Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Análise do pangenoma de *Streptococcus pneumoniae* e comparação de genomas dos serótipos 1 e 3

Adriana Domingos Policarpo

Dissertação orientada por:

Prof. Doutor João Carriço
Prof. Doutor Francisco Pinto

Mestrado em Bioinformática e Biologia Computacional

Especialização em Bioinformática

2015

Agradecimentos

Gostaria de agradecer primeiramente aos meus orientadores, Professor Doutor João Carriço, e Professor Doutor Francisco Pinto, por toda a disponibilidade demonstrada, todo o apoio e aconselhamento.

A toda a Unidade de Microbiologia Molecular e Infeção do Instituto de Medicina Molecular, onde passei grande parte do tempo no decorrer da realização deste trabalho, especialmente ao Professor Doutor Mário Ramirez, por permitir que integrasse a sua unidade no âmbito deste projeto. A todos os colegas, que durante este tempo foram uma companhia e ajuda indispensável, por todos os momentos bem passados na sua companhia.

Agradeço também aos meus pais e às minhas irmãs, por todo o apoio e sacrifícios, que permitiram que eu pudesse chegar até aqui.

A todos os meus colegas e amigos, que sempre me apoiaram e aconselharam, e estiveram disponíveis em todos os momentos. Ao Nuno, por todo o apoio, carinho e paciência incondicionais.

A todos muito obrigada!

Resumo

Streptococcus pneumoniae é uma espécie bacteriana que coloniza a nasofaringe humana, sendo a principal causa de diversas doenças, como infeção respiratória aguda e otite média. Várias estirpes desta espécie apresentam uma cápsula polissacarídica, apresentando diversas variantes composicionais que correspondem a diferentes serótipos, os quais apresentam diferente potencial patogénico. Neste estudo pretende-se analisar o pangenoma – o reportório total de genes de uma espécie microbiana, que poderá ser significativamente maior que o número de genes encontrados em cada uma das estirpes individualmente – de *S. pneumoniae*, que compreende o genoma *core* – conjunto de genes presentes em todas as estirpes – e o genoma acessório – conjunto de genes presentes em duas ou mais estirpes e genes únicos.

Com o desenvolvimento das tecnologias de sequenciação tornou-se fundamental o desenvolvimento de novas ferramentas bioinformáticas para lidar com as grandes quantidades de informação geradas, surgindo a necessidade de efetuar estudos genómicos comparativos a larga escala para tentar extrair informação útil desses dados. Assim, desenvolveu-se neste estudo uma ferramenta bioinformática, denominada SCRAG (Strict CoRe and Accessory Genome) que permite a comparação de vários genomas em simultâneo, obtendo o genoma *core* e acessório. Esta ferramenta foi então utilizada para a análise do genoma de *S. pneumoniae*. O SCRAG tem por base do processo de comparação de sequências o algoritmo BLAST, cujos resultados são depois filtrados por vários parâmetros, dos quais o utilizador pode definir a percentagem de identidade e a percentagem de diferença de tamanho máxima permitida entre sequências de um conjunto de alelos que codificam para um mesmo *locus*. Os resultados obtidos com esta ferramenta são conservadores pois removem possíveis genes parálogos presentes nos genomas e os parâmetros de identidade e diferença de tamanho são determinados de modo a obter elevada confiança nos resultados obtidos.

Utilizaram-se 27 genomas de vários serótipos completamente sequenciados e anotados disponíveis no GenBank e 49 genomas sequenciados pela Unidade de Microbiologia Molecular e Infeção. Estes 49 genomas continham

24 estirpes do serótipo 1 e 25 estirpes do serótipo 3. A escolha da análise destes serótipos prende-se com o facto de serem causadores de doença invasiva em diferentes grupos etários e a sua caracterização genómica ser muito diferente.

Obtiveram-se os resultados para um conjunto de 25 dos 27 genomas disponíveis no GenBank, para os quais estavam disponíveis os ficheiros contendo as regiões codificantes. Obtiveram-se também os resultados para o total dos 76 genomas de *S. pneumoniae*. Foram utilizados diferentes parâmetros de percentagem de identidade e de diferença de tamanho, sendo que para 80% de identidade e 20% de diferença de tamanho se obtém 619 genes *core* e 873 genes acessórios para o conjunto de 25 genomas e 226 genes *core* e 977 genes acessórios, para o conjunto de 76 genomas. No entanto, o número total de genes descobertos não aumenta com o número de genomas analisados, o que será devido ao método utilizado, que se revela bastante estrito na filtragem dos resultados do BLAST.

Para a comparação dos serótipos 1 e 3 utilizou-se também o SCRAG, tendo-se posteriormente comparado os conjuntos de resultados obtidos. Utilizando genes *core*, verificou-se que existem mais genes partilhados entre o serótipo 3 e o grupo de outros serótipos, ao passo que o serótipo 1 parece divergir bastante dos restantes, sendo também o que apresenta menos genes no total, o que era expectável uma vez que apresenta limitada diversidade genética. Já considerando genes acessórios, o maior número de genes partilhado ocorre entre os serótipos 1 e 3, continuando o serótipo 1 a divergir bastante do grupo "outros serótipos".

Futuramente, será importante analisar os dados obtidos com o SCRAG em termos funcionais, para melhor compreender a espécie bacteriana estudada.

Palavras-chave: *Streptococcus pneumoniae*, pangenoma, serótipos, genoma *core*, genoma acessório, BLAST

Abstract

Streptococcus pneumoniae is a bacterial species that colonizes the human nasopharynx and it's the main cause of several diseases, like acute respiratory infection and otitis media. Several strains of this species have a polysaccharide capsule, presenting several compositional variants corresponding to different serotypes, which have different pathogenic potential. The aim of this study is to analyze the pangenome – the total repertoire of genes of a microbial species, which could be significantly larger than the number of genes found in each strain individually – of *Streptococcus pneumoniae*. The pangenome comprises the core genome – the set of genes present in all strains – and the accessory genome – the set of genes present in two or more strains and the unique genes.

With the development of sequencing technologies has become essential the development of new bioinformatics tools to handle the large amounts of information generated, resulting in the need to perform comparative genomic studies on a large scale to try to extract useful information from these data. Thus we developed in this study a bioinformatic tool, called SCRAG (Strict CoRe and Accessory Genome), allowing the comparison of several genomes simultaneously, obtaining the core and accessory genome. This tool was used to analyze the genome of *S. pneumoniae*. SCRAG is based on the sequence comparison process using the BLAST algorithm, whose results are then filtered by various parameters of which the user can define the percentage of identity and the percentage of maximum size difference allowed between sequences of a set of alleles encoding the same *locus*. The results obtained with this tool are conservative because they remove possible paralogous genes present in the genomes and identity and size difference parameters are determined in order to achieve high confidence in the results obtained.

27 genomes of several serotypes, completely sequenced and annotated and available in GenBank and 49 genomes sequenced by the Molecular Microbiology and Infection Unit were used. These 49 genomes contained 24 serotype 1 strains and 25 serotype 3 strains. The choice of the analysis of these serotypes is related with the fact that they cause invasive disease in different

age groups and their genomic characterization is very different.

There were obtained results from a set of 25 of 27 genomes available at GenBank, which have the files containing the coding regions available. The results for the set of all 76 genomes of *S. pneumoniae* were also obtained. Different parameters of percentage of identity and size difference were used. With 80% identity and 20% size difference were obtained 619 core genes and 873 accessory genes for the set of 25 genomes and 226 core genes and 977 accessory genes for the set of all 76 genomes. However, the total number of discovered genes does not increase with the number of analyzed genomes, which could be due the method used, which proved quite strict when filtering the BLAST results. For serotype 1 and 3 comparison SCRAG was also used and the sets of results obtained were compared. Using core genes, it was found that there are more genes shared between serotype 3 and the group of other serotypes, whereas serotype 1 appears to deviate widely from the other, being also the one with fewer genes in total, which was expected since it has limited genetic diversity. Using accessory genes, the highest number of genes shared occurs between serotypes 1 and 3, continuing the serotype 1 to diverge rather from the group "other serotypes". In the future it will be important to analyze the data obtained with the SCRAG in functional terms to better understand the bacterial species studied.

Key words: *Streptococcus pneumoniae*, pangenome, serotypes, core genome, accessory genome, BLAST

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Enquadramento e objectivos

Streptococcus pneumoniae é uma espécie bacteriana que coloniza a nasofaringe humana, identificada como o seu reservatório natural [1], mas que é também a principal causa de infeção respiratória aguda e otite média [2], além de outras doenças menos graves como sinusite e bronquite [3]. É possível proceder à serotipagem de *S. pneumoniae*, ou seja, à determinação dos antigénios de superfície usando um conjunto definido de anticorpos monoclonais ou policlonais [4]. Esta espécie bacteriana causa doença sobretudo em crianças e em idosos [1], sendo que diferentes serótipos apresentam diferente potencial patogénico e diferente distribuição geográfica [5]. Neste estudo, pretende-se analisar o genoma de *Streptococcus pneumoniae*, tentando perceber que genes são partilhados ou não pelos genomas de estirpes diferentes. Uma estirpe diz respeito a uma população ou grupo de populações de células microbianas que derivam de uma única colónia e que apresentam traços genotípicos e/ou fenotípicos comuns, que os permitem distinguir de outros isolados da mesma espécie [4]. Para alcançar os objectivos propostos, foram utilizados 27 genomas de *S. pneumoniae* completos e anotados disponíveis no GenBank e 49 genomas sequenciados pela Unidade de Microbiologia Molecular e Infeção, de estirpes pertencentes a vários serótipos, nomeadamente um conjunto de dados do serótipo 1, um do serótipo 3, e outro conjunto de dados de outros serótipos que não o 1 ou o 3.

Com o desenvolvimento das tecnologias de sequenciação e crescente acessibilidade às mesmas, que resultou num maior uso, e consequentemente a obtenção de cada vez mais dados, tornou-se fundamental o desenvolvimento de novas ferramentas informáticas para lidar com estas grandes quantidades de informação [6]. Surge assim a necessidade de efetuar estudos genómicos com-

parativos a uma larga escala, para tentar extrair informação útil do grande volume de dados disponível [7]. Assim, este estudo visa simultaneamente a construção de uma ferramenta bioinformática que permita a comparação de vários genomas em simultâneo, e a análise do genoma de *S. pneumoniae*, como referido.

1.2 Sequenciação de genomas de alto débito, sua evolução e utilização

A sequenciação de genomas de alto débito ("*High Throughput Sequencing*") tem-se mostrado a ferramenta ideal para revelar diferenças genómicas entre estirpes.[7] A crescente acessibilidade a estas tecnologias, devido sobretudo à queda do preço de sequenciação, permitiu que genomas completos ("*Whole Genome Sequencing*" - WGS) de múltiplas estirpes de variadas espécies bacterianas de interesse clínico, tenham vindo a ser gerados nos últimos anos. Lidar com estes dados genómicos também tem vindo a tornar-se mais fácil, uma vez que têm sido desenvolvidos *software*, bases de dados e algoritmos de análise, permitindo assim a análise de centenas de genomas.[6]

Por outro lado, os estudos genómicos computacionais dependem grandemente da qualidade das anotações dos genomas disponíveis. Algoritmos desenvolvidos para anotação podem beneficiar grandemente com o número de genomas relevantes disponíveis para comparação. Idealmente, as anotações devem ser directamente comparáveis, de modo a possibilitar a análise de elevados números de genomas. No entanto, as anotações genómicas depositadas nas bases de dados da "*International Nucleotide Sequence Database Collaboration*" (INSDC) variam no nível de detalhe, escolha de termos e linguagem e no tipo exacto de ("features") reportado. Seria portanto importante o desenvolvimento de novas bases de dados, que combinem e uniformizem informação de uma variedade de fontes, e apliquem técnicas de reanotação uniformes. No entanto, a uniformização *in silico* não será suficiente, sendo necessário melhorar as anotações empiricamente. Por este motivo, foi criado o "*Genomic Standards Consortium*" (GSC), um grupo internacional que trabalha na criação de um conjunto mais rico de descrições de genomas completos e metagenomas.[7]

Os estudos genómicos comparativos de larga escala permitem detectar entidades e padrões biológicos a nível da organização genómica – como fusão de genes, pseudogenes, RNA não-codificante, bem como estudar genes órfãos e específicos de linhagem – que explicam fenómenos particulares ou excepções de interesse relevantes. [7]

A análise comparativa pode ainda ser estendida para explorar e caracterizar qualquer padrão amplamente partilhado entre micróbios – como a distribuição de uma variedade de características estruturais, a caracterização global dos proteomas, a abundância de sequências repetitivas ou a abundância relativa de tipos específicos de genes – e leva à melhor compreensão da função e evolução de genomas. Pode também levar a resultados práticos, como a aplicação em engenharia de processos de protecção das bactéria usadas em bioprocessos industriais (como a fermentação) contra fagos, o que resulta da compreensão da interacção entre fagos e bactérias. Os estudos comparativos podem ainda ser usados para procurar relações entre características genómicas e ecologia e reconstruir relações evolucionárias entre genomas.[7]

Análises mais aprofundadas entre múltiplos genomas de espécies individuais permitem mesmo explorar o conceito de espécie bacteriana, pois revelam grande diversidade intraespecífica, o que terá implicações em áreas de investigação em Saúde Humana, tais como o desenvolvimento de vacinas ou o estudo do desenvolvimento e propagação intraespécies e interespécies da resistência a antibióticos.[6]

1.3 Pangenoma, genoma *core* e genoma acessório

Numa espécie bacteriana, a "pool" ou reportório de genes indentificados em estirpes distintas dessas espécies aumenta com o número de genomas analisados [7]. Mesmo após a sequenciação de genomas de várias estirpes, em alguns casos, novos genes serão adicionados ao genoma da espécie a cada nova sequência genómica, uma vez que ainda não tinham sido identificados em outros genomas já sequenciados. Modelação matemática prevê que, para algumas espécies, sejam descobertos mais genes mesmo após sequenciação de centenas de genomas. [8] Assim, na teoria, as espécies bacterianas nunca estão completamente descritas, embora fosse uma mais-valia saber quantos genomas são necessários para representar com precisão o reportório de genes de uma espécie [8, 6]. Os fenómenos de mutação e recombinação, sendo essenciais para a evolução e diversividade de uma espécie, vão também contibuir para a capacidade de uma dada espécie adquirir ou gerar novos genes. Espécies clonais – que evoluem sobretudo por mutação – terão assim menor probabilidade de gerar novos genes do que espécies panmíticas – que normalmente apresentam taxas de recombinação elevadas e capacidade de acquisição de DNA de outras espécies. [9, 5]

Assim surge o conceito de "pangenoma" (do grego "pan", que significa "todo") ou supragenoma: o reportório de genes de uma espécie microbiana é significativamente maior que o número de genes encontrados em cada uma das

estirpes individualmente, e uma percentagem significativa de cada genoma é, portanto, específica de cada estirpe individual [6, 7, 8]

Consideramos então que uma espécie pode ser descrita pelo seu pangenoma, que é composto pelo genoma "*core*"e pelo genoma acessório. O genoma *core* contém genes presentes em todas as estirpes, ao passo que o genoma acessório, também denominado por genoma dispensável, contém os genes presentes em duas ou mais estirpes, bem como os genes únicos (encontrados numa única estirpe) [8, 6]. Existe um vasto número de genes únicos, pelo que o pangenoma de uma espécie bacteriana pode ser muito maior do que um genoma de uma única estirpe.[8] Uma possibilidade de explicação para a existência de genomas acessórios tão vastos será a necessidade de adaptação a nichos ecológicos distintos [7]. Os genes "*core*" e "acessórios" representam a essência e diversidade das espécies, respectivamente. [8, 6]

