Start* sofre um alongamento até à posição do primeiro codão *Start* que encontrar.

No fim da aplicação dos filtros de correção das ORFs assinaladas como longas e curtas, a tabela Resultados Finais sofre uma atualização registando as correções elaboradas (Figura 9 ponto 12).

III-Regiões intergénicas

Todas as correções realizadas até ao momento são reportadas para um ficheiro de relatório (neste relatório são reportadas todas as ORFs corrigidas, no entanto as ORFs com mediana de SC inferior a 0,7[d] são assinaladas como boas correções e as restantes como correções com menor grau de certeza) (Figura 9 ponto 8), de modo a que o utilizador consiga averiguar e validar as alterações efetuadas aos resultados do *ab initio*. Para além dos resultados dos filtros, serão impressas as ORFs que para o valor esperado 1E-05 não possuem *Hits* (Figura 9 ponto 7). Essas ORFs, são eliminadas da tabela Resultados Finais criando novas regiões intergénicas. No final deste processo é criado o primeiro ficheiro gff e o primeiro ficheiro fasta com as proteínas produzidos pelo PGP (Figura 9 ponto 13).

A última fase do PGP tenta diminuir os FN com a procura de ORFs que não são previstas, realizando procuras nas regiões intergénicas (os resultados deste filtro são também reportados para o ficheiro de relatório (Figura 9 ponto 22). A primeira etapa deste filtro corresponde à procura de ARNt e ARNr através dos programas tRNAscan-SE e RNAmmer, respetivamente (Figura 9 ponto 14).

Seguidamente, o PGP cria um ficheiro fasta com todas as regiões intergénicas sobre o qual o BLASTx atuará (Figura 9 ponto 16 e 17). Este tipo de BLAST trabalha com um valor esperado (em relação ao BLASTp) mais relaxado de 1E-01 sendo os restantes parâmetros iguais. O seu *output* populará a tabela BLASTx e todos os resultados que possuírem *Hits* serão analisados para averiguar a possibilidade da existência de uma ORF na região intergénica (Figura 9 ponto 18 e 19).

Os resultados das regiões intergénicas com *Hits* saem do XML com as posições referente à *Query* e não referentes ao genoma ou *contig*. Pela Figura 14 facilmente se verifica isso, ou seja, sobre a região intergénica de 50 a 250 nucleótidos é realizado um BLASTx, mas no resultado final os *Hits* são referentes somente à *Query*, com as posições 0 a 200, sendo necessário calcular a posição da *Query* que possui *Hit* em relação ao genoma ou *contig* de estudo.

[d] 0,7-considerou-se este valor uma vez que, quanto mais próximo de 0 for o valor da mediana de SC mais certeza se têm na determinação de ORFs corretas, logo o oposto, ou seja quanto mais afastada de 0 for o valor da mediana de SC mais certeza se terá na determinação das ORFs incorretas.

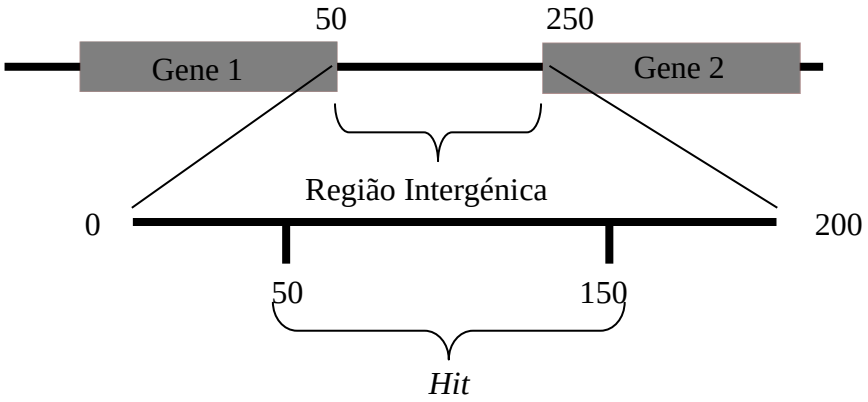


Figura 14- Determinação da ORF contida nas regiões intergénicas. A porção intergénica entre as posições 50 e 250 é avaliada por um BLASTx. Nesse BLASTx a porção intergénica é interpretada como uma *Query* fazendo a posição 50 corresponder à posição 0 nos resultados dos alinhamentos**.**

Assim, o primeiro passo das previsões de ORFs contidas nas regiões intergénicas é o cálculo das posições da *Query* em relação ao genomas/ *contig* através das seguintes fórmulas:

Fórmula 8: *Posição inicial = Posição inicial da região intergénica + Posição inicial da Query*

Fórmula 9: *Posição final = Posição inicial da região intergénica + Posição inicial da Query*

Após este cálculo para cada resultado com *Hit* é efetuada uma procura na tabela Codões de forma a encontrar codões *Start* e *Stop* que estejam nas redondezas das posições iniciais e finais com *Hits*. A procura destas regiões é limitada a um valor escolhido pelo utilizador, por forma a não ultrapassar um determinado número de nucleótidos (por padrão definiu-se o valor de 100 nucleótidos). É necessário ter em consideração que a procura dos codões deverá ter a mesma *frame* e a mesma *strand* que o resultado que possui *Hit* no BLASTx.

A determinação do codão *Start* requer o cumprimento das seguintes condições:

1- O codão *Start* tem de ser compreendido entre o limite escolhido pelo utilizador e a posição inicial dada pela fórmula 8;
2- A *frame* da posição do codão *Start* tem de possuir a mesma *frame* e *strand* entre a região prevista e codão;
3- O nome do *contig*/ genoma do codão *Start* tem de ser igual ao nome do *Contig*/ genoma da *Contig*/ Genoma que possui a região intergénica;

Nas situações em que se encontram vários codões *Start* com estas condições, o primeiro é o escolhido para determinar a posição inicial. Caso não encontre nenhum codão *Start* a iteração para a determinação desta ORF pára. Se se encontrar um codão *Start* passar-se-á para a procura de um codão *Stop*, que terá de preencher os seguintes requisitos:

1- O codão *Stop* tem de ser compreendido entre o limite escolhido pelo utilizador e a posição Final dada pela fórmula 9;
2- A *frame* da posição do codão *Stop* tem de possuir a mesma *frame* e *strand* entre a região prevista e codão;
3- O nome do *contig*/ genoma do codão *Stop* tem de ser igual ao nome do *contig*/ genoma que possui a região intergénica;

Por fim, somente as ORFs previstas nas regiões intergénicas que não apresentem codões *Stop* no meio e que tenham um tamanho do alinhamento do *Hit* superior a 90% da *Query* são demarcadas como possíveis ORFs e introduzidas na tabela Resultados Finais e é através desta tabela que é construído o ficheiro gff e ficheiro fasta com as proteinas com as previsões elaboradas pelo PGP (Figura 9 ponto 20 e 21).

3.3.2 *Métodos*

Para avaliar o algoritmo implementado foi elaborado um conjunto de 4 análises abaixo descritas (encontra-se em anexo um manual de utilização do PGP). Nas duas primeiras procurou-se saber quais os valores que possuem maior impacto nos filtros de correção, ou seja, determinar os parâmetros para os quais os filtros de correção o PGP mostra-se mais eficiente, assim como mostrar as melhorias provocadas por esses impactos em diferentes bactérias comparativamente ao Prodigal. A terceira análise correspondeu à demonstração do resultado final com inclusão das ORFs previstas nas regiões intergénicas para um conjunto de genomas modelo, demonstrando a eficácia com que o PGP pode realizar as suas previsões. Por fim na quarta análise procurou-se avaliar o comportamento do PGP contra um conjunto de *pipelines* existentes na literatura, demonstrando os benefícios do uso do PGP em relação a esses *pipelines*.