1.3.1 Pangenoma aberto e fechado

O pangenoma de uma espécie bacteriana pode ser aberto ou fechado. Um pangenoma aberto será aquele que aumenta a cada estirpe sequenciada, apresentando grande diversidade genética. Estudos anteriores apontam para 20% a 35% de genes específicos de estirpes únicas, em média. Este tipo de pangenoma é típico de espécies que colonizam múltiplos ambientes e têm múltiplas vias de trocar material genético. *Streptococcus pneumoniae*, *Streptococcus agalactiae* (*Streptococcus* do grupo B – GBS), *Streptococcus pyogenes*, *Staphylococcus aureus*, *Neisseria meningitidis* (*Meningococci*), *Helicobacter pylori*, *Sallmonellae* e *Escherichia coli* são exemplos de espécies bacterianas que apresentam estas características, observando-se nelas um pangenoma aberto. [8] Um pangenoma fechado considera-se aquele em que o número de genes específicos adicionados por cada genoma tende a convergir para zero ao fim de poucos genomas. É típico de espécies que vivem isoladas, com acesso limitado ao reportório de genes microbiano global, sendo mais conservadas, com menos capacidade de adquirir genes provenientes de outras espécies [8]. Quando ocorrem menos eventos de recombinação – como é o caso – isto será mais provável de ocorrer, como referido. Um exemplo é o *Bacillus anthracis*, em que apenas quatro genomas são suficientes para caracterizar a espécie, de acordo com estudos prévios, sendo este um dos mais extremos pangenomas fechados. [8, 6] Outros exemplos são os genomas de *Mycobacterium tuberculosis* e *Chlamydia trachomatis*. Um exemplo ainda mais extremo é o caso do *Buchnera aphidicola*, cujo genoma não sofreu alterações nos últimos 50 milhões de anos, demonstrando a mais extrema estabilidade genómica observada até à data. [8]

Em alguns casos, espécies com pangenoma fechado e aberto tendem a ser

muito semelhantes – por exemplo, *B. anthracis* (pangenoma fechado) e *B. cereus* (pangenoma aberto) – parecendo clones e não espécies verdadeiramente independentes, e a principal característica que os distingue é a aquisição de factores de virulência – no exemplo dado, dois plasmídeos, um dos quais codifica para a toxina do atrax. A classificação do *B. anthracis* como espécie independente é então, geneticamente, apenas um traço fenotípico codificado pelo genoma acessório. Os critérios usados para definir uma espécie microbiana podem assim ser inconsistentes com a informação genética. [8]

De acordo com estudos anteriores de análise do pangenoma de uma espécie, a lei de Heaps é aplicável como modelo para pangenomas abertos. A lei de Heaps determina que o número n de atributos distintos cresce de acordo com uma lei de potências ("power law") sub-linear do número N de entidades consideradas, e à medida que a amostragem continua, descobrir novos atributos torna-se mais difícil. Ou seja, no caso do pangenoma, temos que o número de genes (atributos) aumenta com o número de genomas (entidades), sendo que à medida que são considerados mais e mais genomas, o número de novos genes descobertos diminui. De acordo com esta "power law", se o tamanho de o pangenoma tende para uma constante quantos mais forem os genomas considerados, então trata-se de um pangenoma fechado. Se o tamanho do pangenoma é uma função que aumenta, tendendo para o infinito com o número de genomas considerados – ou seja, o tamanho do pangenoma segue a lei de Heaps – trata-se de um pangenoma aberto. Se o tamanho do pangenoma segue tendência logarítmica, ou seja, cresce muito lentamente, tecnicamente também é infinito, considerando-se que o pangenoma é aberto. [6]

1.3.2 Caracterização dos genes *core*, genes acessórios e genes do pangenoma

O genoma *core* inclui todos os genes responsáveis pelos aspectos básicos da biologia de uma espécie e os seus traços fenotípicos mais importantes. Já o genoma acessório é constituído pelos genes que contribuem para a diversidade das espécies, e podem codificar vias bioquímicas e funções suplementares que não são essenciais para o crescimento bacteriano, mas que conferem vantagens selectivas, como a adaptação a diferentes nichos ecológicos, resistência a antibióticos ou a capacidade de colonizar novos hospedeiros [8, 6]. São geralmente genes agrupados em grandes ilhas genómicas, tipicamente flanqueados por repetições curtas de ADN, e caracterizados por um conteúdo G+C anormal. Investigação e anotação funcional dos genes acessórios revela que os genes hipotéticos, de origem fágica e de transposões, contam para a

vasta maioria das descobertas, enquanto que num genoma típico este tipo de genes representam percentagens muito mais pequenas. O facto de estarem maioritariamente associados com um número limitado de estirpes indica uma fraca selecção positiva para estas funções, e mostra que os elementos móveis contribuem pobremente para o *fitness* geral e diferenciação da espécie, apesar de, por vezes, poderem conter genes importantes. Dado que estes genes não são essenciais para a sobrevivência e manutenção das espécies, podem eventualmente ser perdidos do genoma de uma estirpe. No entanto, nas espécies patogénicas, esta perda é por vezes acompanhada paralelamente por uma redução da virulência. Assim, a necessidade de sequenciar múltiplos genomas de cada espécie para melhor compreender a diversidade das espécies bacterianas não é apenas teórica. [8]

O pangenoma de uma espécie pode ser continuamente trocados dentro de uma dada espécie ou entre espécies por três processos principais: transformação (material genético retirado do ambiente), transdução (ADN é transferido por um vírus) e conjugação (ADN directamente trocado entre células bacterianas), sendo que nos casos de transformação e conjugação o organismo fonte e o organismo alvo vivem em estreito contacto. Este vasto reportório genético sugere então que, durante a evolução, a grande maioria das novas funções foi gerada no mundo microbiano, e não nos grandes animais, que apresentam um reportório mais pequeno. Micróbios e grandes animais terão assim papéis totalmente diferentes na evolução, sugerindo que os micróbios geram novos genes e módulos funcionais, ao passo que os grandes animais rearranjam os módulos provenientes das bactérias de muitas maneiras diferentes, dentro do próprio genoma, e por *splicing* alternativo de RNA mensageiro (mRNA). [8]

Os métodos clássicos para catalogar espécies bacterianas são baseados no conhecimento de traços fenotípicos convenientes, e assumem que todas as estirpes do mesmo serótipo são similares. Mas técnicas mais recentes, como MLST (*MultiLocus Sequence Typing*), baseadas na detecção da variabilidade associada com os genes *housekeeping*, levam à classificação das estirpes em complexos clonais (CC) e tipo de sequência (ST), respectivamente. Contudo, a comparação das sequências dos genomas completos das estirpes de GBS mostram que a diversidade genómica poderá não estar relacionada com os serótipos ou STs MLST. Por vezes, isolados provenientes de diferentes serótipos são mais estreitamente relacionados do que isolados do mesmo serogrupo, e estirpes do mesmo ST podem ser geneticamente muito distantes. A razão para poder existir esta ausência de correlação entre serótipos e diversidade genética reside provavelmente no facto de genes de especificidade capsular estarem presentes no genoma acessório que é trocado livremente entre estirpes com diferentes *backgrounds* genéticos. Por contraste, os ge-

nes usados para determinar o tipo MLST pertencem ao genoma *core*, e não apresentam semelhanças presentes no genoma acessório, que está frequentemente ligado a características patogénicas. Isto demonstra não haver total congruência entre os métodos tradicionais, como serotipagem e MLST,e a caraterização genómica de genomas completos [8]

1.4 Comparação de genomas

1.4.1 Comparação de sequências

Quando se descrevem comparações de sequências é frequente usar vários termos diferentes: identidade, similaridade e homologia. Estes três termos, apesar de por vezes serem utilizados indistintamente, têm diferentes significados. [10]

Identidade de sequências refere-se à ocorrência de exactamente o mesmo nucleótido ou aminoácido na mesma posição, em sequências alinhadas. Similaridade ou semelhança de sequências tem em conta as correspondências aproximadas, e é significativa apenas quando tais substituições são pontuadas de acordo com alguma medida de "diferença"ou de "igualdade", com substituições de alta probabilidade ou conservativas a obterem pontuações mais elevadas do que substituições não-conservativas ou improváveis. O termo "homologia de sequências" é, por sua vez, o mais importante dos três. Quando dizemos que duas sequências têm elevada homologia, estamos a afirmar não só que as duas sequências parecem a mesma, como que os seus ancestrais também pareciam o mesmo. A segunda afirmação é mais difícil de confirmar. Apesar de por vezes a comparação de duas sequências ser sumarizada como uma percentagem de homologia de sequências, este uso é geralmente incorreto, uma vez que o valor indica a identidade ou similaridade, não refletindo necessariamente uma relação evolucionária. Algoritmos de comparação de sequências, como BLAST e FASTA – que aplicam algoritmos heurísticos para procurar numa base de dados de sequências as correspondências mais aproximadas a uma sequência de interrogação – e SSEARCH – que procede a um alinhamento local completo para cada par de sequências por um método de programação dinâmico – não medem a homologia das sequências, mas sim a similaridade e identidade. Inferências acerca da homologia apenas podem ser fornecidas pelo utilizador. [10]

É importante saber o quão similares são duas sequências pois isto permite-nos atribuir informação conhecida sobre uma sequência a outras sequências: a Natureza resolveu os mesmos casos muitas vezes, por vezes com significativa semelhança entre as soluções.[10]

1.4.2 Alinhamentos e matrizes de substituição

Antes de poder ser calculada computacionalmente a similaridade de duas sequências, é necessário determinar o seu alinhamento mais conveniente. No entanto, avaliar um determinado alinhamento envolve o cálculo de similaridade – trata-se portanto de um "problema circular" [10, 11]. Para determinar o quão semelhantes são duas sequências é necessário avaliar fatores como se é ou não uma correspondência perfeita, qual o melhor alinhamento, como pontuar os alinhamentos, ou como devem ser pontuados os intervalos (*gaps*), se forem permitidos. Torna-se assim necessário ter uma forma de pontuar as correspondências ou não-correspondências, bem como um método de usar ambos para avaliar os numerosos alinhamentos possíveis. [10]

Aquando da avaliação de um alinhamento de sequências, para saber o quão significativo ele é, requer-se uma matriz de pontuação: uma tabela de valores que descrevem a probabilidade de um aminoácido ou nucleótido biologicamente significativo ocorrer num alinhamento. Tipicamente, quando duas sequências de nucleótidos são comparadas, o que é pontuado é se duas bases numa dada posição são ou não a mesma. A todas as correspondências é atribuída a mesma pontuação (tipicamente +1 ou +5), assim como a todas as não-correspondências (tipicamente -1 ou -4) [10, 11]. No entanto, com proteínas, a situação é diferente. As matrizes de substituição para aminoácidos são mais complicadas e implicitamente têm em conta tudo o que possa afetar a frequência com que qualquer aminoácido é substituído por outro, tal como a natureza química e a frequência de ocorrência dos aminoácidos. O objetivo é providenciar uma penalidade relativamente pesada por alinhar dois resíduos que têm uma baixa probabilidade de serem homólogos. Existem dois fatores principais que fazem com que as taxas de substituição de aminoácidos se distanciem da uniformidade: nem todas as substituições ocorrem com a mesma frequência, e algumas substituições são menos toleradas funcionalmente do que outras. [10]

Entre as matrizes de substituição mais utilizadas incluem-se a matriz de substituição de blocos (BLOSUM – *BLOcks of Amino Acid SUbstitution Matrix*) e matriz de mutações pontuais aceites (PAM – *Point Accepted Mutation*). Ambas são baseadas na utilização de conjuntos de alinhamentos de alta confiança de muitas proteínas homólogas e na avaliação das frequências de todas as substituições, mas são computadas utilizando métodos diferentes. [10]

1.4.2.1 PAM

As matrizes PAM são calculadas com base num modelo de distância evolucionária a partir de alinhamentos de sequências estreitamente relacionadas (com pelo menos 85% de identidade) de 34 super-famílias agrupadas em 71 árvores evolucionárias e contendo 1572 mudanças, ou mutações pontuais. O limiar de similaridade foi escolhido para minimizar tanto erros nos alinhamentos como mutações coincidentes. Para determinar a sequência ancestral para para cada alinhamento, foram reconstruídas árvores filogenéticas para essas sequências. As substituições foram calculadas por tipo, normalizadas para frequências de uso e convertidas para *log odd scores*. A matriz resultante foi chamada M1 ou PAM1 e define a unidade de mudança evolucionária. Portanto, os valores na matriz M1 representam a probabilidade de um aminoácido em 100 vir a sofrer uma substituição. Multiplicando a matriz PAM1 por si própria geram-se matrizes de pontuação para graus de relação arbitrários. Multiplicando-a por si própria n vezes dá uma matriz de pontuação para proteínas que tenham sofrido n mutações múltiplas e independentes. A matriz PAM120 é considerada uma boa matriz de pontuação para sequências estreitamente relacionadas, enquanto que a PAM250 é mais apropriada para sequências mais distantemente relacionadas. A multiplicação infelizmente também multiplica o erro associado a cada estimativa de probabilidade de substituição de aminoácidos, significando que as matrizes PAM de ordem superior são mais propensas a erros. [10]

1.4.2.2 BLOSUM

As matrizes BLOSUM foram construídas de maneira similar às matrizes PAM, mas a partir de sequências selecionadas para evitar sequências altamente relacionadas que ocorrem frequentemente. Essa informação é derivada da base de dados BLOCKS, que consiste num conjunto de alinhamentos sem *gaps* de sequências provenientes de famílias de proteínas relacionadas. Utilizando cerca de 2000 blocos de segmentos de sequências alinhados, caracterizando mais de 500 grupos de proteínas relacionadas, as sequências em cada bloco são organizadas em grupos (*clusters*) estreitamente relacionados, e as frequências de substituição entre esses *clusters* dentro de uma família são utilizadas para calcular a probabilidade de uma substituição significativa. O número associado a uma matriz BLOSUM (como BLOSUM62 ou BLOSUM80), indica o valor limite (*cutoff*) da percentagem de identidade das sequências que definem os *clusters*. Um *cutoff* mais baixo permite assim mais diversidade de sequências nos *clusters*, e as matrizes correspondentes são apropriadas para avaliar relações mais distantes. [10]

1.4.3 BLAST

Com o aumento da quantidade de dados genómicos gerados nos últimos anos, foi necessária a criação de ferramentas de alinhamento e procura de similaridade, para explorar efetivamente esses dados para investigação médica e biológica [12]. A procura de semelhanças, incluindo a comparação de sequências, é uma das principais técnicas usadas pelos biólogos computacionais e amplamente utilizada entre os biólogos no geral, sendo a principal forma pela qual a bioinformática contribui para o nosso entendimento da biologia [10]. Várias ferramentas foram geradas para este propósito nas últimas décadas, incluindo BLAST, FASTA, sim4 ou BLAT [10], tendo vindo a ser extremamente úteis, sobretudo na área da genómica comparativa, em que genomas tanto de espécies estreitamente relacionadas como de espécies geneticamente distantes são comparados, uma vez que o conhecimento do genoma de uma espécie pode ser usado para compreender o genoma de outras espécies [12]. Das várias ferramentas desenvolvidas, o BLAST (*Basic Local Alignment Search Tool*) é a mais amplamente utilizada, e o seu uso tornou-se fundamental na biologia [10].

Estas ferramentas funcionam, de uma forma geral, com base na identificação de uma lista de segmentos de uma sequência genómica alvo (*target*) numa base de dados, que mostre semelhanças com a sequência de interrogação (*query*) [12]. O BLAST, por exemplo, procede a comparações entre estes pares de sequências, procurando por regiões de similaridade local. Existem várias implementações deste algoritmo, sendo o NCBI BLAST e o WU-BLAST as que adquiriram um uso mais vasto. O NCBI BLAST está disponível através do Centro Nacional para Informação Biotecnológica – NCBI (*Nacional Center for Biotechnology Information*), e o WU-BLAST pela Universidade Washington, em Saint Louis. Neste estudo foi utilizado o NCBI BLAST, em cujos princípios e aplicações nos iremos focar [10].

O BLAST apresenta vários programas, que se encontram listados e descritos na tabela 1.1 [10].

No BLAST, cada correspondência entre um fragmento da sequência de interrogação e um fragmento da sequência alvo é reportado como um par de alta pontuação (HSP – "*high-scoring segment pair*"), que consiste num par de segmentos do mesmo tamanho (Q, T), onde Q é um segmento da sequência de interrogação ou *query* e T é o segmento correspondente de uma sequência alvo ou *target*. A pontuação de similaridade para um par de segmentos alinhados é a soma dos valores de similaridade para cada par de resíduos alinhados. O par de segmentos com a pontuação mais alta é chamado "par de segmentos de pontuação máxima" (*maximal-scoring segment pair* – MSP) e o seu alinhamento não pode ser melhorado por extensão ou encurtamento Uma

Programa	**Tipo de sequência interrogatória (“query”)**	**Tipo de sequência alvo (“target”)**	**Descrição**
BLASTP	Proteína	Proteína	Compara uma sequência de aminoácidos (“query”) contra uma base de dados de sequências de proteínas.
BLASTN	Nucleótido	Nucleótido	Compara uma sequência de nucleótidos (“query”) contra uma base de dados de sequências de nucleótidos.
BLASTX	Nucleótido (traduzido)	Proteína	Compara uma sequência de nucleótidos traduzida em todas as janelas de leitura (“query”) contra uma base de dados de sequências de proteínas.
TBLASTN	Proteína	Nucleótido (traduzido)	Compara uma sequência de aminoácidos (“query”) contra uma base de dados de sequências de nucleótidos dinamicamente traduzidas em todas as janelas de leitura.
TBLASTX	Nucleótido (traduzido)	Nucleótido (traduzido)	Compara uma sequência de nucleótidos traduzida com base numa janela de 6 nucleótidos (“query”) contra uma base de dados de sequências de nucleótidos traduzidas com base numa janela de 6 nucleótidos.

Tabela 1.1: Programas baseados no BLAST [10]

pesquisa usando BLAST pode retornar vários HSPs para uma sequência de interrogação no genoma alvo, sugerindo a existência de um ou mais genes homólogos nesse genoma (ou base de dados de nucleótidos), geralmente correspondendo cada HSP com um exão. [10, 11, 12] O BLAST atribui a cada HSP uma pontuação (*bit score*), um valor esperado (*E-value*) e valores de identidade e similaridade. Quando são reportados vários HSPs estes podem ser todos eles únicos, com um *E-value* e identidade correspondente, sendo que alguns deles podem representar genes candidatos e providenciar um ponto de partida significativo para pesquisas adicionais, enquanto que outros são *hits* aleatórios [12].

Geralmente, o BLAST é a ferramenta de eleição, não só pela sua melhor precisão, como devido à sua disponibilidade e vasta aceitação como *standard* [10].

1.4.3.1 Inserções e delecções: penalidades de *gaps*

Eventos mutacionais incluem não apenas substituições, mas também inserções e deleções. Em relação ao alinhamento e comparação de sequências, a consequência é a necessidade de introduzir *gaps* (espaços vazios) numa ou em ambas as sequências, de modo a produzir um alinhamento adequado. A penalidade pela criação de um *gap* deve ser grande o suficiente para para que estes sejam introduzidos apenas onde são necessários, e a penalidade por prolongar um *gap* deve ter em conta a probabilidade de as inserções e deleções ocorrerem ao longo de vários resíduos ao mesmo tempo [10].

1.4.3.2 Alinhamento óptimo. Algoritmos de programação dinâmica. Alinhamentos locais. Heurísticas.

É necessário ter um método de encontrar o alinhamento ótimo de entre as numerosas alternativas. No entanto, esse método deve ser consistente e biologicamente significativo. Para garantir o melhor alinhamento, muitos devem ser gerados e avaliados. Para duas sequências longas, pode verificar-se que isto demora um tempo considerável [10]. Apesar da crescente evolução da rapidez dos computadores e da eficiência dos algoritmos desenvolvidos, estes continuam a não ser suficientemente rápidos para permitir procuras exaustivas em enormes repositórios de sequências, como o GenBank ou SWISS-PROT [10, 11]. Acresce ainda a este problema o crescimento das bases de dados de sequências, que ultrapassa as melhorias na velocidade de computação [10]. No entanto, examinando ao detalhe os cálculos, verifica-se que a grande maioria do tempo é passado na avaliação repetida das mesmas porções dos alinhamentos candidatos. Este aspeto redundante da comparação de

sequências permite, por outro lado, poupar tempo, utilizando programação dinâmica [10].

Os métodos de programação dinâmica foram descritos pela primeira vez nos anos 50, fora do contexto da bioinformática, sendo aplicados pela primeira vez neste contexto por Needleman e Wunsch em 1970. Estes métodos procuram uma solução ótima para um dado problema, partindo o problema original em subproblemas cada vez mais pequenos, até que estes tenham uma solução trivial. Utilizam-se então essas soluções para construir soluções para porções cada vez maiores do problema original. Na comparação de sequências, o problema é determinar o alinhamento ótimo de duas sequências. São então gerados alinhamentos cada vez mais pequenos de partes de uma sequência com partes de outra sequência, até ao caso menor, que consiste no alinhamento de um único resíduo de uma sequência com um único resíduo de outra sequência. Para este caso, a solução é conhecida, sendo obtida da matriz de pontuação. [10] Estes "*hits*" de elevada pontuação são usados como "*seeds*" (sementes) para os algoritmos de programação dinâmica mais sofisticados e mais demorados [10].

Uma generalização de uma abordagem de programação dinâmica recursiva é o algoritmo de Smith-Waterman, utilizado pelo BLAST, que é um método exaustivo e matematicamente ótimo, que garante a descoberta do alinhamento de pontuação mais elevada. O algoritmo incorpora os conceitos de não-correspondências (*mismatches*) e intervalos (*gaps*), e identifica alinhamentos ótimos locais. Os alinhamentos locais, em que partes de uma sequência são alinhadas com partes de outra sequência, são mais relevantes biologicamente que os alinhamentos globais, onde sequências completas são alinhadas, uma vez que regiões longas de alta similaridade são uma exceção e não uma regra [10, 11].

1.4.3.3 Funcionamento do BLAST

No algoritmo do BLAST existem três passos principais: compilação de uma lista de palavras de alta pontuação, procura na base de dados por correspondências (*hits*) e extensão dos *hits* [10, 11]. No primeiro passo, o BLAST filtra as regiões de baixa complexidade – por exemplo, repetições "CA" – e remove-as da sequência de interrogação. Estas repetições de baixa complexidade em termos de composição podem gerar um número muito grande de resultados estatisticamente significativos, mas desinteressantes biologicamente. Em seguida é gerada uma lista de todas as sequências curtas, ou "palavras", que compõem a sequência de interrogação. O tamanho das palavras definido por defeito é 3 para sequências de aminoácidos e 11 para sequências de nucleótidos. BLAST usa então uma matriz de pontuação – BLOSUM62 por

defeito, para aminoácidos – para determinar todas as correspondências de alta pontuação para cada palavra na sequência de interrogação. Nesta fase não são permitidos intervalos (*gaps*). A lista de correspondências é reduzida considerando-se apenas aquelas que pontuam acima de uma dado limite ou *threshold*, T. [10] Há no entanto um dilema entre velocidade e sensibilidade: um *threshold* T maior proporciona uma maior velocidade mas aumenta a probabilidade de falhar pares relevantes, enquanto que um *threshold* T menor aumenta a probabilidade de um par de segmentos com uma dada pontuação conter um par de palavras com uma pontuação de pelo menos T. No entanto, um T baixo irá aumentar, deste modo, o número de *hits*, e consequentemente o tempo de execução do algoritmo será maior [10, 11].

No segundo passo, é efetuada uma procura na base de dados alvo de correspondências exatas da lista de palavras gerada. Uma vez que o BLAST já pré-processou e indexou as bases de dados para a ocorrência de todas as palavras em cada sequência na base de dados, a procura torna-se extremamente rápida. Quando é encontrada uma correspondência, esta é usada para semear um possível alinhamento entre a sequência de interrogação e a sequência da base de dados [10].

No terceiro passo, BLAST tenta estender o alinhamento a partir das palavras correspondentes em ambas as direções, enquanto a pontuação continuar a aumentar. O alinhamento resultante é então um "par de alta pontuação" ou HSP. Em seguida, BLAST determina se cada pontuação encontrada é maior que uma dada pontuação limite (*cutoff*), determinada empiricamente por examinação do leque de pontuações dadas por comparação de sequências aleatórias e escolhendo então um valor significativamente maior. Por fim, é determinada a significância estatística de cada pontuação, inicialmente por cálculo da probabilidade de duas sequências aleatórias, uma do tamanho da sequência de interrogação, e a outra do tamanho da base de dados (a soma dos tamanhos de todas as sequências na base de dados) poderem produzir a pontuação calculada. Quando o valor esperado, E, para uma dada sequência de uma base de dados satisfaz o valor limite selecionado pelo utilizador, a correspondência é reportada. Tipicamente são utilizados valores entre 0.1 e 0.001 [10].

1.5 *Streptococcus pneumoniae*

Streptococcus pneumoniae, também conhecido por *pneumococcus*, coloniza frequentemente o trato respiratório superior, sendo a nasofaringe humana o único reservatório natural conhecido para esta bactéria. Quando estes agentes patogénicos são aspirados para os pulmões, poderão causar

doença. *S. pneumoniae* causa tanto doenças invasivas como não invasivas, em todas as faixas etárias, particularmente em crianças com menos de 5 anos e adultos com mais de 65 anos. Além disso, pessoas com certas condições médicas como doenças crónicas do coração, pulmão ou fígado, ou anemia falciforme, têm também risco aumentado de sofrer doenças pneumocócicas. Pessoas com HIV/SIDA, ou pessoas que receberam transplantes de órgãos e se encontram a tomar medicação que diminui a sua imunidade, estão também em elevado risco de adquirir estas doenças [1].

S. pneumoniae é assim, enquanto agente patogénico humano, a causa mais comum de infecção respiratória aguda e otite média, estimando-se que seja causador de mais de três milhões de mortes em crianças todos os anos, por todo o mundo, por pneumonia, bacterémia ou meningite. Entre a população idosa ocorrem ainda mais mortes, sendo *S. pneomoniae* a principal causa de pneumonia e meningite adquiridas na comunidade [2]. Estas doenças causadas por *pneumococci* constituem assim um importante problema de saúde pública global. Além das já referidas, *S. pneumoniae* causa ainda outras doenças bastante comuns, mas menos graves, como sinusite e bronquite [3].

S. pneumoniae é transmitido por contacto directo com as secreções respiratórias de doentes e portadores saudáveis. Estima-se que, no ano 2000, tivessem ocorrido cerca de 14,5 milhões de episódios de doença pneumocócica grave, resultando em cerca de 826 000 mortes em crianças com idades compreendidas entre 1 e 59 meses. No mundo desenvolvido, casos de doença grave ocorrem principalmente em crianças com menos de dois anos e em idosos. Nos países em desenvolvimento, são afectadas sobretudo crianças com menos de dois anos, incluindo recém-nascidos, enquanto que as taxas da doença na população idosa são em grande parte desconhecidas [3].

As vacinas existentes foram desenhadas para cobrir os serótipos mais frequentemente associados com doença pneumocócica severa. Existem correntemente três vacinas conjugadas, abrangendo 7, 10 e 13 serótipos, e uma vacina não conjugada de polissacárido, que cobre 23 serótipos, comercializadas mundialmente [3].

1.5.1 Serótipo 1 e serótipo 3

Os isolados de *Streptococcus pneumoniae* são tradicionalmente caracterizados em termos da composição química das suas cápsulas de polissacárido, existindo mais de 90 tipos de serótipos capsulares diferentes [5, 13, 14]. Diferentes serótipos apresentam diferentes potenciais patogénicos e distribuição geográfica [5]. Infecções com o serótipo 3 são associadas a um risco relativo de morte aumentado, enquanto que infecções com o serótipo 1 foram

associadas a um risco de morte diminuído, apesar de ambos os serótipos causarem doença pneumocócica invasiva, independentemente da idade e outros marcadores de severidade de doença, de acordo com estudos anteriores [13].

O serótipo 1 está entre os mais comummente isolados em doença pneumocócica invasiva e raramente causa colonização nasofaringeal assintomática [14, 15]. Comparativamente a outros serótipos, a infecção por serótipo 1 é mais provável de ser identificada em jovens pacientes sem comorbilidades, apesar de ser geralmente associado com baixa mortalidade [14]. Este serótipo está também associado a surtos de doença, sobretudo em comunidades fechadas de jovens adultos (como prisões ou abrigos para sem-abrigo), sendo também as epidemias particularmente comuns na África sub-saariana [14, 15]. O serótipo 1 foi também identificado como sendo uma causa comum de doença pneumocócica invasiva no periodo neonatal [14].

Estudos anteriores apontam para a existência de linhagens geograficamente distintas do serótipo 1, tendo sido identificadas por *Multilocus Sequence Typing* (MLST) três linhagens, encontradas na Europa e América do Norte, África e Israel, e na América do Sul, respectivamente. A cápsula do serótipo 1 protege os organismos da fagocitose e destruição por células do sistema imunitário inato. [14] Apresenta também uma baixa incidência de resistência antimicrobiana [14, 15]. A colonização por estirpes do serótipo 1 é eficientemente passada por contacto próximo, mas é muito mais curta em duração e/ou apresentando menor densidade bacteriana do que aquelas associadas com outros serótipos, o que reduz a oportunidade do organismo ser disseminado internacionalmente e consequentemente a oportunidade de trocas genéticas, o que explica a sua limitada diversidade genética e elevada similaridade intra-serótipo [5, 14, 15].

Por sua vez, pneumococos do serótipo 3 são colonizadores comuns e causam frequentemente doença nasofaringeal. O serótipo 3 destaca-se dos restantes serótipos pela sua cápsula de polissacárido mais grossa, que lhe permite inibir a fagocitose, sendo mais virulento e estando assim mais frequentemente associado a casos fatais. O serótipo 3 está assim associado a uma elevada incidência de choque séptico e a uma elevada mortalidade.[16]

Enquanto que serótipos altamente invasivos, como o serótipo 1, causam doença em doentes mais jovens e saudáveis previamente, serótipos pouco invasivos, como o serótipo 3, causam doença apenas em doentes mais idosos e com mais comorbilidades, podendo ser considerados oportunísticos. [16]

Capítulo 2

Trabalho desenvolvido

2.1 Características do estudo

No presente estudo pretende-se, através do desenvolvimento de uma ferramenta de procura de genes semelhantes, escrita em Python, obter o genoma *core* e o genoma acessório de *Streptococcus pneumoniae* a partir dos genomas de várias estirpes desta espécie. Utilizaram-se genomas depositados na base de dados do GenBank e genomas sequenciados pela Unidade de Microbiologia Molecular e Infeção do Instituto de Medicina Molecular,em Lisboa. As estirpes sequenciadas pelo Instituto de Medicina Molecular, bem como algumas das estipres retiradas do NCBI, foram identificadas como pertencendo ao serótipo 1 ou serótipo 3, pretendendo-se também comparar estes dois serótipos de modo a permitir uma melhor compreensão das diferenças genómicas entre si.

Foram realizadas análises considerando diferentes parâmetros no que toca à percentagem de identidade mínima permitida entre alelos que definem um mesmo *locus* – que passaremos a designar por Conjuntos de Variantes Alélicas Possíveis (CVAPs) – e relativamente à percentagem de diferença de tamanho permitida entre essas sequências. Em termos estatísticos, foi calculado para cada conjunto de sequências encontrado o mínimo, a média e o desvio padrão relativamente à similaridade entre as sequências, a partir das distâncias obtidas através de uma matriz de distâncias, calculada por sua vez por alinhamento múltiplo. Estes valores permitiram a construção de um gráfico da média em função do mínimo, com o desvio padrão representado por uma escala de cores, ilustrando deste modo as estatísticas para cada conjunto de alelos no mesmo gráfico de pontos.