3.3.2.1 Parametrizações do Alfa e do Beta

Para garantir o melhor resultado do PGP foram efetuadas parametrizações para os filtros de correção de ORFs longas e curtas. Como evidenciado no capítulo 3 secção 3.3.1 os filtros de correção só são aplicados nos genes considerados como IN. Como a sinalização é influenciada pelos valores médios dos SC (α) e pelos valores do cálculo do valor absoluto da diferença entre a média e a mediana do *score* de alinhamento (β), foram testados diferentes valores de α e β com intuito de perceber quais os parâmetros que possuem maior impacto nos filtros de correção. Os parâmetros de α e β nesta fase do trabalho foram avaliados em 7 genomas já trabalhados na avaliação da previsão do Prodigal, GeneMark e Glimmer (capítulo 3 secção 3.2): *B. subtillis*, *E. coli*, *H. influenzae, L. lactis, P. putida*, *S. coelicolor* e *M. tuberculosis*. Foram analisados os seguintes parâmetros de α e β, respetivamente: 0,5 e 0,4; 0,5 e 0,3 (estes últimos 2 valores de α e β correspondem ao valor padrão escolhido pelo GenePrimp); 0,5 e 0,25; 0,4 e 0,4; 0,4 e 0,3 e 0,4 e 0,25. O parâmetro estatístico averiguado nesta análise foi o impacto que é dado pela fórmula 5 do capítulo 2 secção 2.6.

A execução do programa PGP contou ainda com permissão de correção de genes até um máximo limite de 20 aminoácidos. A escolha deste parâmetro deve-se ao facto do PGP necessitar de um limite para procurar os novos codões em relação as ORFs iniciais. Assim para não permitir que se alonge ou encurte demasiado as ORFs, assinaladas como incorretas, somente os novos codões que se

encontram dentro de uma distância de 20 aminoácidos da posição assinalada é que são escolhidos para efetuar a correção.

O ficheiro produzido pelo PGP para a elaboração desta análise é o primeiro ficheiro gff que ainda não contém os possíveis genes determinados nas regiões intergénicas. Nesta análise apenas se consideraram os genes codificantes que se encontravam nos genomas de referência por forma a classificá-los como FN, FP, IN e VP.

3.3.2.2 Melhorias provocadas pelo PGP em relação ao Prodigal

Conforme os parâmetros de α e β, que apresentaram maior impacto na analise do capítulo 3 secção 3.3.2.1, procurou-se verificar se estes produziam melhorias em relação ao *software* Prodigal (os resultados do Prodigal são automaticamente produzidos pelo PGP, sendo por isso descritos no capítulo 3 secção 3.3.1). Foram utilizados os genomas *B. subtillis*, *E. coli*, *H. influenzae, L. lactis, P. putida*, *S. coelicolor* e *M. tuberculosis* trabalhados no capítulo 3 secção 3.2. Para reproduzir um resultado global foi realizado o cálculo das médias para a determinação dos FN, FP,VP, taxa de erro (Fórmula 2), especificidade (Fórmula 3) e sensibilidade (Fórmula 4).

À semelhança da análise anterior apenas se consideraram os genes codificantes que se encontravam nos genomas de referência por forma a classificar os genes como FN, FP, IN e VP. Também se considerou o máximo limite de 20 aminoácidos com que o PGP executa os filtros de correção de *ORFs* longas e curtas. O ficheiro do PGP para se realizar esta abordagem foi o ficheiro gff ainda sem as regiões intergénicas.

3.3.2.3 Resultado final com a incorporação das previsões efetuadas nas regiões intergénicas

Após a verificação dos parâmetros α e β que possuíam maior impacto nos filtros de correção de possíveis ORFs longas e curtas, implementados no PGP e após a averiguação de que estes filtros traduzem uma melhoria em relação ao Prodigal, procedeu-se à avaliação de todas as previsões efetuadas, com a inclusão de ORFs previstas nas regiões intergénicas, para todos os genomas abordados no capítulo 3 secção 3.2. Deste modo os valores de α e β foram escolhidos, conforme os melhores resultados obtidos nas análises do capítulo 3 secção 3.3.2.1. Os restantes parâmetros do

PGP mantiveram-se semelhantes às análises efetuadas no capítulo 3 secções 3.3.2.1 e 3.3.2.2, sendo incluídos mais dois parâmetros que influenciam a determinação de ORFs nas regiões intergénicas. Um desses parâmetros corresponde à percentagem de alinhamento que o *Hit* deve ter em relação à *Query*, no qual foi atribuído 0,9 por forma a que apenas se escolham os *Hits* com identidades iguais ou superiores a 90% em relação à *Query*. O outro parâmetro corresponde ao número limite máximo de 100 nucleótidos que os filtros de correção das regiões intergénicas podem procurar codões *Start* e *Stop* em relação às posições dos *Hits* verificados nas regiões intergénicas.

O ficheiro de análise corresponde ao ficheiro gff com as regiões intergénicas produzido pelo PGP. Nesta análise apenas se consideraram os genes codificantes que se encontravam nos genomas de referência (ficheiro GenBank) por forma a classificar os genes FN, FP, IN e VP.

3.3.2.4 Comparação do PGP com ISGA, xBASE e Consensus predictions

Como última análise decidiu-se aprofundar o estudo, avaliando o processo decorrido através da comparação dos dados produzidos pelo PGP com os dados produzidos por Ederveen *et al.,* em 2013[65]. Neste estudo foram abordadas comparações para diferentes *pipelines* de previsão (BASys, ISGA, RAST e xBASE) de forma a criar um consenso entre previsores, ou seja, a criação de um novo método de previsão (Consensus predictions).

Utilizando os dados obtidos por Ederveen *et al.,* em 2013[65], comparou-se os *pipelines* ISGA, xBASE e Consensus predictions com o PGP. Os *pipelines* RAST e BASys estudados por Ederveen *et al.*[65] não foram incluídos nesta análise, porque estes obtiveram piores taxas de sucesso na previsão dos genomas *L. lactis* (I), *H. influenzae* (II), *S. pneumoniae* (III), *B. subtilis* (IV)*, L. plantarum* (V), *E. coli* (VI), *N. meningites* (VI), *S. entérica* (VII) (genomas trabalhados no capítulo 3 secção 3.2) em comparação com o ISGA e xBASE.

Dentro dos genomas abordados nesta análise, as previsões de alguns genomas possuíram melhor precisão para o ISGA em detrimento do xBASE. Por isso ISGA foi escolhido para as comparações das previsões nos genomas I, II, III, IV, VI e o xBASE para previsões dos genomas V, VII e VIII.

Para esta análise, foram considerados nos genomas de referência os genes codificadores e os ARNt para classificar as previsões efetuadas em FN, FP, IN e VP, indo assim ao encontro dos

resultados produzidos por Ederveen *et al.,* em 2013[65]. O PGP foi executado com os mesmos parâmetros descritos no capítulo 3 secção 3.3.2.3.

3.3.3 *Resultados e discussão*

3.3.3.1 Parametrizações do Alfa e do Beta

Com esta análise tentou-se compreender quais os valores de α e β são mais adequados para os filtros de correção de ORFs longas e curtas.

Pela Tabela 12 verifica-se que o impacto é mais constante para o parâmetro 0,5 (α) e 0,3 (β) quando comparado com os restantes, sendo classificado 2 vezes como primeiro, 2 vezes como o segundo, 1 vez como o terceiro e 2 vezes como o quarto parâmetro que possui maior impacto.

Tabela 12- Impacto das diferentes parametrizações α e β e tabela das médias do impacto. Análise em 7 genomas (I- *L. lactis*, II-*H. influenzae*, III-*B. subtilis*, IV-*E. coli* (conteúdo em moderado) e VI-*P. putida*, VII-*M. tuberculosis*, VIII-*S. coelicolor* (G+C% extremo)). Nos genomas de referência providenciados pelo ficheiro GenBank apenas se consideraram as posições marcadas com a *Tag* das regiões codificantes (*coding sequence*-CDS) por forma a realizar esta análise. Os números correspondem ao impacto produzido pela aplicação dos filtros de correção para os parametros de α e β.

Parâmetros		Impacto								Média do Impacto	
α	β	*I-L. lactis*	*II-H. influenzae*	*III-B. subtilis*	*IV-E. coli*	*V-P. putida*	*VI-M. tuberculosis*	*VII-S. coelicolor*	Total	G+C% Moderado	G+C% Extremo
0,5	0,4	0,020	0,050	0,052	0,029	0,080	0,263	0,067	0,080	0,038	0,136
0,5	0,3	0,097	0,037	0,103	0,055	0,060	0,075	0,099	0,075	0,073	0,078
0,5	0,25	0,049	0,030	0,034	0,107	0,040	0,046	0,040	0,049	0,055	0,042
0,4	0,4	0,017	0,148	0,021	0,022	0,237	0,117	0,198	0,108	0,052	0,184
0,4	0,3	0,024	0,074	0,017	0,024	0,119	0,057	0,050	0,052	0,035	0,075
0,4	0,25	0,032	0,025	0,026	0,039	0,048	0,039	0,034	0,035	0,030	0,040

Esta parametrização vai de encontro aos valores apresentados no GenePrimp. No caso deste software os valores de α e β são fechados, contudo para a sua determinação foi realizada uma distribuição para os valores da média e da mediana para o *score* de alinhamento para os genes considerados como curtos, longos e corretos demonstrando que tanto para a média como para a mediana do *score* de alinhamento os genes considerados como corretos apresentam-se entre -0,5(α) e 0,5(α), enquanto que os genes considerados curtos abaixo de -0,5(α) e os longos acima de 0,5(α)[63].

Nesta análise denotou-se ainda que para os genomas extremos o parâmetro que se apresentou melhor classificado foi o 0,4 (α) e 0,4 (β), sendo 2 vezes o primeiro classificado e 1 vez o segundo qualificado como parâmetro de maior impacto. Para os restantes parâmetros, por norma quando se diminui o α (neste caso de 0,5 para 0,4) os resultados tendem a piorar, como se pode averiguar para as parametrizações 0,4 (α) e 0,3 (β) e 0,4(α) e 0,25 (β), sendo estes parâmetros em geral os que apresentam piores resultados para os genomas em estudo.

Apesar do parâmetro 0,5 (α) e 0,3 (β) apresentar os valores mais constantes, nem sempre foi o melhor, como já referido. Por forma a comprovar este facto, procedeu-se à elaboração das médias de impacto representadas na Tabela 12, na qual constam as médias do conjunto de parâmetros (α) e (β) dos impactos totais, as médias de impactos do conjunto de parâmetros (α) e (β) para os genomas com conteúdos G+C% moderados e as médias do impacto do conjunto de parâmetros (α) e (β) para conteúdos G+C% extremos. Através da tabela 12 é possível constatar que o parâmetro 0,5 (α) e 0,3 (β) na totalidade apenas se apresenta como o terceiro classificado (com um impacto de 0,075) ocorrendo maior impacto para os parâmetros 0,5 (α) e 0,4 (β) (com um impacto de 0,080) e 0,4(α) e 0,4 (β) (com um impacto de 0,184). Verificaram-se estes resultados porque o PGP trabalha sobre os resultados previstos pelo Prodigal. Como a previsão das ORFs por parte do Prodigal ocorre com maior eficácia quando os conteúdos em G+C% são moderados, o algoritmo produzido pelo PGP não possui tanta influência. Por isso, para os genomas com conteúdo G+C% extremo, onde as previsões do Prodigal não são tão exatas, o algoritmo do PGP apresentou um maior impacto.

Analisando separadamente as médias do impacto para genomas moderados e extremos, o parâmetro 0,5 (α) e 0,3 (β) apresentou uma melhor relação para os genomas moderados (este parâmetro reduz a possibilidade de alteração às ORFs, devido a maior permeabilidade na aceitação de corretos, ou seja, pela Figura 11 que representa a árvore de decisão, todos os α com valores entre -0,5 e 0,5 ou β superiores a 0,3 são considerados como corretos). Contrariamente, quando o genoma possuiu um conteúdo em G+C% extremo, o parâmetro 0,4 (α) e 0,4 (β) apresentou os melhores resultados (este parâmetro aumenta a possibilidade de correção de ORFs, uma vez que o conjunto de ORFs consideradas como corretas tem de obedecer a uma regra mais específica, ou seja, os parâmetros 0,4 (α) e 0,4 (β) apresentam uma menor permeabilidade na aceitação de genes assinalados como corretos).

3.3.3.2 Melhorias provocadas pelo PGP em relação ao Prodigal

Pela leitura da Tabela 13 é possível verificar melhorias quando o parâmetro 0,5(α) e 0,3 (β) foi aplicado a todos os genomas. Com a implementação conjunta dos filtros de correção ao *software* Prodigal, averiguou-se que o número de FN mantém-se quando o PGP é empregue. No entanto, em termos médios o número de FP (aproximadamente 28) e IN (aproximadamente 44) foram menores, levando a um aumento do número de VP (sensivelmente 44). Por esta forma ocorreram melhorias de ≈ 1,0% em relação à taxa de erro e especificidade com a utilização do PGP, isto para o mesmo valor de sensibilidade com que as ORFs foram previstas (encontra-se em anexo a Tabela 1 com os resultados totais).

Tabela 13- Variação de valores pela aplicação do PGP comparativamente ao Prodigal. FN- falsos negativos, FP- falsos positivos, VP- verdadeiros positivos, IN- incorretos. PGP 1 referente aos parâmetros 0,5 (α) e 0,3 (β). PGP2 referente aos parâmetros 0,4 (α) e 0,4 (β). Média do Prodigal* referente às médias dos parâmetros dos genomas em estudo (*B. subtillis*, *E. coli*, *H. influenzae, L. lactis, P. putida*, *S. coelicolor* e *M. tuberculosis*). Média Prodigal ** referente aos genomas com conteúdo em G+C% normal (*B. subtillis*, *E. coli*, *H. influenzae, L. lactis*). Média Prodigal*** referente aos genomas com conteúdo em G+C % extremo *P. putida*, *S. coelicolor* e *M. tuberculosis*. Nos genomas de referência providenciados pelo ficheiro GenBank apenas se consideraram as posições marcadas com a *Tag* das regiões codificantes (*coding sequence*- CDS) por forma a realizar esta análise.