Este estudo resultou assim na construção de uma ferramenta bioinformática, que se designou por "SCRAG" – *Strict CoRe and Accessory Genome* – e

que procede à identificação dos CVAPs constituintes do genoma *core* e do genoma acessório de uma forma bastante conservadora, como explicado em seguida. Pretende-se ainda a aplicação desta mesma ferramenta a estirpes de *S. pneumoniae*, permitindo deste modo também a comparação do conteúdo genómico entre os serótipos 1 e 3 desta espécie. Os dados obtidos permitem sustentar uma melhor compreensão da biologia e funcionalidade da espécie, em que poderão ser focados estudos futuros.

2.2 SCRAG – Strict CoRe and Accessory Genome

O algoritmo desenvolvido, implementado em Python 2.6, ao qual se deu o nome de SCRAG – Strict CoRe and Accessory Genome, encontra-se disponível para acesso no repositório `https://github.com/adpolicarpo/SCRAG`.

2.2.1 Estrutura do algoritmo

SCRAG é composto por 5 *scripts*, escritos em Python, um dos quais funciona como *script* principal, a partir do qual se correm os restantes pela ordem correta. Foram ainda desenvolvidos *scripts* adicionais, como um *script* para descarregar genomas do NCBI, um *script* para verificar se cada sequência foi encontrada unicamente num CVAP, um *script* para obter sequências codificantes utilizando o Prodigal e ainda um *script* para comparação de conjuntos de resultados (conforme explicado adiante, no capítulo 4).

O *script* "principal", designado por "run_all.py" tem como função ler os ficheiros dos genomas a analisar e concatenar os mesmos num só ficheiro, que será posteriormente usado para a construção da base de dados do BLAST e para interrogar a mesma, bem como correr os quatro *scripts* que constituem o algoritmo propriamente dito, fornecendo-lhe os parâmetros necessários. O esquema representativo dos quatro *scripts* constituintes do algoritmo e do *script* principal, bem como as relações entre eles, podem ser observadas na figura 2.1.

Para correr o *script* "run_all.py" é necessário que o utilizador forneça três parâmetros:

- O tipo de análise a que quer proceder - genoma *core* (1) ou genoma acessório (2);
- A percentagem de identidade mínima entre as sequências de cada CVAP, ou seja, quão similares devem ser estas sequências, que portanto são

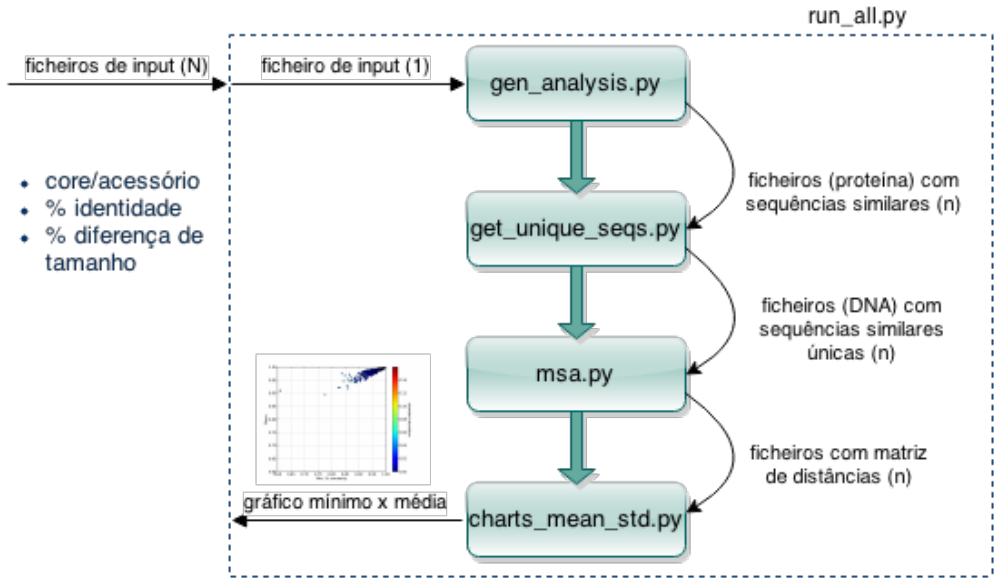

Figura 2.1: Esquema de relação entre os *scripts* constituintes do algoritmo (versão final)

consideradas como sendo "o mesmo" gene - embora possam não ser 100% semelhantes, mas representam o mesmo *locus*;

- A percentagem de diferença de tamanho permitida entre sequências do mesmo CVAP.

É então realizada uma procura de sequências semelhantes e identificação dos CVAPs entre os genomas em análise, utilizando o BLAST, sendo depois os resultados filtrados consoante a percentagem de identidade e a percentagem de diferença de tamanho entre sequências do mesmo CVAP escolhidas. Os resultados da filtragem por percentagem de identidade e percentagem de diferença de tamanho são guardados em ficheiros .fasta, obtendo-se um ficheiro para cada CVAP. A partir destes ficheiros são gerados novos ficheiros, contendo apenas as sequências únicas para cada CVAP, ou seja, eliminando as sequências que são repetidas. Por exemplo, analisando 25 genomas, um gene *core* é aquele que é encontrado, tendo em conta a percentagem de identidade, em todos os 25 genomas, pelo que se obtém um CVAP de 25 sequências semelhantes, das quais apenas algumas são exatamente iguais. Os ficheiros com as sequências únicas servem novamente de *input* para o MUSCLE **??**, procedendo-se ao alinhamento das sequências de cada CVAP e posteriormente à construção de uma matriz de distâncias utilizando o ClustalW

/citeclustalW. As matrizes de distâncias, guardadas em ficheiros, servem por sua vez de *input* para a construção de um gráfico da média em função do mínimo da similaridade (1 - distância) entre cada sequência do CVAP, que permite a visualização dos resultados, representando cada CVAP como um ponto no gráfico - à exceção dos CVAPs com apenas uma sequência única, os quais não é possível alinhar com o MUSCLE, mas cujos pontos estariam sempre representados no canto superior direito do gráfico, correspondendo aos valores de média e mínimo 100%.

2.2.2 Explicação do algoritmo

O algoritmo desenvolvido tem por base o processo de comparação de sequências utilizando o BLAST. Foram testadas várias abordagens, sendo que as duas primeiras, apresentadas no capítulo 2.3, apresentavam algumas falhas ou dificuldade na interpretação dos resultados. A terceira abordagem, aqui descrita, foi a abordagem final que resulta de melhorias nas duas abordagens prévias. De notar que, em cada uma das versões, foram decorrendo várias mudanças ao longo do desenvolvimento, sendo que o que aqui se apresenta pretende resumir as alterações efetuadas de uma forma simples, de forma a que se compreenda o fundamento e os principais desenvolvimentos de cada uma das versões.

Nesta terceira e última versão é apresentado um *script* "run_all.py", que recebe como parâmetros o tipo de análise (1 para genoma *core* e 2 para genoma acessório), a percentagem de identidade e a percentagem de diferença de tamanho máxima permitida entre sequências de cada CVAP, como referido acima. Assim, para correr o algoritmo na linha de comandos, executa-se o comando "python run_all.py «tipo de análise» «percentagem de identidade» «percentagem de diferença de tamanho»". Por exemplo, se pretendermos fazer uma análise do genoma *core* e escolhermos como percentagem de identidade 80% e como percentagem de diferença de tamanho 10%, corremos o algoritmo do seguinte modo: "python run_all.py 1 80 10". Antes da comparação de sequências, o primeiro passo a ser executado é a concatenação de todos os ficheiros correspondentes a genomas ou proteomas de estirpes diferentes num só ficheiro. Este processo é efetuado também por este *script*, que lê ficheiros numa diretoria, concatenando-os num só ficheiro FASTA. São utilizados neste passo ficheiros em formato .faa (*FASTA Amino Accid*), uma vez que se pretendem utilizar na comparação sequências de aminoácidos, em vez de sequências de ADN, devido à redundância do código genético – codões diferentes podem codificar para o mesmo aminoácido – o que vai ter implicações ao nível da identidade e similaridade entre sequências e do significado da mesma a nível biológico. É ainda gerado um outro ficheiro

correspondente à concatenação de ficheiros .ffn (*FASTA nucleotide coding regions*), para que posteriormente possam ser obtidas as sequências de ADN correspondentes, uma vez que o objetivo final deste tipo de análise será a aplicação a estudos de tipagem microbiana, e como tal pretende-se discriminar o mais possível. De notar que os ficheiros com sequências de aminoácidos e os ficheiros com sequências de ácidos nucleicos devem corresponder às mesmas estirpes e as proteínas/genomas devem estar pela mesma ordem nos dois ficheiros relativos ao proteoma/genoma de cada estirpe. Em seguida, são executados os quatro *scripts* restantes, através da execução da linha de comandos utilizando o python. Ou seja, apenas é requerido ao utilizador que corra o *script* principal, fornecendo os parâmetros necessários, sendo que os restantes *scripts* e respetivos parâmetros são corridos automaticamente através do *script* principal.

Em seguida são executados os passos de construção e interrogação da base de dados do BLAST, leitura e filtragem dos resultados da interrogação consoante os parâmetros definidos, como ilustrado na figura 2.2. Estes são os passos fundamentais para a obtenção de resultados, ocorrendo no *script* "gen_analysis.py", tendo-se utilizado as funções disponíveis no biopython para os efetuar. Este *script* recebe como parâmetros, além do tipo de análise, da percentagem de identidade e da percentagem de diferença de tamanho permitida, o diretório onde se encontram os ficheiros com sequências de aminoácidos, designado por "faa_files", onde se encontra também o ficheiro .fasta resultante da concatenação dos ficheiros .faa. A partir do ficheiro contendo todas as sequências de aminoácidos de todos os proteomas em análise, é criada localmente uma base de dados do BLAST, e em seguida interroga-se esta base de dados a partir do mesmo ficheiro usado para a criar. É gerado um ficheiro .xml com os resultados do BLAST, que é lido em seguida. De notar que apenas se efetua tanto a construção da base de dados como a interrogação da mesma se não existir, para aquele conjunto de genomas em análise, a base de dados e o ficheiro .xml com os resultados do BLAST, respetivamente. Esta verificação efetua-se para tornar o processo mais rápido, no caso de se pretender realizar análises múltiplas, considerando diferentes percentagens de identidade e de diferença de tamanho.

Ao executar o BLAST, para cada sequência de interrogação encontram-se sequências semelhantes, sendo que neste caso pretendemos obter não apenas a sequência mais semelhante, ou seja, aquela que obtém a maior pontuação, mas várias sequências semelhantes cuja identidade seja igual ou superior à percentagem definida. São utilizadas nesta fase sequências de aminoácidos, ou seja, proteomas, mas designaremos de uma forma geral "genoma", uma vez que estas sequências correspondem às sequências de ácidos nucleicos, apenas se usando as primeiras devido à redundância do código genético, como

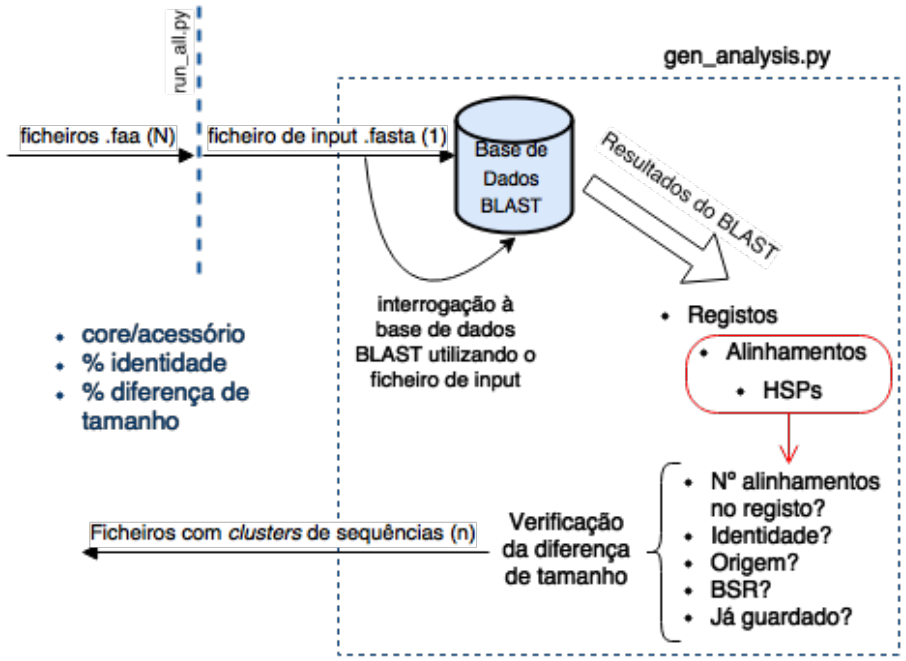

Figura 2.2: Esquema do *script* "gen_analysis.py", onde ocorrem os principais passos da análise, no algoritmo

referido acima. Assim, para o genoma *core*, pretende-se encontrar tantas sequências que correspondem à sequência de interrogação, com percentagem de identidade igual ou superior à definida, quanto o número de genomas em análise. Já para o genoma acessório pretende-se encontrar menos sequências que o número de genomas em análise. Ou seja, quando se executa o BLAST, para N genomas uma sequência de interrogação encontra na base de dados várias sequências, que com ela são alinhadas, sendo que quando são encontradas N sequências estas se consideram candidatas a representar um gene *core*, e quando são encontradas entre uma a N-1 sequências semelhantes, um gene acessório. De notar que há uma verificação de vários outros parâmetros, pelo que apenas os CVAPs que passam na verificação de todos os parâmetros são considerados genes *core* ou acessórios, consoante o número de alelos no CVAP. No entanto, a primeira verificação a ser efectuada é o número de alinhamentos encontrados numa interrogação do BLAST, ou seja, com quantas sequências a sequência de interrogação é alinhada, verificando-se aí se um alinhamento é ao não candidato a representar um gene *core* ou acessório. Desta forma, não são considerados os CVAPs em que se encontram mais do que N alinhamentos, sendo que este método, apesar de mais restritivo, apenas nos

fornece os alinhamentos que à partida são bons candidatos a genes *core* ou acessórios, sendo estes ainda filtrados por outros parâmetros considerados. Desta forma, são eliminados também todos os genes duplicados ou parálogos.

Ao analisar e filtrar os resultados do BLAST, procuram-se assim sequências correspondentes à sequência de interrogação, obtendo-se CVAPs com N (gene *core*) ou menos de N (gene acessório) sequências, como já referido, mas pretende-se que as sequências encontradas cumpram vários outros parâmetros (figura 2.2), sendo que quando não cumprem um deles, são excluídas do conjunto de resultados. Ou seja, apenas são guardados como resultados os grupos de genes ou CVAPs que cumprem os vários requisitos definidos que são, além de terem N ou menos de N sequências:

- Apresentam uma percentagem de identidade igual ou superior à percentagem de identidade definida pelo utilizador;
- Provêm de estirpes diferentes – só desta forma se garante que um gene pertence ao genoma *core*, ou que não se tratam de duplicações de genes, por exemplo;
- Os valores do *BLAST Score Ratio* (BSR) são superiores a 0.6 – aspeto que se encontra explicado mais abaixo;
- Não estão ainda no conjunto de resultados – ou seja, se um CVAP já foi encontrado e passou nos restantes parâmetros não é guardado novamente no conjunto de resultados, pois pretende-se que não haja resultados redundantes.