α	β	Teste	FN	FP	VP	IN	Taxa de erro	Especificidade	Sensibilidade
0,5	0,3	Prodigal*	118	187	3481	647	0,179	0,821	0,974
		PGP 1*	118	159	3525	602	0,165	0,835	0,974
0,5	0,3	Prodigal**	77	126	2822	251	0,123	0,877	0,978
		PGP 1**	77	116	2828	245	0,117	0,883	0,978
0,4	0,4	Prodigal***	174	270	4358	1175	0,254	0,746	0,969
		PGP 2***	174	216	4460	1073	0,228	0,772	0,969

Relativamente a genomas com conteúdos em G+C% extremo, como referido anteriormente, o impacto aplicado pelo PGP foi maior para os filtros de 0,4(α) e 0,4(β). Verificou-se que para estes parâmetros ocorreu uma correção média de 102 IN, fazendo aumentar os VP de 4358 (previsões efetuadas pelo Prodigal) para 4460 (previsões efetuadas com o PGP) e elevando a precisão para ≈77,0%.

Assim sendo, ocorreu uma melhoria média percentual de especificidade relativamente ao Prodigal de ≈2,6%. Em relação aos genomas com G+C% moderado, com os parâmetros 0,5(α) e 0,3(β) verificaram-se melhorias médias mais baixas. Em média, para os 4 genomas em G+C% moderado, são

corrigidos 6 ORFs perfazendo um aumento de especificidade ≈0,6%, contudo para estes genomas o Prodigal apresenta uma especificidade de ≈90,0 %.

3.3.3.3 Resultado final com a incorporação das previsões efetuadas nas regiões intergénicas

Na Tabela 14 verifica-se que o PGP em termos médios corrigiu 153 genes por genoma, com maior incidência nos genomas com conteúdos em G+C% extremo (por exemplo, verificou-se a alteração de 705 genes no genoma *S. coelicolor* com conteúdo G+C% extremo), fazendo a remoção em média de 29 genes por genoma sem valor esperado significativo.

Tabela 14- Análise do PGP com a inserção das regiões intergénicas. α- *score* de alinhamento dado pela fórmula 1. β- valor absoluto da diferença entre a média do *score* de alinhamento com a mediana do *score* de alinhamento. Alterações- corresponde às modificações efetuadas pelo PGP nos genes considerados como curtos e longos. Removidos- genes deletados devido a não possuírem um valor esperado significativo. Taxa de erro, especificidade e sensibilidade dada pelas fórmulas 2, 3 e 4, respetivamente. Nos genomas de referência providenciados pelo ficheiro GenBank apenas se consideraram as posições marcadas com a *Tag* das regiões codificantes (*coding sequence*-CDS) por forma a realizar esta análise.

α	β	Genoma	Alterações	Removidos	Intergénicas	Taxa de erro	Especificidade	Sensibilidade
0,5	0,3	*L. lactis*	43	19	0	0,099	0,901	0,988
0,5	0,3	*H. influenzae*	30	2	6	0,154	0,846	0,988
0,5	0,3	*S. pneumoniae*	33	2	8	0,211	0,789	0,916
0,5	0,3	*B. subtillis*	54	15	11	0,107	0,893	0,985
0,5	0,3	*L. Plantarum*	22	22	2	0,177	0,823	0,975
0,5	0,3	*E. coli*	105	2	36	0,112	0,889	0,999
0,5	0,3	*N. meningitidis*	66	16	18	0,225	0,775	0,914
0,5	0,3	*S. enterica*	107	12	0	0,165	0,835	0,966
0,4	0,4	*M. mobile*	24	94	0	0,891	0,109	0,591
0,4	0,4	*P. putida*	207	81	7	0,246	0,754	0,967
0,4	0,4	*M. tuberculosis*	447	38	3	0,236	0,764	0,967
0,4	0,4	*S. coelicolor*	705	41	19	0,205	0,795	0,973
Médias			153	29	9	0,236	0,763	0,935

A aplicação dos filtros de correção juntamente com a deleção de genes previstos pelo Prodigal que não possuem valor esperado significativo (inferior a $1E^{-5}$) permite a formação de novas regiões intergénicas em relação aos dados produzidos pelo Prodigal. Deste modo, nas novas regiões intergénicas formadas pelo PGP, em termos médios por genoma foram encontrados 9 novos genes. Contudo estas previsões averiguaram-se como ORFs que não se encontraram no genoma de referência, sendo assim consideradas FP. Em termos finais com a inclusão de regiões intergénicas o PGP apresentou em média uma taxa de erro de ≈24%, uma de especificidade ≈76% e uma

sensibilidade de ≈93%. Nestas análises também foram incluídos os resultados obtidos para a *M. mobile* e pelos motivos redigidos no capítulo 3 secção 3.2.2 as previsões averiguaram-se erradas. Se excluirmos os dados deste genoma, a precisão com que o PGP realiza as suas previsões os termos de taxa de erro, especificidade e sensibilidade atingem ainda melhores resultados, neste caso de ≈18%, ≈82% e ≈97%, respetivamente.

3.3.3.4 Comparação do PGP com ISGA, xBASE e Consensus predictions

O PGP para 3 genomas de estudo (II, III, VIII) apresentou o menor número de ORFs FP. Para os genomas I e V ostentou apenas uma diferença de 4 e 9 FP, respetivamente, em relação à melhor previsão de FP, neste caso realizada por parte do Consensus predictions. Para os restantes 3 genomas (IV, VI e VII) apresentou 113, 235 e 203 FP, respectivamente (as previsões de FP foram favoráveis ao *software* ISGA e xBASE com previsão de 30, 76 e 97 em relação aos mesmos genomas IV, VI e VII).

O PGP obteve ainda uma melhor previsão de genes VP, como se pode analisar na Figura 15, para os genomas II, III, IV, VI, VII e VIII, sendo a previsão dos genomas I e V favorável ao ISGA e xBASE.

Pôde-se ainda constatar que o PGP foi o *software* que apresentou menos IN, pois para os genomas I, II, III, IV, VI, VII e VIII previu menos 61 a 221 IN em detrimento do ISGA e xBASE e menos 20 a 69 IN comparativamente ao Consensus predictions. Para o genoma V o PGP apresentou piores resultados em termos de IN, existindo benefícios para a previsão efetuada pelo xBASE. Em termos de cobertura ao genoma de referência, ou seja, em termos totais (soma dos FP, IN e VP), o PGP nesta fase produziu resultados mais baixos para os genomas I, II, VI e VII.

Em anexo encontra-se a Tabela 2, que possibilitou a construção da Figura 15. Nesta tabela, para além dos dados já apresentados, também se procedeu ao cálculo da sensibilidade, especificidade e taxa de erro (Fórmulas 2, 3 e 4 abordadas anteriormente), bem como ao cálculo das suas médias e desvios padrões representado na Tabela 3, igualmente em anexo. Em relação aos dados já analisados, como esperado, o PGP apresentou estatísticas promissoras. Pelo cálculo das médias conseguiu-se perceber que o PGP apresentou uma melhoria na ordem de ≈3,1 para o valor de taxa de erro e especificidade relativamente às previsões levadas a cabo pelo Consensus prediction.

Após uma comparação com as melhores previsões produzidas com o xBASE e ISGA relativamente aos resultados do PGP, os últimos apresentaram melhores valores, obtendo-se uma percentagem de ganho de ≈4,4% referente à taxa de erro e especificidade. Contudo, em termos de

sensibilidade, o PGP foi um pouco superior (apresentou uma sensibilidade de 95,9±3%), uma vez que os valores médios foram melhores quando os *pipelines* ISGA e xBASE foram utilizados (neste caso com 95±3,6%).