Mais detalhadamente, podemos dizer que o BLAST gera como resultados vários registos, correspondendo cada um a uma interrogação de uma sequência contra a base de dados, e fornece-nos para cada registo uma série de informações relevantes sobre a sequência de interrogação e as sequências encontradas a partir desta, bem como os vários HSPs (pares de elevada pontuação) encontrados para cada sequência que corresponde à sequência de interrogação. Para analisar os resultados, começamos por escolher apenas os registos em que foram encontradas N sequências ou uma até N menos uma, como já referido. Em seguida escolhemos como melhor HSP – o que vamos considerar – aquele que apresenta melhor pontuação no alinhamento. É calculado em seguida o valor da identidade para esse HSP, dividindo o número de letras (aminoácidos) que correspondem, no alinhamento, pelo tamanho total do alinhamento, ou seja, obtém-se uma percentagem do número de correspondências exatas. É efetuada uma contagem dos alinhamentos em que esta percentagem de identidade é superior ao valor definido, em cada registo.

Ou seja, sempre que um alinhamento apresenta uma identidade superior ao valor escolhido pelo utilizador, a contagem aumenta, sendo que no final o valor da contagem deve corresponder ao número de alinhamentos naquele registo – portanto, todos os alinhamentos do registo possuem uma percentagem de identidade superior à definida. É também verificada a proveniência (estirpe) da sequência encontrada em cada alinhamento, que consta na definição da mesma, sendo esta origem guardada numa lista para cada registo. É ainda calculado o *BLAST Score Ratio* (BSR), que como o nome indica é um rácio de pontuações do BLAST. Mais concretamente, obtém-se o BSR dividindo a pontuação obtida num alinhamento de uma sequência contra outra, pela pontuação de referência, que é a pontuação do alinhamento quando uma sequência é alinhada contra ela própria [17, 18] (equação 2.1).

$$BSR = \frac{Pontuação\ da\ interrogação}{Pontuação\ de\ referência} \tag{2.1}$$

Os valores do BSR variam assim entre 0 e 1, sendo que um valor 0 indica que não há de todo correspondência entre as sequências, e uma valor de 1 indica uma correspondência perfeita – como quando a sequência se encontra a si própria, ou outra sequência igual. Um valor de BSR elevado significa portanto que um alinhamento é bom, enquanto um valor de BSR baixo representa um mau alinhamento. [17, 18] Considerou-se como valor limite 0.6, pois embora tradicionalmente seja usado o valor de 0.4 – que corresponde a 30% de semelhança em 30% do tamanho da sequência [17] – verificou-se por simulação que um valor de 0.6 corresponde a um valor de similaridade de cerca de 80%, que representa um bom valor para uma análise deste tipo, em que se pretende encontrar sequências semelhantes. Deste modo, há uma dupla verificação do grau de semelhança das sequências, tornando o método mais robusto.

Assim, para cada registo apresentado nos resultados do BLAST estão presentes as várias sequências semelhantes à sequência de interrogação, e obtemos para cada registo o valor de identidade do alinhamento de cada uma das sequências, a origem de todas as sequências encontradas e os valores de BSR para estes alinhamentos. É então verificado se são cumpridos todos os requisitos para se considerar que as sequências de cada registo, ou seja, cada CVAP obtido, representam de facto um gene *core* ou um gene acessório.

Após a filtragem dos resultados do BLAST, são obtidas as sequências originais – através do ficheiro com todas as sequências, utilizado para construir a base de dados e para a interrogar – uma vez que o BLAST apenas nos dá a parte das sequências que corresponde, num alinhamento. Obtemos assim o nosso conjunto de resultados, ao qual é ainda efectuada uma ex-

clusão por tamanho de sequências, consoante a percentagem de diferença de tamanho definida. Ou seja, os CVAPs cujas sequências apresentam entre si uma diferença de tamanho superior à percentagem definida, são excluídos. Este passo é importante precisamente devido ao facto de recuperarmos as sequências completas, uma vez que o BLAST apenas nos dá a parte que corresponde, que nem sempre é correspondente ao tamanho total da sequência. Ou seja, apesar de o BLAST nos fornecer sequências bastante idênticas, como verificado através dos alinhamentos, na verdade, se tivermos em conta o tamanho das sequências originais, elas poderão não ser assim tão semelhantes. Assim, em cada CVAP é verificado o tamanho da sequência mais pequena e da sequência maior, e em seguida obtém-se o rácio, dividindo o mínimo pelo máximo. Se escolhermos como percentagem de diferença de tamanho permitida 20%, por exemplo, vamos então considerar apenas os CVAPs em que o rácio é igual ou superior a 0,8 (80%).

No final destes passos obtemos então os ficheiros com os resultados: um ficheiro .fasta para cada CVAP. Estes ficheiro são guardados em pastas que identificam os parâmetros opcionais utilizados na análise, dentro da pasta do programa. Assim, dentro da pasta dos resultados, obtemos uma pasta que indica se obtemos genes *core* ou acessórios e a percentagem de identidade utilizada na análise, e dentro dessa pasta localizam-se outras duas, uma indicando a percentagem de identidade e a percentagem de diferença de tamanho, e outra contendo os ficheiros referentes aos CVAPs antes do passo da exclusão por tamanho, ou seja, considerando todos os genes encontrados tendo em conta apenas a percentagem de identidade, e permitindo qualquer diferença de tamanho – o mesmo que se obteria ao escolher como percentagem de diferença de tamanho 100%. Esta última pasta é gerada para acelerar o processo, no caso de o utilizador querer testar os seus dados para várias diferenças de tamanho diferentes, bem como para permitir uma comparação entre efetuar ou não a exclusão por tamanho. Ou seja, se pretendermos testar diferentes percentagens de diferença de tamanho para uma mesma percentagem de identidade, apenas da primeira vez são realizados os passos de filtragem dos resultados do BLAST, sendo nas seguintes utilizados os ficheiros já existentes e o processo continua a partir do passo de exclusão por tamanho. Estes ficheiros, contendo os CVAPs obtidos, serão utilizados como *input* pelo *script* seguinte.

O *script* "get_unique_seqs.py" recebe como parâmetros o diretório com os ficheiros .faa ("faa_files"), o diretório com os ficheiros .ffn ("ffn_files") e o tipo de análise, a percentagem de identidade e a percentagem de diferença de tamanho, definidos pelo utilizador. Mais uma vez, o utilizador não precisa de fornecer estes parâmetros, uma vez que não corre o *script* diretamente. Nesta fase, são lidos os ficheiros relativos aos CVAPs encontrados, sendo

estes filtrados de modo a que fiquem apenas as sequências únicas, ou seja, as sequências repetidas em cada CVAP são eliminadas. São gerados novos ficheiros contendo apenas as sequências únicas, mas em vez de se manterem as sequências de aminoácidos, são recuperadas as sequências de ácidos nucleicos correspondentes – que serão utilizadas a partir deste passo – a partir do ficheiro gerado inicialmente, resultante da concatenação dos vários ficheiros contendo genomas.

Utilizando agora as sequências únicas, e de forma a analisar os resultados gerados, procede-se ao alinhamento múltiplo das sequências de cada CVAP, correndo o *script* "msa.py". Este recebe como parâmetros o tipo de análise, a percentagem de identidade e a percentagem de diferença de tamanho máxima permitida. É realizado um alinhamento múltiplo utilizando o MUSCLE, sendo utilizados como *input* os ficheiros com as sequências únicas, e obtendo como *output* os ficheiros relativos aos alinhamentos, no formato ClustalW, cuja extensão é ".aln". Nesta fase é utilizado o MUSCLE, pois é mais rápido do que o ClustalW para um elevado número de sequências [19]. Assim, se desejarmos realizar uma análise a muitos genomas será mais vantajoso usar o MUSCLE para o alinhamento das sequências. De seguida, utilizando os ficheiros dos alinhamentos, obtém-se uma matriz de distâncias para cada alinhamento, correndo o ClustalW. Cada matriz de distâncias corresponde a um ficheiro gerado com extensão ".dst". São ainda gerados automaticamente ficheiros de extensão ".ph", que poderão ser utilizados posteriormente na construção de árvores filogenéticas.

As matrizes de distâncias geradas são utilizadas para a obtenção de estatísticas para cada CVAP. Sendo que cada distância na matriz corresponde inversamente à similaridade (semelhante à identidade) entre as sequências (pois as distâncias são calculadas a partir das árvores filogenéticas, e os alelos de cada CVAP podem ter tamanhos diferentes) podem utilizar-se estes valores para calcular o mínimo, a média e o desvio padrão da similaridade entre as sequências de cada CVAP. Este processo efetua-se no *script* "charts_mean_std.py", que recebe como parâmetros o tipo de análise, a percentagem de identidade e a percentagem de diferença de tamanho permitida. Tendo obtido o mínimo, média e desvio padrão da similaridade entre as sequências de cada CVAP, é possível gerar-se um gráfico de pontos, cada ponto correspondendo a cada um dos CVAPs. O mínimo da similaridade é assim representado no eixo das abcissas e a média no eixo das ordenadas, enquanto que o desvio padrão é representado pela cor do ponto, de acordo com uma escala de cores. Para os CVAPs em que se obtém apenas uma sequência única, e como não é possível produzir o alinhamento de uma sequência isolada, não pode ser gerada uma matriz de distâncias, pelo que estes não aparecem representados no gráfico – embora correspondam às coordenadas (100, 100). De notar

que nesta fase são utilizados valores de similaridade, calculados a partir das distâncias, que não são exatamente iguais aos valores de identidade calculados inicialmente e que correspondem ao número de correspondências exatas de resíduos num alinhamento. Além disso, são utilizadas as sequências completas para proceder aos alinhamentos múltiplos, enquanto que para o cálculo da identidade foram utilizados valores fornecidos pelo BLAST, que procede a alinhamentos locais e apenas nos dá uma parte da sequência (HSP). O aspeto do gráfico gerado pode ser observado na figura 2.3, que corresponde a uma análise do genoma *core*, para 25 genomas, considerando 70% de identidade e 30% de diferença de tamanho máxima permitida entre sequências. Tanto os ficheiros com as sequências únicas, como os ficheiros dos alinhamentos, das matrizes de distância e o ficheiro do gráfico gerado, são gravados no diretório que identifica os parâmetros utilizados na análise, conforme explicado anteriormente.

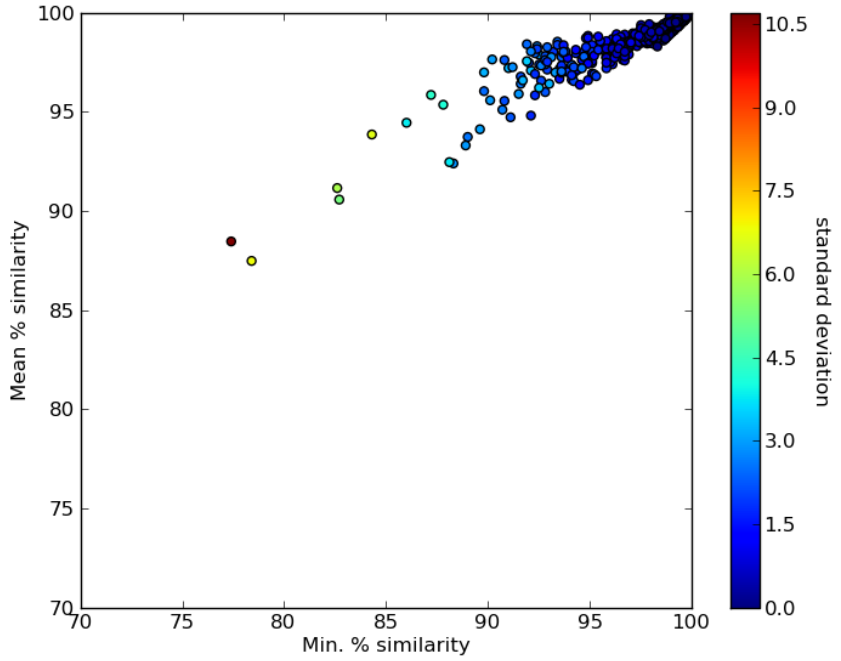

Figura 2.3: Gráfico obtido para uma análise do genoma *core*, utilizando 25 genomas de *Streptococcus pneumoniae*, considerado como parâmetros para obter cada CVAP 70% de identidade e 30% de diferença de tamanho máxima permitida.

2.3 Versões anteriores

O algoritmo desenvolvido, descrito na secção anterior, e cujos resultados de uma análise a genomas de *S. pneumoniae* serão apresentados no capítulo 3, foi elaborado após serem testadas várias alternativas e considerações diferentes. De um modo geral, foram desenvolvidas duas versões antes da versão atual, sendo que a última resulta de melhorias nas versões prévias. A primeira versão foi desenvolvida utilizando um processo iterativo, que se verificou não funcionar conforme pretendido, pelo que se testou um outro método, que consistia em comparar um genoma contra todos os genomas, e comparar os resultados que se obtinham utilizando diferentes genomas. Mostrando-se um processo pouco prático e intuitivo, decidiu-se comparar diretamente todos os genomas contra todos os genomas – método utilizado na terceira versão – o que, embora se mostre muito demorado computacionalmente quando se analisam muito genomas, se revelou o método mais prático, intuitivo e com resultados consistentes.

2.3.1 Primeira versão

A primeira abordagem para desenvolver uma ferramenta de comparação de genomas que nos permitisse obter o genoma *core* (posteriormente também o genoma acessório) consistia na comparação de um genoma com outro genoma, e obtidos os resultados dessa comparação em que havia um valor de identidade superior ao definido, comparava-se um outro genoma com os resultados da comparação anterior. Esta primeira versão compreendia assim um processo iterativo, obtendo-se no final o resultado das iterações de todos os genomas, ou seja, os genes *core*. Nesta fase, começou por se considerar um valor fixo de identidade de 80%, e obtinham-se apenas as sequências de tamanho igual. Também começou por se calcular o valor de identidade dividindo a pontuação do HSP pelo tamanho total do alinhamento. Esta versão era constituída por apenas um *script*, que efetuava todos os passos.

O processo iniciava-se utilizando os ficheiros em formato GenBank (.gbk) obtidos da base de dados do NCBI, e convertendo-os para ficheiros de formato FASTA, lendo a informação necessária dos ficheiros .gbk – utilizando apenas a parte referente às sequências que codificam para genes – e gerando uma designação ou nome para identificar cada sequência. Assim, os ficheiros em formato FASTA gerados continham apenas as designações dos genes – linhas ímpares – e a sequência de cada gene – linhas pares. A partir de um destes ficheiros em formato FASTA – correspondente a um genoma – era gerada um base de dados do BLAST e um outro ficheiro correspondente a um genoma era utilizado para interrogar esta base de dados, como ilustrado

na figura 2.4. Obtinham-se assim os genes em comum aos dois genomas comparados, sendo que dos resultados do BLAST, apenas era considerado o primeiro alinhamento para cada sequência de interrogação e o primeiro HSP de cada alinhamento, e apenas eram considerados os alinhamentos com um valor de identidade acima do valor definido – calculado da maneira acima referida – e em que a sequência de interrogação e a sequência correspondente no alinhamento tinham o mesmo tamanho. Como resultado desta primeira iteração eram guardadas as sequências de interrogação num ficheiro em formato FASTA, com a mesma estrutura dos ficheiros com genomas, que seria utilizado em seguida para interrogar uma nova base de dados, gerada utilizando um terceiro ficheiro .fasta com um genoma (figura 2.4). Esta seria assim a segunda iteração, sendo os resultados novamente gravados num ficheiro em formato FASTA, que mais uma vez seria utilizado para interrogar uma base de dados construída com um quarto ficheiro de um genoma, constituindo a terceira iteração, e assim por diante, até todos os ficheiros terem sido comparados entre si. No final, obtinha-se um ficheiro em formato FASTA resultante da comparação de todos os genomas, contendo apenas uma sequência de cada CVAP, representando assim todas as sequências do mesmo, apresentando entre si mais de 80% de similaridade e igual tamanho. Esta sequência seria a sequência de interrogação utilizada na primeira comparação, e que se mantém como sequência de interrogação durante todas as iterações, uma vez que era o ficheiro com resultados que era novamente usado como ficheiro de interrogação.

2.3.1.1 Problemas e limitações

Ao desenvolver e testar esta primeira versão foram sendo descobertos alguns problemas e limitações no método utilizado, nomeadamente:

- O BLAST não retorna as sequências completas, mas apenas a porção da sequência em que há correspondência com a outra sequência, o que pode levar a diferenças a nível da identidade, quando comparando novamente a sequência retornada pelo BLAST;
- Ao alterar a ordem porque se correm os ficheiros dos genomas, originam-se resultados diferentes, ou seja, não se chega a um número de genes fixo, que seja igual para os mesmos genomas em análise.

Para resolver o problema enunciado no primeiro ponto, o processo é bastante simples: foi necessário criar um "dicionário" que permitisse recuperar as sequências originais, que estão nos ficheiros dos genomas. Já para o exposto no segundo ponto, não se conseguiu encontrar uma solução, pelo que

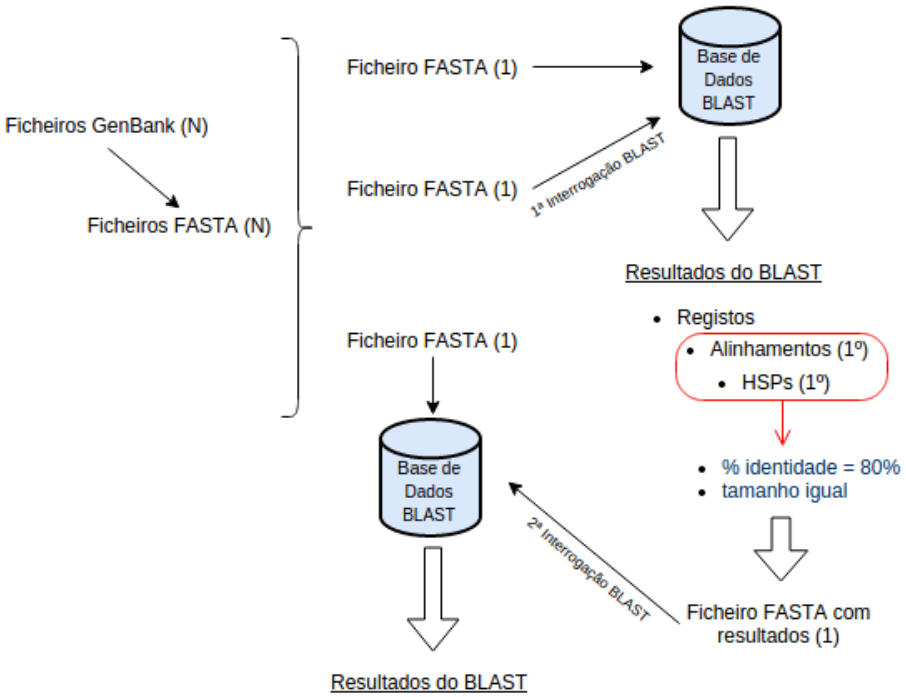

Figura 2.4: Esquema explicativo do algoritmo na sua primeira versão. O algoritmo era constituído por apenas um *script*, tratando-se de uma abordagem mais simples para obter o genoma *core*, com base num processo iterativo, considerando apenas sequências com tamanho igual e mais de 80% de identidade. A imagem representa a primeira e segunda iterações do processo, utilizando três genomas.

este constitui uma limitação do método utilizado, e que levou a que novos métodos fossem testados para tentar melhorar este aspeto. No entanto, foi possível perceber porque isto acontece. Como ilustrado na figura 2.5, ao comparar duas sequências, estas podem apresentar entre si similaridade acima do limiar definido – por exemplo, considerando um limiar de 70%, duas sequências apresentam entre si 75% – e ao comparar a segunda sequência com uma terceira, elas também podem apresentar um valor de similaridade de 75%, que também é acima do limiar dos 70%. No entanto, se trocarmos a ordem por que efetuamos as comparações, e compararmos em primeiro lugar a primeira sequência com a terceira, o alinhamento pode apresentar um valor de identidade abaixo dos 70% – por exemplo 50% – e desta forma vamos excluir uma sequência ou CVAP que antes não excluíamos. Desta forma, o número de CVAPs que obtemos no final vai depender da ordem por que efetuamos as comparações das sequências de cada CVAP.

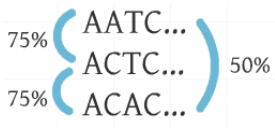

Figura 2.5: Problema na comparação de sequências usando um processo iterativo: os resultados mudam consoante a ordem por que as sequências são comparadas, pois consideramos uma percentagem de identidade inferior a 100%.

Devido à limitação apresentada no terceiro ponto, e sem se encontrar um solução óbvia que não alterasse por completo o algoritmo, decidiu-se assim procurar um método que permitisse obter um resultado fixo, independentemente da ordem por que estão os ficheiros dos genomas. Deste modo, testou-se o método apresentado em seguida, numa segunda versão do algoritmo.

2.3.2 Segunda versão

Tendo em conta a limitação apresentada anteriormente, tentou-se uma nova abordagem, que consistia em gerar uma base de dados do BLAST com todos os genomas, e comparar cada genoma com esta base de dados, obtendo não apenas o melhor alinhamento do BLAST, mas todos os bons alinhamentos para cada sequência, em vez do processo iterativo em que se adiciona um genoma de cada vez à análise.

Nesta nova versão (figuras 2.6 e 2.7), após o passo de conversão dos ficheiros do GenBank (.gbk) para ficheiros FASTA (.fasta) – que passou a constituir um novo *script*, independente dos restantes – começava-se por gerar a base de dados do BLAST, a partir de um ficheiro resultante da concatenação de todos os ficheiros FASTA contendo genomas. Em seguida, cada um dos ficheiros contendo genomas era usado como ficheiro de interrogação, e os resultados de cada uma destas interrogações (ficheiros .xml) eram lidos e filtrados de acordo com os requisitos pretendidos (figura 2.6). Começava-se por verificar o número de alinhamentos em cada registo dos resultados do BLAST, para averiguar se estávamos perante um gene *core* (tantos alinhamentos quanto genomas em análise) ou um gene acessório (menos alinhamentos que o número de genomas em análise), sendo que nesta fase foram gerados dois *scripts* que diferiam neste aspeto, uma para obter o genoma *core* e outro para obter o genoma acessório – designados respetivamente “vsall.py” e “vsall_accessory.py”. Quando o número de alinhamentos é superior ao número de genomas em análise, as sequências eram sempre excluídas do conjunto de

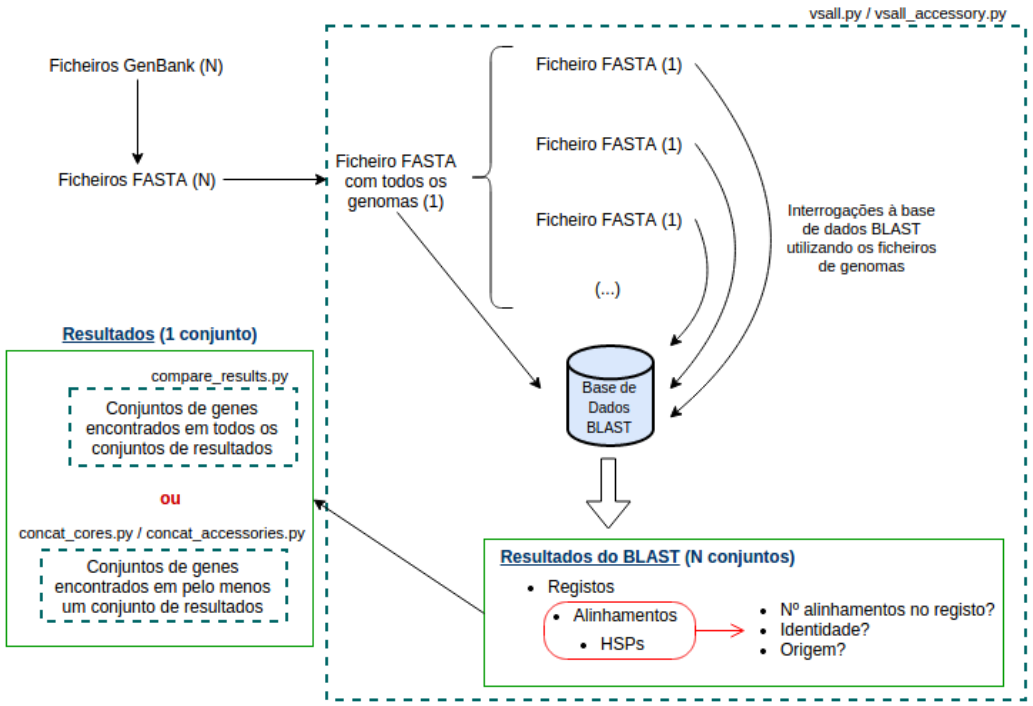

Figura 2.6: Esquema dos principais passos da segunda versão do algoritmo.

resultados. Considerando apenas o HSP com pontuação mais elevada para cada alinhamento, verificava-se então se a percentagem de identidade desse alinhamento era superior ao valor definido, sendo que todos os alinhamentos deveriam cumprir este requisito. A percentagem de identidade era calculada dividindo o número de letras (ácidos nucleicos) que tinham correspondência no alinhamento pelo tamanho total do alinhamento (número total de pares de ácidos nucleicos alinhados). Era também verificada a proveniência das sequências, ou seja, quais os genomas de que eram originárias, de modo a verificar que todas as sequências provinham de genomas diferentes, e portanto que não se tratavam de duplicações de um gene no mesmo genoma, por exemplo, conduzindo dessa forma ao enviesamento dos resultados.

Após serem filtrados os resultados do BLAST, eram então recuperadas as sequências originais, completas – utilizando um dicionário obtido através do ficheiros com todas as sequências, utilizado para construir a base de dados – uma vez que o BLAST apenas devolve a parte da sequência que alinha, e por vezes contendo intervalos (*gaps*). Os resultados da comparação de cada genoma com a base de dados consoante os parâmetros definidos eram então guardados num único ficheiro com todas as sequências, quando se procedia à

análise do genoma *core*, e em ficheiros separados, um para cada CVAP, cada conjunto dentro de uma pasta referente àquele genoma, aquando da análise do genoma acessório. Para cada genoma era assim gerado um conjunto de resultados, sendo que cada conjunto era diferente dos outros, uma vez que cada genoma possui genes diferentes dos outros. Era portanto necessário proceder a uma uniformização dos resultados, de forma a obter apenas um conjunto de resultados final, em vez de termos tantos conjuntos de resultados diferentes quanto o número de genomas em análise. Assim, pensou-se na melhor forma de juntar os conjuntos de resultados num só, e que seria também independente da ordem por que se analisam os genomas, uma vez que todos os genomas são comparados com todos os genomas. Surgiram então duas hipóteses para os genes *core* (figura 2.6):

- A partir do ficheiro de resultados com menos CVAPs encontrados, verificar se cada CVAP está também nos outros conjuntos de resultados, e se não estiver, removê-lo do nosso conjunto de resultados final;
- Comparar todos os conjuntos de resultados, um a um, e ir juntando todos os CVAPs diferentes que aparecem – ou seja, se um CVAP ainda não está no conjunto de resultados final, passa a estar, e se já estava não se adiciona, para não ficar repetido.

De entre estas duas hipóteses, a primeira mostra-se mais restritiva, garantindo que cada CVAP representa de facto um gene *core*, uma vez que o mesmo CVAP foi encontrado ao comparar cada genoma com a base de dados com todos os genomas, e é portanto independente da ordem de comparação das sequências. A segunda hipótese, que foi o modelo adotado também para a análise do genoma acessório – uma vez que nesse caso queremos os genes que não estão em todos os genomas, incluindo os genes únicos, característicos de cada estirpe – não é tão restritiva como a primeira, juntando assim todos os CVAPs que, pelo menos uma vez foram encontrados como tendo similaridade acima da percentagem definida. Ambos os métodos foram testados, num *script* designado "compare_results.py" para o método mais restritivo – o primeiro a ser testado – e "concat_cores.py" para o método menos restritivo (genoma *core*). Nesta fase, além de se juntar os conjuntos de resultados num só, também se efetuava a exclusão dos CVAPa em que havia uma diferença de tamanho entre sequências superior à percentagem definida, sendo que no final se obtinha um ficheiro com todas as sequências de todos os CVAPs de genes *core*, bem como ficheiros individuais relativos a cada CVAP obtido. O *script* equivalente para o genoma acessório foi designado "concat_accessories.py", mas gerava uma lista com os nomes dos genes a manter no conjunto de resultados, em vez de gerar novos ficheiros para cada CVAP.

Para completar a análise, procedeu-se à obtenção das sequências únicas de cada CVAP, gerando-se um ficheiro relativo a cada um, em que constavam apenas as sequências diferentes, identificadas por um nome genérico, atribuído sequencialmente, numerando o CVAP e a variante da sequência representante de cada gene. Este passo ocorria também em *scripts* diferentes para o genoma *core* e o genoma acessório designados "read_genes.py" e "read_genes_accessory.py", respetivamente (figura 2.7). Em seguida procedia-se ao alinhamento das sequências de cada CVAP com o MUSCLE e à geração de uma matriz de distâncias com o ClustalW a partir de cada ficheiro de alinhamento múltiplo. O *script* onde ocorriam estes passos, designado "msa.py" era já comum à análise do genoma *core* e do genoma acessório, bem como o seguinte e último, designado "charts_mean_std.py", em que era gerado um gráfico com os pontos correspondentes às estatísticas de cada CVAP, conforme descrito no capítulo 2.2, para a versão final do algoritmo.

Todos os *scripts* constituintes desta versão eram corridos sucessivamente, através de um outro *script* a que se atribuiu o nome "run_all.py" e que recebia como argumentos o tipo de análise (1 para o genoma *core* e 2 para o genoma acessório), a percentagem de identidade e a percentagem de diferença de tamanho entre sequências permitida. Os restantes *scripts* referidos eram corridos automaticamente, recebendo como argumentos a percentagem de identidade o primeiro *script* ("vsall.py" ou "vsall_accessory"), o tipo de análise, a percentagem de identidade e a percentagem de diferença de tamanho os dois últimos, comuns à análise do genoma *core* e do genoma acessório ("msa.py" e "charts_mean_std.py"), e a percentagem de identidade e percentagem de diferença de tamanho os restantes *scripts*. O esquema do algoritmo pode ser melhor compreendido observando a figura 2.7.