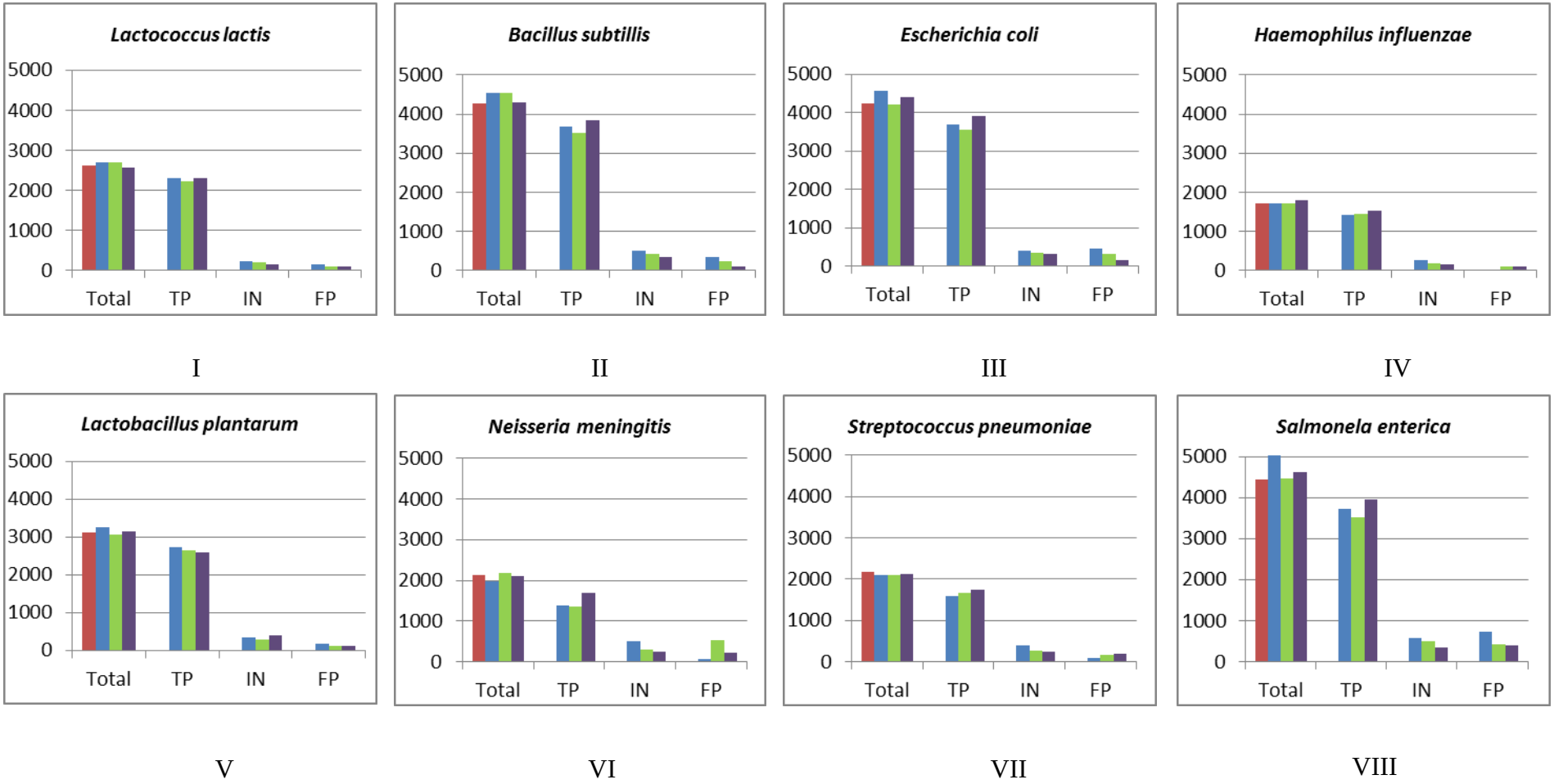


Figura 15- Variação de previsão de ORFs entre os *pipelines* ISGA ou xBASE, Consensus predictions e PGP para 8 genomas com conteúdo em G+C % moderado. ORFs anotadas do genoma de referência (Total de proteínas codificantes e genes ARNt), ORFs previstas pelo ISGA (I,II,III,IV,VI) ou xBASE (V, VII e VIII), ORFs previstas pelo Consensus predictions, ORFs previstas pelo PGP. Total corresponde ao numero total de previsões realizadas pelos *pipelines*, TP verdadeiros positivos, IN corresponde aos incorrectos, FP corresponde aos falsos negativos. Nos genomas de referência providenciados pelo ficheiro GenBank consideraram-se as posições marcadas com as *Tags* das regiões codificantes (*coding sequence*-CDS) e tARN por forma a realizar esta análise.

CAPÍTULO 4

4.1 CONCLUSÕES

4.1.1 Síntese geral do trabalho

O trabalho aqui apresentado teve como principal motivação desenvolver uma ferramenta de previsão. Deste modo, foram desenvolvidos estudos por forma a encontrar as maiores carências nos previsores de maior utilização na comunidade científica. Em suma o trabalho elaborado potencia a automatização da anotação genética tornando-a cada vez mais eficaz. Revisitando os objetivos formulados o trabalho pretendeu essencialmente:

1) Selecionar o previsor ab initio que atualmente oferece melhor capacidade de previsão;
2) Identificar os tipos e características de genes que são previstos de forma incorreta;
3) Desenvolver um *software* baseado em homologia de sequências para a correção de genes previstos incorretamente;
4) Desenvolver uma *pipeline* que englobe todas as ferramentas necessárias para a previsão de genes.

Em termos teóricos entendeu-se que os genes são os objetos de estudo da previsão, sendo que existem dois tipos de genes, os eucariotas e os procariotas. Quanto aos processos de anotação de genes estes compreendem sistemas estatísticos e de procura por homologias. Um dos principais processos de anotação consiste na previsão de genes, objeto principal do trabalho empírico concretizado. O processo de previsão compreende quer métodos intrínsecos como extrínsecos, assentes em ferramentas e *softwares* especializados como, por exemplo, o Glimmer, o GeneMark e o Prodigal.

O trabalho prático desenvolveu-se em torno da previsão de genes procariotas recorrendo-se a métodos de *ab initio* e homologia. Neste sentido, apresenta-se de seguida as conclusões inerentes aos objetivos formulados.

4.1.2 Conclusões do trabalho

Este trabalho proporcionou o desenvolvimento de uma nova ferramenta de previsão de genes por incorporação de metodologias de procura por *ab initio* e homologia. Inicialmente foram realizadas diferentes análises de modo a avaliar o comportamento de diferentes previsores *ab initio*. Concluiu-se que o Prodigal, por norma, tende a ser o mais eficaz para os vários campos de análise. Em termos médios, para 12 genomas modelo, o Prodigal apresentou uma sensibilidade de ≈93,3%, uma taxa de erro de ≈23,9% e uma especificidade de ≈76,1%; valores estes superiores quando comparados com os obtidos pelo GeneMark (≈35,3% de taxa de erro, ≈64,7% de especificidade e ≈92,3% de sensibilidade) e com o Glimmer (≈58,1% de taxa de erro, ≈41,9% de especificidade e ≈66,9% de sensibilidade).

A partir das previsões elaboradas pelo Prodigal foi desenvolvido uma primeira abordagem por homologia, com vista ao seu melhoramento, implementando-se um algoritmo que corrige e valida as previsões automaticamente. Sucintamente, o algoritmo de correção calcula um *score* de alinhamento (cálculo baseado nos *Hits* obtidos pela aplicação da ferramenta BLASTp) sobre o qual é possível determinar se um dado gene previsto é curto, longo ou correto relativamente à sua posição *Start*.