2.3.2.1 Problemas e limitações.

Nesta versão, deparamo-nos com dois problemas principais:

- A necessidade de obter um único conjunto de resultados, uma vez que comparamos cada gene de cada genoma com uma base de dados com todos os genomas, correndo o BLAST, sendo assim produzidos tantos conjuntos de resultados diferentes quanto o número de genomas que analisamos;

- A dimensão e quantidade dos ficheiros de resultados obtidos: são obtidos tantos conjuntos de resultados quanto o número de genomas em análise, obtendo-se um ficheiro para cada genoma, quando realizamos uma análise ao genoma *core*, e um ficheiro para cada gene de um ge-

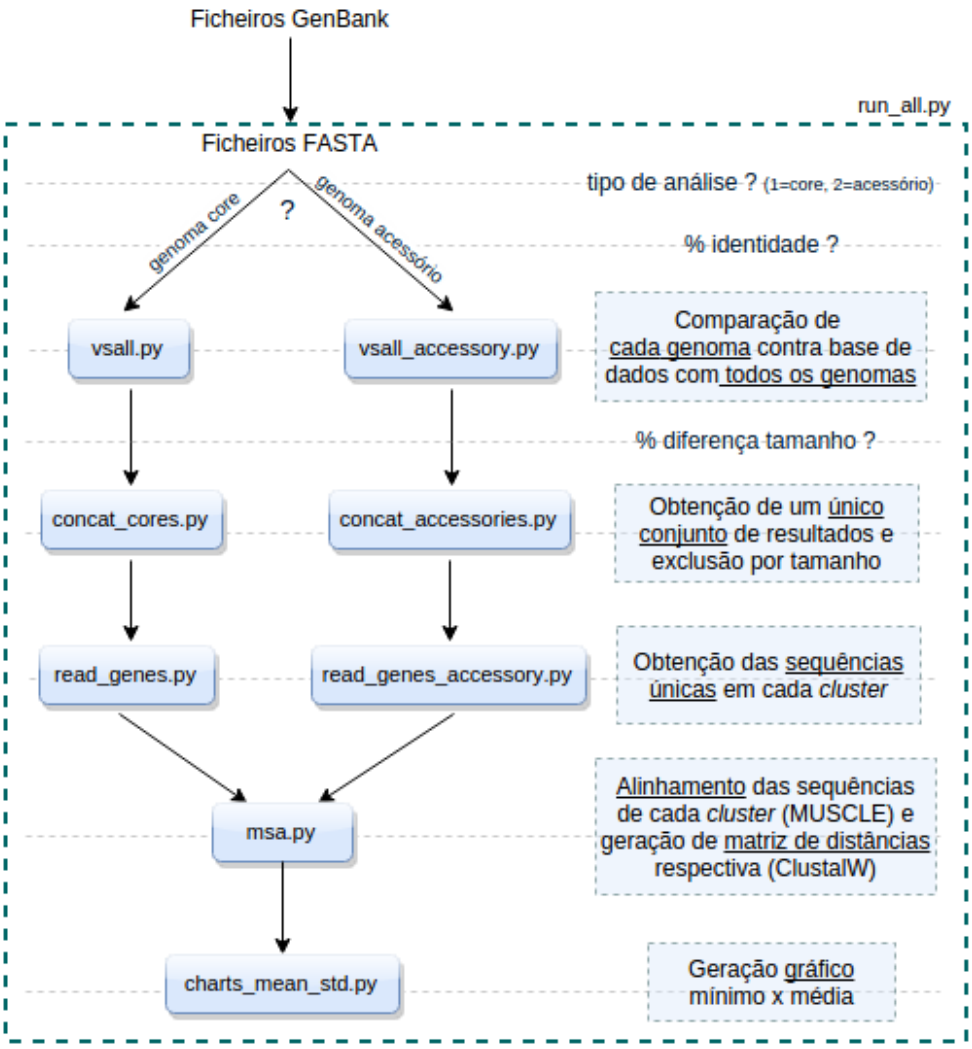

Figura 2.7: Esquema explicativo da estrutura do algoritmo na sua segunda versão (considerando o *script* "concat_cores.py" no passo de uniformização dos resultados).

noma, dentro de uma pasta relativa a esse genoma, quando pretendemos analisar o genoma acessório.

Para o primeiro problema apresentado, dos dois métodos testados para o genoma *core*, começou-se por desenvolver a versão mais restritiva, uma vez que é a única que garante que todos os genes comuns a todos os genomas são sempre encontrados, qualquer que seja o ficheiro de genoma que corremos contra a base de dados, e portanto independentemente da ordem por que as comparações são feitas. Ou seja, cada uma das sequências de um CVAP, ao ser utilizada como sequência de interrogação pelo BLAST, encontra as mesmas sequências constituintes desse CVAP, com uma percentagem de identidade superior à escolhida. Assim seria mais fácil afirmar que um determinado gene é um gene *core*, tendo em conta a percentagem de identidade definida. No entanto, é preciso pensar se faz sentido excluir do genoma *core* os CVAPs que não são encontrados na comparação de todos os genomas com a base de dados. Uma vez que estamos a considerar uma percentagem de identidade, geralmente diferente de 100%, mas alta o suficiente para podermos considerar que um gene é "o mesmo", podemos pensar que quando um CVAP é encontrado ao comparar um dos genomas com uma base de dados com todos os genomas, isto significa que aquela sequência é suficientemente semelhante com as restantes para poder ser considerada um gene *core* – mesmo que as restantes sequências obtenham alinhamentos com percentagens de identidade ligeiramente inferiores.

Devido às limitações encontradas e tendo em conta a complexidade de tentar juntar os vários conjuntos de resultados num único conjunto de resultados final, optou-se por tentar encontrar outra solução, pelo que se desenvolveu a terceira versão – que é também a versão final apresentada neste estudo no capítulo 2.2, e para a qual são apresentados os resultados dos dados analisados nos capítulos 3 e 4.

2.4 Alterações e melhorias da terceira versão

Na terceira versão, optou-se por comparar todos os genomas (ou proteomas, para facilitar as comparações de sequências), concatenados num só ficheiro, com a base de dados com todas as sequências de todos os genomas. Deste modo, é produzido apenas um ficheiro de resultados do BLAST, a partir do qual os resultados são filtrados consoante o que é requerido para a análise em questão, e no final obtemos um único conjunto de resultados, facilitando a análise e compreensão dos mesmos. De um modo geral, podemos dizer que a terceira versão consegue resolver os problemas anteriormente enunciados:

- Obtenção das sequências originais completas, em vez de apenas a parte que foi alinhada com o BLAST;

- Obtenção de um conjunto de resultados único, independente da ordem porque os ficheiros são adicionados à análise e portanto da ordem por que as comparações de sequências são efetuadas;

- Obtenção de apenas um conjunto de resultados final, para o mesmo conjunto de genomas em análise, em vez de um conjunto de resultados por cada genoma que teria de ser novamente analisado de forma a fazer os resultados convergir – portanto, menos ficheiros de resultados, maior facilidade em trabalhar com os mesmos.

De notar que nesta terceira versão optou-se por obter todos os genes *core* encontrados pelo menos uma vez – o que corresponde ao método menos restritivo apresentado na segunda versão – uma vez que não faz sentido ser muito restritivo quando não trabalhamos com percentagens de 100% de identidade, e como explicado anteriormente, para um gene ser encontrado, todas as sequências do CVAP têm de ter uma percentagem de identidade superior à percentagem escolhida com a sequência de interrogação.

Foram ainda realizadas outras melhorias na terceira versão, relativamente às versões anteriores, como resolução de pequenos erros não detetados anteriormente, melhoria e otimização do código, menos *scripts*, mais concisos e com código menos redundante e ainda a paralelização do código, para os processos serem divididos pelo número de cores do CPU. Também foram efetuadas posteriormente alterações ao nível da otimização do algoritmo para permitir realizar mais eficientemente análises para os mesmos dados, mas usando conjuntos de parâmetros diferentes, uma vez que inicialmente eram repetidos todos os passos de comparação de sequências com o BLAST e filtração dos resultados. Após estas alterações passou a ser possível repetir a análise para outra percentagem de identidade sem repetir o BLAST, ou para outra percentagem de diferença de tamanho (para a mesma percentagem de identidade) utilizando os ficheiros relativos aos CVAPs antes da exclusão por tamanho, que também são guardados. Estas melhorias permitem poupar um tempo considerável, quando se pretende analisar um conjunto de dados testando vários parâmetros.

Foi ainda desenvolvido um *script* adicional ("verify_clusters.py") que permite verificar que num conjunto de resultados (genoma *core*, genoma acessório, ou ambos, para as mesmas percentagens de identidade e de diferença de tamanho), cada sequência só se encontra num CVAP. Isto permite garantir a integridade dos resultados e que os CVAPs identificados estão bem definidos.

Uma vez que se confirmou este fator para os conjuntos de dados testados, é possível afirmar que o método é robusto na obtenção dos CVAPs.

Assim, esta versão satisfaz todos os requisitos que se verificou serem necessários cumprir aquando das implementações das duas versões anteriores. Nos próximos capítulos serão apresentados os resultados e discussão dos mesmos da utilização desta ferramenta para proceder à análise de um conjunto e vários sub-conjuntos de genomas de *Streptococcus pneumoniae*.

Capítulo 3

Análise do pangenoma de *S. pneumoniae*

3.1 Conjuntos de dados utilizados

De modo a obter o genoma *core* e o genoma acessório de *Streptococcus pneumoniae*, foram analisados 76 genomas desta espécie, utilizando a ferramenta desenvolvida. Com base nos resultados obtidos foram ainda efetuadas comparações entre serótipos da mesma espécie, cujos resultados se apresentam no capítulo seguinte. No total, foram utilizados genomas de 76 estirpes, dos quais 27 foram obtidos através da base de dados do GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) e 49 foram sequenciados pela Unidade de Microbiologia Molecular e Infeção do Instituto de Medicina Molecular.

Dos 27 genomas obtidos através do GenBank, para dois deles (estirpes SPN032672 e SPN033038) não existiam na base de dados os ficheiros .ffn ou .faa, que contém as várias regiões codificantes encontradas no genoma ou proteínas por estas codificadas, respectivamente, mas apenas os ficheiros contendo o genoma completo, que apresentam a extensão .fna (*Fasta Nucleic Acid*). Deste modo, foi necessário utilizar o Prodigal (*Prokaryotic Dynamic Programming Genefinding Algorithm*) [20] para obter as regiões codificantes e respetivas proteínas para estes dois genomas – ou seja, obter os ficheiros com extensão .ffn e .faa, que são utilizados pelo SCRAG. Também para os genomas sequenciados no Instituto de Medicina Molecular foi necessário obter as regiões codificantes e respectivas proteínas, pelo que também se correu o Prodigal para estes 49 genomas, uma vez que apenas tinham sido gerados previamente os ficheiros FASTA correspondentes aos vários *contigs*, obtidos com o programa de assemblagem de genomas SPAdes [21]. Para correr o

Prodigal (versão 2.6.1) para todos os genomas em que era necessário obter os ficheiros .ffn e .faa, desenvolveu-se um *script* em Python, que lê um directório com ficheiros, e corre o Prodigal para esses ficheiros, com os parâmetros necessários, gerando os novos ficheiros em duas novas pastas correspondentes aos ficheiros de ácidos nucleicos e de aminoácidos.

Para proceder a uma análise do pangenoma de *S. pneumoniae*, ou seja, para obter o genoma *core* e genoma acessório da espécie, começou-se por testar uma amostra dos genomas das 25 estirpes cujos ficheiros .ffn e .faa foram retirados do GenBank. Só posteriormente se obtiveram as regiões codificantes e proteínas correspondentes para os restantes dois genomas do GenBank e para os 49 genomas sequenciados pelo Instituto de Medicina Molecular. Na tabela 3.1 encontram-se listadas as 27 estirpes cujos genomas ou proteomas foram descarregados da base de dados do GenBank, disponíveis em ftp://*ftp.ncbi.nih.gov/genbank/genomes/Bacteria/*, bem como a identificação dos respetivos serótipos.

Posteriormente procedeu-se a uma comparação entre o serótipo 1 e serótipo 3 de *S. pneumoniae*, cujos resultados se apresentam no capítulo 4. Dos 49 genomas sequenciados no Instituto de Medicina Molecular, 24 pertenciam ao serótipo 1 e 25 ao serótipo 3. Também se identificaram, dos genomas descarregados do GenBank, 5 estirpes pertencentes ao serótipo 1 e 6 pertencentes ao serótipo 3 [5, 22, 23, 24, 25], conforme se pode observar na tabela 3.1. Obtemos então, de um modo geral, cinco conjuntos de dados diferentes, cujo genoma *core* e genoma acessório foram obtidos com o SCRAG:

- 25 genomas obtidos do GenBank (excluíram-se as duas estirpes para que apenas se descarregaram os ficheiros .fna: SPN032672 e SPN033038);
- 76 genomas (total)
- 29 genomas do serótipo 1;
- 31 genomas do serótipo 3;
- 16 genomas de outros serótipos ("outros").

Os dois primeiros conjuntos de dados (25 genomas e 76 genomas) foram utilizados para obter o genoma *core* e acessório de *S. pneumoniae*, e assim testar também a ferramenta desenvolvida – e cujos resultados são apresentados em seguida – ao passo que os três restantes foram utilizados para comparar geneticamente os serótipos 1, 3 e "outros", conforme explicado no capítulo 4.

Estirpe	**Número de acesso**	**Serótipo**
Streptococcus pneumoniae 670-6B	CP002176.1	6B
Streptococcus pneumoniae 70585	CP000918.1	5
Streptococcus pneumoniae A026	CP006844.1	19F
Streptococcus pneumoniae AP200	CP002121.1	11A
Streptococcus pneumoniae ATCC 700669	FM211187.1	23F
Streptococcus pneumoniae CGSP14	CP001033.1	14
Streptococcus pneumoniae D39	CP000410.1	2
Streptococcus pneumoniae G54	CP001015.1	19F
Sreptococcus pneumoniae Hungary19A-6	CP000936.1	19A
Streptococcus pneumoniae INV104	FQ312030.1	1
Streptococcus pneumoniae INV200	FQ312029.1	14
Streptococcus pneumoniae JJA	CP000919.1	14
Streptococcus pneumoniae OXC141	FQ312027.1	3
Streptococcus pneumoniae P1031	CP000920.1	1
Sreptococcus pneumoniae R6	AE007317.1	2
Streptococcus pneumoniae SPN032672	FQ312039.1	1
Streptococcus pneumoniae SPN033038	FQ312042.1	1
Streptococcus pneumoniae SPN034156	FQ312045.1	3
Streptococcus pneumoniae SPN034183	FQ312043.1	3
Streptococcus pneumoniae SPN994038	FQ312041.1	3
Streptococcus pneumoniae SPN994039	FQ312044.2	3
Streptococcus pneumoniae SPNA45	HE983624.1	3
Streptococcus pneumoniae ST556	CP003357.1	19F
Streptococcus pneumoniae TCH8431/19A	CP001993.1	19A
Streptococcus pneumoniae TIGR4	AE005672.3	4
Streptococcus pneumoniae Taiwan19F-14	CP000921.1	19F
Streptococcus pneumoniae gamPNI0373	CP001845.1	1

Tabela 3.1: Lista de genomas/proteomas descarregados do GenBank

3.2 Pangenoma de *S. pneumoniae* – 25 estirpes

3.2.1 Genoma *core*

Utilizando 25 dos 27 genomas extraídos do GenBank (excluindo SPN032672 e SPN033038) foi possível utilizar o SCRAG para encontrar o genoma *core* e o genoma acessório dessas 25 estirpes. Foram testadas várias percentagens de identidade – 70%, 80%, 90% e 100% – e várias percentagens de diferença de tamanho máxima permitida entre as sequências de um CVAP – 0%, 10%, 20% e 30% – sendo que também se obtêm os ficheiros relativos a não se realizar a exclusão por tamanho, ou seja, todas as diferenças de tamanho são permitidas. Também foi verificada a integridade dos resultados e robustez do método, verificando que cada sequência apenas é atribuída a um CVAP, conforme referido no capítulo 2 (2.4). Os resultados obtidos, relativos ao número de genes *core* encontrados para estes 25 genomas, de acordo com as várias percentagens testadas, podem ser observados na figura 3.1.