Esta metodologia apresentou resultados positivos, provando-se que com a introdução da correção automática de genes é possível obter melhores previsões. Deste modo, pelas análises efetuadas, averiguou-se que o PGP possui um impacto maior quando o conteúdo em G+C% é extremo, apresentando em média uma taxa de erro de ≈22,8%, mais baixo cerca de ≈2,6% que a taxa de erro do Prodigal (≈25,4%), tendo obtido o mesmo valor de sensibilidade (≈96,9%) entre PGP e Prodigal. Relativamente aos conteúdos G+C% moderados, o impacto é relativamente inferior ao impacto analisado para genomas de G+C% extremo. Contudo, ainda assim, o PGP apresentou resultados significativos de melhoramento em relação ao Prodigal, apresentando para 4 genomas em conteúdo em G+C% moderado uma taxa de erro de ≈16,5%, significativamente inferior ao do Prodigal (≈17,9%), e com a mesma sensibilidade (≈97,8%).

Outro teste foi elaborado para avaliar o comportamento do PGP comparativamente a outros *pipelines* de anotação como ISGA, xBASE e Consensus Predictions, neste caso considerando 8 genomas de conteúdo em G+C% moderado. Relativamente ao ISGA e xBASE o PGP mostrou um ganho de ≈4,4% de taxa de erro, ou seja, o PGP possuí uma taxa de erro médio de ≈15,8% e o ISGA e xBASE de ≈20,2%. Em média a sensibilidade do PGP correspondeu ≈95,9% enquanto o ISGA e xBASE a ≈95,1%.

Comparando o PGP com o Consensus Predictions, o PGP mostrou-se novamente mais assertivo, apresentando valores médios de ganho relativamente à taxa de erro de ≈3,1% e de ≈5,4% para a sensibilidade (Consensus Predictions em média previu os genes com uma taxa de erro de ≈18,9% e uma sensibilidade de ≈90,5%).

A segunda abordagem do algoritmo implementado passou pela análise das regiões intergénicas, uma vez que um dos problemas averiguados no *ab initio* passa pela não deteção de genes (em média o Prodigal tende a não prever 137 genes por genoma). Contudo, pelo algoritmo implementado a abordagem criada não se mostrou muito significativa, uma vez que introduz em média 9 falsos positivos por genoma, não conseguindo ainda descobrir os genes verdadeiros.

Encerra-se as conclusões com um balanço positivo de todo o trabalho desenvolvido, pois foi possível estudar e comparar diferentes ferramentas de previsão, decidindo-se sobre qual apresenta maior número de vantagens para a anotação de genes procariotas. Conclui-se que de todas, o PGP, embora careça de melhorias, foi o que apresentou melhores resultados e menos erros. Considera-se por isso que os benefícios aqui encontrados são significativos para a comunidade científica e para a Bioinformática.

4.2 Recomendações para Trabalho Futuro

No decurso deste trabalho desenvolveu-se uma ferramenta de previsão e correção de genes, tendo-se obtido resultados bastante satisfatórios e promissores, como concluído anteriormente. No entanto, alguns aspetos poderão ser aprimorados de forma a melhorar o desempenho futuro do PGP.

Um dos problemas verificados no PGP encontra-se na sinalização de genes que deveriam ser considerados como incorretos. Isto é, apesar dar abordagens mostrarem benefícios na identificação e posterior correção, futuros novos métodos deveram ser encontrados, de modo a que se consiga englobar mais genes com posições erradas, tornando assim o PGP ainda mais eficiente.

Outro dos possíveis trabalhos futuros poderá passar pela implementação de estratégias mais fortes na deteção de genes nas regiões intergénicas. A implementação criada para este tipo de problemas não surtiu efeito, surgindo assim a real necessidade de implementação de novas estratégias que suplementem este problema.

O trabalho efetuado consistiu no desenvolvimento de novas estratégias de previsão de genes, não se tendo concluído todo o processo de anotação. Embora o processo mais difícil esteja concluído

com as previsões, faltam ainda as caracterizações funcionais dos genes previstos, com dados anotados em diferentes bases de dados, pelo que, futuramente deverão ser implementadas plataformas que acedam automaticamente às bases de dados e atribuam significado aos genes previstos.

Essencialmente, o trabalho desenvolvido foi a primeira abordagem ao estudo do genoma bacteriano com a descoberta de "pistas" para a sua caracterização.

"Aventure-se, pois da mais insignificante pista, surgiu toda a riqueza que o homem já conheceu." John Masefield (1878-1967) So long to learn – 1952.

REFERÊNCIAS

1. Thomson, R. C., Wang, I. J. & Johnson, J. R (2010). Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular ecology* 19, 2184–2195.
2. Schneider, M. V. *et al* (2010). Bioinformatics training: a review of challenges, actions and support requirements. *Briefings in bioinformatics* 11, 544–551.
3. Fleischmann, R. D. *et al* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)* 269, 496–512.
4. Poptsova, M. S. & Gogarten, J. P. (2010) Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology (Reading, England)* 156, 1909–1917.
5. Margulies, M. *et al.* (2006). Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature biotechnology* 437, 376–380.
6. Shendure, J. *et al.* (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* 309, 1728–1732.
7. Bentley, D. R. *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59.
8. Korlach, J., Officer, C. S. & Biosciences, P. Understanding Accuracy in SMRT Sequencing. at < http://www.pacificbiosciences.com/pdf/Perspective_UnderstandingAccuracySMRTSequencing.pdf>. Acedido em junho de 2013.
9. Sanger, F., Nicklen, S. (1977). DNA sequencing with chain-terminating. 74, 5463–5467.
10. Bräutigam, A., Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant biology (Stuttgart, Germany)* 12, 831–832.
11. Egan, A. N., Schlueter, J., Spooner, D. M. (2012).Applications of next-generation sequencing in plant biology. *American journal of botany* 99, 175–85.
12. Quail, M. a *et al.* (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* 13, 341.
13. Kircher, M., Kelso, J. (2010). High-throughput DNA sequencing-concepts and limitations. *BioEssays: news and reviews in molecular, cellular and developmental biology* 32, 524–36.
14. Misawa, K. (2013). RF: A method for filtering short reads with tandem repeats for genome mapping. *Genomics* 102, 35–37.
15. Koren, S. *et al.* (2013). Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome biology* 14, R101.
16. Kahvejian, A., Quackenbush, J., Thompson, J. F. (2008). What would you do if you could sequence everything? *Nature biotechnology* 26, 1125–33.
17. Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends in biotechnology* 26, 602–11.
18. Angelova, M., Kalajdziski, S., Kocarev, L. (2010). Computational Methods for Gene Finding in Prokaryotes. *ICT Innovations 2010 Web Proceedings ISSN* 11–20.
19. Richardson, E. J., Watson, M. (2012). The automatic annotation of bacterial genomes. *Briefings in bioinformatics* 14, 1–12.
20. Alberts, Bruce; Bray, Dennis *et al*. (2007). *Fundamentos da Biologia Celular*. Artmed.
21. Rimoin, D.L., Connor, J.M., Pyeritz, R.E. and Korf, B. R. (2013).*Emery & Rimoin's Principles and Practice of Medical Genetics*. Churchill Livingstone.

22. Venter, J. C. *et al* (2001). The Sequence of the Human Genome. *Science (New York, N.Y.)* 291, 1304-1350.
23. Frewer, L. J. *et al.* (2013). Public perceptions of agri-food applications of genetic modification – A systematic review and meta-analysis. *Trends in Food Science & Technology* 30, 142–152.
24. Hernández, M. L. O., Salinas, E. S., González, E. D., Godínez, M. L. C. (2013). Pesticide Biod,degradation: Mechanisms, Genetics and Strategies to Enhance the Process, Biodegradation. *Life of Science*.
25. Pafford, B. W., Petti, C. (2013). Diagnostic medical home: a model for health and well-being. *Archives of pathology & laboratory medicine* 137, 884–885.
26. Misra, S. (2013). Human Gene Therapy: A Brief Overview of the Genetic Revolution. *Journal of the Association of Physicians of India* 61, 127-133.
27. Fontinha Vieira, C. S. (2007). Estudo de Variáveis Discretas: um contributo ao Ensino e à Genética. Tese de mestrado. Departamento de Matemática. Universidade de Aveiro.
28. Hartl, D. L.; Jones E. W. (2006). *Essential Genetics: A Genomics Perspective*. Jones & Bartlett Learning.
29. Regateiro, F. J. (2003). *Manual de Genética Médica*. Imprensa da Universidade de Coimbra.
30. R.Blattner, F; Plunket, Guy *et al*. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science* 277, 1453-1462.
31. Griffiths, A. J.F., Lewontin, R. C., Carroll, S.B., Wessler, S. R., William D. F. (2008). *Introduction to Genetic Analysis*. W. H. Freeman.
32. Akhtar, M., Al., E. (2008). Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. *IEEE Journal of Selected Topics in Signal Processing* 2, 310–321.
33. Felder, Y. (2007). Analysis of Biological Networks: Transcriptional Networks - Promoter Sequence Analysis. 1–15. <at http://www.cs.tau.ac.il/~roded/courses/bnet-a06/lec11.pdf>. Acedido em março de 2013.
34. Srebrow, A., & Kornblihtt, A. R. (2006).The connection between splicing and cancer. *Journal of Cell Science*, 119, 2635-2641.
35. Cenik, C., Derti, A., Mellor, J. C., Berriz, G. F. & Roth, F. P. (2010). Genome-wide functional analysis of human 5' untranslated region introns. *Genome biology* 11, R29.
36. Castellana, N., Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: a computational perspective. *Journal of proteomics* 73, 2124–2135.
37. Yus, E. *et al.* (2012). Transcription start site associated RNAs in bacteria. *Molecular systems biology* 8, 585.
38. Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America* 72, 784–788.
39. Uemura, S. *et al.* (2007). Peptide bond formation destabilizes Shine-Dalgarno interaction on the ribosome. *Nature* 446, 454–457.
40. Lukashin, a V & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic acids research* 26, 1107–1115.
41. Stein, L. (2001). Reviews genome annotation: from sequence to biology. *Nature reviews* 2, 493–503.
42. Richardson, E. J., Watson, M. (2013). The automatic annotation of bacterial genomes. *Briefings in bioinformatics* 14, 1–12.
43. Stothard, P., Wishart, D. S. (2006). Automated bacterial genome analysis and annotation. *Current opinion in microbiology* 9, 505–510.

44. Besemer, J., Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research* 33, W451–454.
45. Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics* 5, 59.
46. Hyatt, D. *et al.* (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11, 119.
47. Boratyn, G. M. *et al.* (2013). BLAST: a more efficient report with usability improvements. *Nucleic acids research* 41, W29–33.
48. Pertsemlidis, a & Fondon, J. W. (2001). Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome biology* 2.
49. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC bioinformatics* 7, 62.
50. Badger, J. H. & Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Molecular biology and evolution* 16, 512–524.
51. Maji, S. & Garg, D. (2013). Progress in Gene Prediction: Principles and Challenges. *Current Bioinformatics* 8, 226–243.
52. Delcher, A. L., Bratke, K. a, Powers, E. C. & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)* 23, 673–679.
53. Stanke, M. & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* 33, W465–467.
54. Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* 25, 955–964.
55. Lagesen, K. *et al.* (2007). RNAmmer : consistent and rapid annotation of ribosomal RNA genes. 35, 3100–3108.
56. Mavromatis, K. *et al.* (2009). The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *Standards in genomic sciences* 1, 63–67.
57. Rutherford, K. *et al.* (2000). Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)* 16, 944–945.
58. Cantarel, B. L. *et al.* (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* 18, 188–196.
59. Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K. V & Dong, Q. (2010). An Ergatis-based prokaryotic genome annotation web server. *Bioinformatics (Oxford, England)* 26, 1122–1124.
60. Aziz, R. K. *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics* 9, 75.
61. Chaudhuri, R. R. & Pallen, M. J. (2006). xBASE, a collection of online databases for bacterial comparative genomics. *Nucleic acids research* 34, D335–7.
62. Van Domselaar, G. H. *et al.* (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic acids research* 33, W455–9.
63. Pati, A. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nature Methods* 7, 1–6.
64. frameshift mutation / frame-shift mutation; frameshift. at <http://www.nature.com/scitable/definition/frameshift-mutation-frame-shift-mutation-frameshift-203>. Acedido em junho de 2013.
65. Ederveen, T. H. a., Overmars, L., Van Hijum, S. a. F. T (2013). Reduce Manual Curation by Combining Gene Predictions from Multiple Annotation Engines, a Case Study of Start Codon Prediction. *PLoS ONE* 8, e63523.

66. Snipen, L.-G. & Ussery, D. W. (2012). A domain sequence approach to pangenomics: applications to Escherichia coli. *F1000Research* 19, 1–17.
67. Acinetobacter baumannii ABNIH20 contig00011, whole genome shotgun sequ - Nucleotide - NCBI. at <http://www.ncbi.nlm.nih.gov/nuccore/APBI01000011.1>. Acedido em março de 2013.
68. Escherichia coli O157:H43 str. T22 contig25, whole genome shotgun sequ - Nucleotide - NCBI. at <http://www.ncbi.nlm.nih.gov/nuccore/NZ_AHZD02000025.1>. Acedido em março de 2013.
69. Audic, S. & Claverie, J. M. (1998). Self-identification of protein-coding regions in microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America* 95, 10026–10031.
70. Besemer, J., Lomsadze, a & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research* 29, 2607–2618.
71. Prodigal: Microbial Gene Prediction Algorithm Description. at http://prodigal.ornl.gov/algorithm.html>. Acedido em julho de 2013.

ANEXOS

MANUAL DE UTILIZAÇÃO PGP

Este programa é um sistema hibrido que faz a previsão de genes. Começa por executar o *ab initio* Prodigal e sobre este tenta validar e ou corrigir a informação gerada. A validação e correção ocorre por métodos de homologia, nos quais se realizam dois tipos de BLAST. Os BLAST geram informação suficiente para a sinalização de genes considerados incorretos e para uma posterior correção dos genes assinalados como incorretos. Os parâmetros estatísticos que são utilizados no PGP para a correção de ORFs, encontram-se em aberto por forma a dar liberdade de escolha ao utilizador sobre os parâmetros que achar mais conveniente aos seus processos de previsão. Os parâmetros são os seguintes:

-a Média do SC permitida para a sinalização das ORFs incorretas (valor de α);
-b Número de aminoácidos máximos permitidos para a correção dos genes assinalados como incorretos (Valor padrão 20 aminoácidos);
-c Número de Cpus utilizados para correr o programa;
-d Nome da base de dados;
-e Nome do ficheiro de treino do Prodigal;
-f Nome do ficheiro de potenciais genes em format txt previstos pelo Prodigal;
-g Nome do ficheiro de relatório que assinala todas as alterações provocadas pelo PGP;
-h Menu inicial (ajuda);
-i Ficheiro fasta de entrada com a sequência de ADN do genoma/ *contig*;
-j Nome do ficheiro fasta de Proteínas criado pelo Prodigal;
-k Nome do ficheiro fasta de nucleótidos criado pelo Prodigal;
-l Nome do ficheiro gff com as previsões iniciais criado pelo Prodigal;
-m Valor absoluto da diferença entre a média e a mediana do *score* de alinhamento permitido para a sinalização das ORFs incorretas (valor de β);
-n Número de nucleótidos que se utilizam para procurar codões *Start* e *Stop* a partir das regiões intergénicas com *Hit* (por padrão 100 nucleótidos)*;*
-o Nome do ficheiro gff sem as regiões intergénicas criado pelo PGP;
-p Utilização da base de dados filtrada do NCBI (T-Verdadeiro, F-Falso);

-q Nome do ficheiro gff com as regiões intergénicas criado pelo PGP;

-r Nome para o ficheiro XML produzido pelo blastP;

-s Nome da pasta onde se guardaram os resultados finais;

-t Tamanho do alinhamento do *Hit* em relação à *Query* referente às regiões intergénicas (por padrão 0,9);

-v Nome do ficheiro fasta de nucleótidos das regiões intergénicas;

-x ficheiro com os tRNA;

-y ficheiro com os rRNA.

Opções extra:

-w Entrada do ficheiro XML por forma a que não seja executado o BLASTp;

-z Entrada do ficheiro XML por forma a que não seja executado o BLASTx.

Exemplo de uma execução:

run.pl –a [valor_de_α] –b [20] –c [16] –d [nome_DB] –e [ficheiro_treino_prodigal] –f [Genes_potencias_Prodigal] –g [Nome_do_relatório] –h [ajuda] -i [Ficheiro_fasta_inicial] –j [[Nome_fasta_proteínas_prodigal] –k [Nome_fasta_nucleótidos_prodigal] –l [Nome_gff_prodigal] –m [valor_de_β] –n [100] –o [PGP_sem_intergenicas.gff] –p [F] –q [PGP_com_intergenicas.gff] –r [xml_blatp] –s [Nome_Pasta] –t[0.9] –v[fasta_intergénica] –x[ficherio_tRNA] –y [ficherio_rRNA] –w[xml_blastp] –z [xml_blastx]

Tabela anexo 1-Variação de valores pela aplicação do PGP comparativamente com o Prodigal. FN- falsos negativos, FP- falsos positivos, VP verdadeiros positivos, IN-incorretos. PGP 1 referente aos parâmetros 0,5 (α) e 0,3 (β). PGP2 referente aos parâmetros 0,4 (α) e 0,4 (β). Nos genomas de referência apenas se consideraram as posições marcadas como regiões codificantes por forma a realizar esta análise.

Genomas	*Software*	FN	FP	VP	IN	Taxa de erro	Especificidade	Sensibilidade
B. subtillis	Prodigal	63	108	3753	360	0,112	0,888	0,985
	PGP 1	63	92	3765	348	0,105	0,895	0,985
E. coli	Prodigal	176	166	3827	318	0,112	0,888	0,959
	PGP 1	176	164	3818	327	0,114	0,886	0,959
H. influenzae	Prodigal	20	109	1474	163	0,156	0,844	0,988
	PGP 1	20	107	1480	157	0,151	0,849	0,988
L. lactis	Prodigal	47	119	2235	162	0,112	0,888	0,981
	PGP 1	47	100	2250	147	0,099	0,901	0,981
P. putida	Prodigal	175	409	4107	1068	0,265	0,736	0,967
	PGP 1	175	327	4143	1032	0,247	0,753	0,967
S. coelicolor	Prodigal	213	186	5995	1560	0,226	0,774	0,973
	PGP 1	213	144	6131	1424	0,204	0,796	0,973
M. tuberculosis	Prodigal	134	215	2973	896	0,272	0,728	0,967
	PGP 1	134	176	3089	780	0,236	0,764	0,967
P. putida	Prodigal	175	409	4107	1068	0,265	0,736	0,967
	PGP 2	175	327	4153	1022	0,245	0,755	0,967
S. coelicolor	Prodigal	213	186	5995	1560	0,226	0,774	0,973
	PGP 2	213	144	6135	1420	0,203	0,797	0,973
M. tuberculosis	Prodigal	134	215	2973	896	0,272	0,728	0,967
	PGP 2	134	176	3092	777	0,236	0,764	0,967

Tabela anexo 2- Variação de previsão de ORFs entre os *pipelines* ISGA ou xBASE, Consensus predictions e PGP para 8 genomas com conteúdo em G+C % moderado. O número total de ORFs de referência é dado pelo número de regiões codificantes de proteína mais os genes de ARNt. VP- verdadeiros positivos, IN- incorretos, FP-falsos positivos.

	ORFs de referência	*Software*	Total	VP	IN	FP	Taxa de erro	Especificidade	Sensibilidade
L. plantarum	3128	ISGA	3267	2728	350	189	0,172	0,828	0,984
		C. Predictions	3076	2650	293	133	0,136	0,864	0,941
		PGP	3142	2586	416	142	0,178	0,822	0,960
L. lactis	2605	ISGA	2691	2313	24	138	0,062	0,938	0,897
		C. Predictions	2691	2221	198	96	0,113	0,887	0,929
		PGP	2563	2316	147	100	0,095	0,905	0,945
B. subtilis	4262	ISGA	4540	3691	494	355	0,199	0,801	0,982
		C. Predictions	4540	3519	417	223	0,150	0,850	0,924
		PAGe	4290	3850	348	103	0,106	0,894	0,985
E. coli	4235	ISGA	4572	3692	410	470	0,208	0,792	0,969
		C. Predictions	4215	3556	347	312	0,156	0,844	0,922
		PGP	4394	3903	327	201	0,125	0,875	0,999
S. pneumoniae	2163	xBASE	2098	1597	404	97	0,232	0,768	0,925
		C. Predictions	2106	1676	260	170	0,199	0,801	0,895
		PGP	2114	1731	245	203	0,207	0,793	0,914
S. enterica	4448	ISGA	5038	3732	573	733	0,294	0,706	0,968
		C. Predictions	4455	3516	512	427	0,211	0,789	0,906
		PGP	4625	3942	357	404	0,171	0,829	0,967
N. meningitis	2122	xBASE	1979	1390	513	76	0,278	0,722	0,897
		C. Predictions	2191	1372	292	527	0,386	0,614	0,784
		PGP	2103	1703	241	235	0,224	0,776	0,916
H. influenzae	1715	xBASE	1721	1429	262	30	0,170	0,830	0,986
		C. Predictions	1709	1436	179	94	0,159	0,841	0,942
		PGP	1802	1538	157	113	0,157	0,843	0,988

Tabela anexo 3- Variação média das previsão de ORFs entre os *pipelines* ISGA ou xBASE, Consensus predictions e PGP para 8 genomas com conteúdo em G+C % moderado.

Fórmulas	*Software*	Média	Desvio padrão
Taxa de erro	ISGA-xBASE	0,202	0,068
	C. predictions	0,189	0,080
	PGP	0,158	0,044
Especificidade	ISGA-xBASE	0,798	0,068
	C. predictions	0,811	0,080
	PGP	0,842	0,044
Sensibilidade	ISGA-xBASE	0,951	0,036
	C. predictions	0,905	0,048
	PGP	0,959	0,030