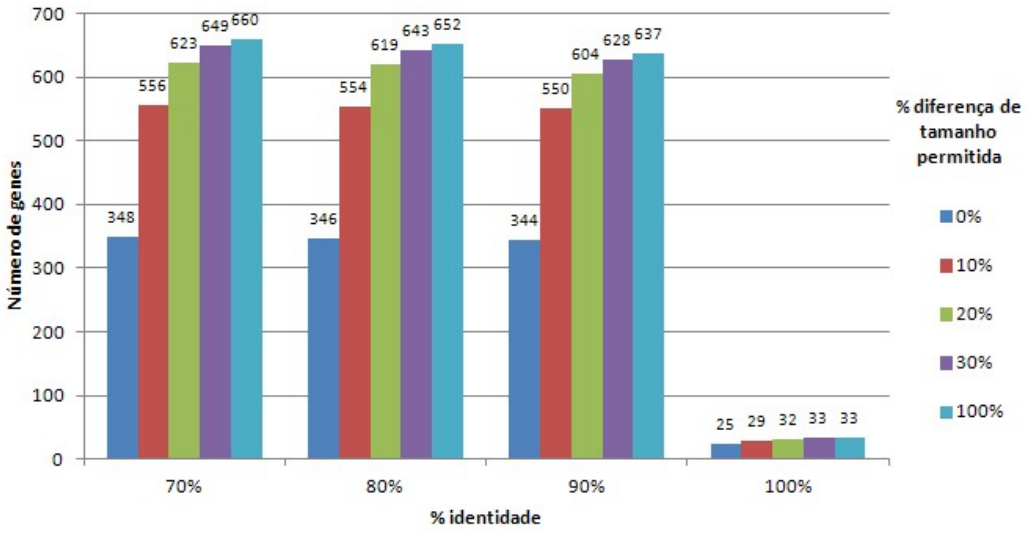

Figura 3.1: Resultados para o genoma *core* de *S. pneumoniae*, utilizando 25 genomas. Foram testados vários parâmetros de percentagem de identidade e diferença de tamanho permitida. Os resultados para a percentagem de diferença de tamanho de 100% correspondem a não ter efetuado o passo da exclusão por tamanho.

Pode verificar-se que são encontrados mais genes considerando percen-

tagens de identidade mais baixas e percentagens de diferença de tamanho mais altas. Ou seja, quanto mais restritivos forem os parâmetros, considerando sequências muito semelhantes e pequenas diferenças de tamanho, menos genes vão ser encontrados, como seria de esperar. Para percentagens de identidade de 100%, os números decrescem abruptamente, uma vez que se consideram apenas os casos em que o BLAST atribui uma correspondência perfeita em todos os 25 alinhamentos correspondentes a cada sequência de um CVAP pertencente ao genoma *core*. Também se verifica que para percentagens de diferença de tamanho de 0%, o número de genes encontrados diminui bastante, ao passo que considerar uma diferença de tamanho de 30% ou não efetuar a exclusão por tamanho de todo (100%) não apresenta grandes diferenças – ou seja, são poucos os genes encontrados, para uma dada percentagem de identidade, em que existem diferenças de tamanho superiores a 30% entre as sequências de um mesmo CVAP.

Tendo em conta os resultados obtidos para as percentagens testadas, verifica-se que os valores mais intermédios são obtidos para 80% de identidade e 20% de diferença de tamanho. Abaixo é apresentado o gráfico de pontos gerado pelo programa considerando essas percentagens (figura 3.2). O gráfico para uma análise ao genoma *core*, utilizando os mesmos 25 genomas, mas considerando como parâmetros 70% de identidade e 30% de diferença de tamanho encontra-se no capítulo 2 (figura 2.3).

De notar que os gráficos de pontos são obtidos através do cálculo do mínimo, média e desvio padrão da percentagem de similaridade entre as sequências de cada CVAP, sendo que os valores de similaridade correspondem por sua vez ao inverso das distâncias, obtidas da matriz de distâncias gerada com o ClustalW, através do alinhamento múltiplo realizado com o MUSCLE. Deste modo, e tendo em conta também que nesta fase são recuperadas as sequências de ADN, completas, os valores de percentagem de similaridade não são exatamente iguais aos valores de percentagem de identidade obtidos originalmente, embora possam ser aproximados. Os valores de percentagem de identidade, por sua vez, foram calculados através dos valores de identidade (ou seja, as correspondências exatas) fornecidos pelo BLAST para cada alinhamento, onde foram utilizadas sequências de aminoácidos. De notar também que um alinhamento do BLAST (alinhamentos locais) pode não corresponder à sequência completa, ao contrário do alinhamento múltiplo de sequências, que considera as sequências globalmente, e não apenas parcialmente.

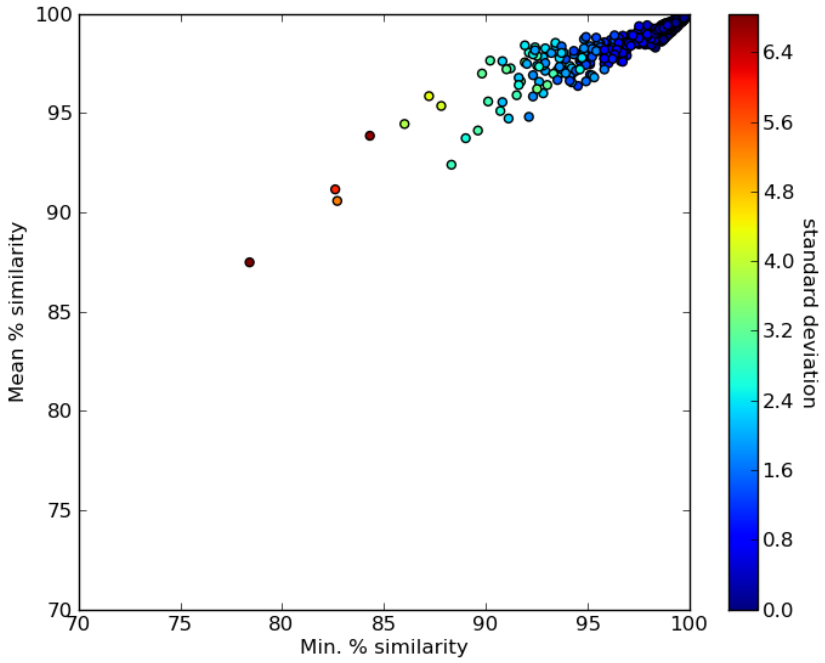

Figura 3.2: Gráfico obtido para uma análise do genoma *core*, utilizando 25 genomas de *S. pneumoniae*, considerado como parâmetros para obter cada CVAP 80% de identidade e 20% de diferença de tamanho máxima permitida.

3.2.2 Genoma acessório

Também para o genoma acessório dos 25 genomas foi utilizado o programa, para as mesmas percentagens de identidade e de diferença de tamanho. Os resultados obtidos podem ser observados na figura 3.3. Mais uma vez, verifica-se que parâmetros mais restritivos levam a um menor número de CVAPs encontrados, ao passo que percentagens de identidade inferiores e percentagens de diferença de tamanho maiores levam à descoberta de mais genes acessórios. É também apresentado o gráfico de pontos gerado, considerando como parâmetros 80% de identidade e 20% de diferença de tamanho (figura 3.4).

Analisando os resultados, é possível observar que são encontrados mais genes acessórios do que genes *core*. Tal facto seria de esperar, uma vez que o número de genes *core* descobertos diminui com o número de genomas adicionados à análise. Já o número de genes acessórios vai aumentando quantos mais genomas são analisados – são descobertos mais CVAPs que não estão

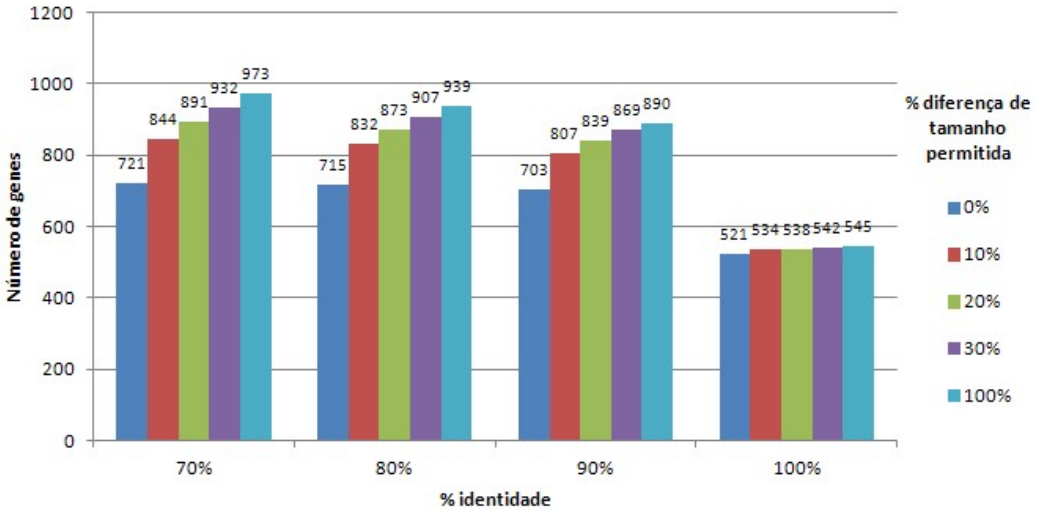

Figura 3.3: Resultados para o genoma acessório de *S. pneumoniae*, utilizando 25 genomas. Foram testados vários parâmetros de percentagem de identidade e diferença de tamanho permitida. Os resultados para a percentagem de diferença de tamanho de 100% correspondem a não ter efetuado o passo da exclusão por tamanho.

representados em todos os genomas, bem como genes únicos, característicos de cada genoma, sendo que o tamanho do pangenoma também aumenta com o número de genomas [6, 8, 9]. Assim sendo, repetindo esta análise para apenas alguns dos 25 genomas considerados, seria de esperar encontrar mais genes *core* e menos genes acessórios do que para os 25 genomas. Da mesma forma, considerando mais genomas do que os 25, como acontece na análise aos 76 genomas, apresentada na secção seguinte, são encontrados menos genes *core* e mais genes acessórios do que para os 25 genomas cujos resultados aqui são apresentados, como veremos adiante.

3.3 Pangenoma de *S. pneumoniae* – 76 estirpes

3.3.1 Genoma *core*

Também se utilizou o SCRAG no conjunto de 76 genomas de *S. pneumoniae*, considerando diversas percentagens de identidade e diferença de ta-

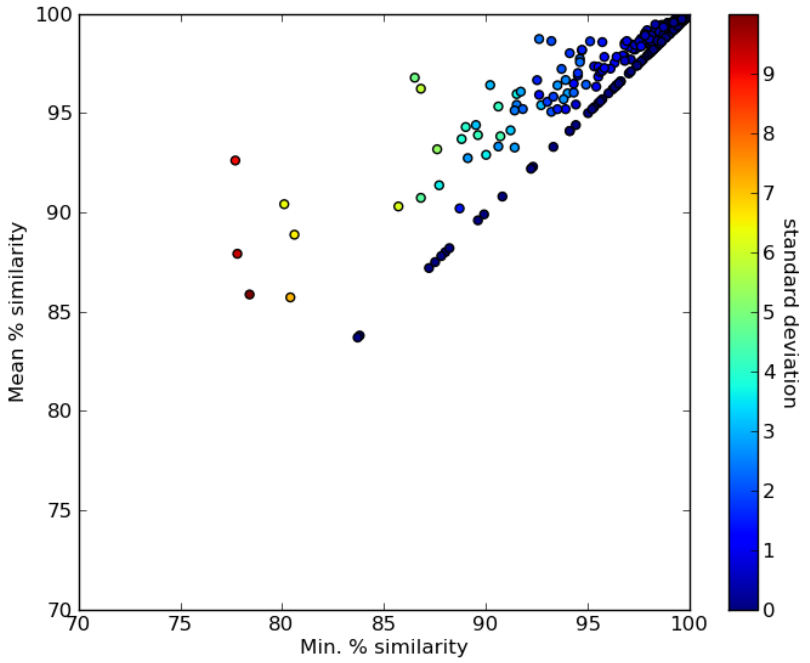

Figura 3.4: Gráfico obtido para uma análise do genoma acessório, utilizando 25 genomas de *S. pneumoniae*, considerado como parâmetros para obter cada CVAP 80% de identidade e 20% de diferença de tamanho máxima permitida.

manho. Para o genoma *core*, obtiveram-se menos genes do que quando se utilizaram apenas 25 genomas, como era esperado, conforme explicado anteriormente. Na figura 3.5 podem observar-se os resultados obtidos.

Mais uma vez é possível verificar que quanto maior a percentagem de identidade e quanto menor a percentagem de diferença de tamanho consideradas – considerando portanto parâmetros mais restritivos – menos CVAPs são encontrados. Para 100% de identidade, são encontrados poucos CVAPs *core*, uma vez que será difícil obter uma correspondência perfeita para todas as 76 sequências. Também a maior diferença no número de CVAPs encontrados, tendo em conta as percentagens de diferença de tamanho, é para percentagens de 0%, para as quais se encontram muito menos CVAPs que para as restantes percentagens testadas. Observa-se ainda que à medida que se vai aumentando a percentagem de diferença de tamanho, menos vão ser as diferenças relativamente ao número de CVAPs encontrados. Considerando os valores mais intermédios, obtidos para 80% de identidade e 20% de diferença de tamanho, é apresentado na figura 3.6 o gráfico de pontos gerado,

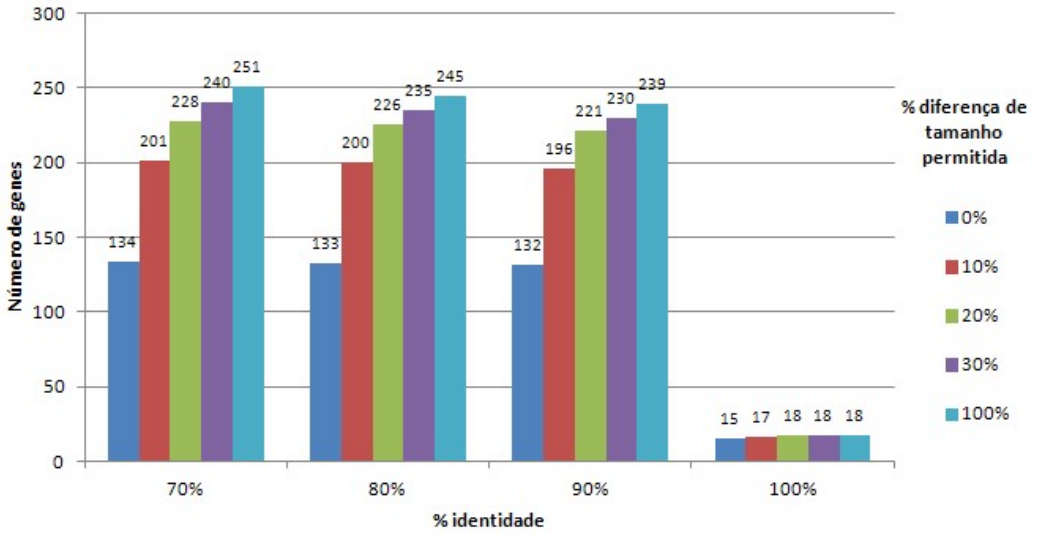

Figura 3.5: Resultados para o genoma *core* de *S. pneumoniae*, utilizando 76 genomas. Foram testados vários parâmetros de percentagem de identidade e diferença de tamanho permitida. Os resultados para a percentagem de diferença de tamanho de 100% correspondem a não ter efetuado o passo da exclusão por tamanho.

considerando estes parâmetros.

3.3.2 Genoma acessório

Para o genoma acessório, considerando 76 genomas, e para as diferentes percentagens de identidade e de diferença de tamanho utilizadas, encontra-se um maior número de CVAPs do que quando se utilizaram apenas 25 genomas, conforme se pode observar na figura 3.7. Isto vai de encontro aos resultados esperados, como referido acima.

Também se obteve o gráfico de pontos para os parâmetros de 80% de identidade e 20% de diferença de tamanho (figura 3.8).

3.4 Tempos de corrida do algoritmo

Relativamente aos tempos de corrida do algoritmo, verificou-se que o processo é mais demorado quando é necessário efetuar o passo de interrogação a base de dados do BLAST, sendo que a duração deste processo aumenta

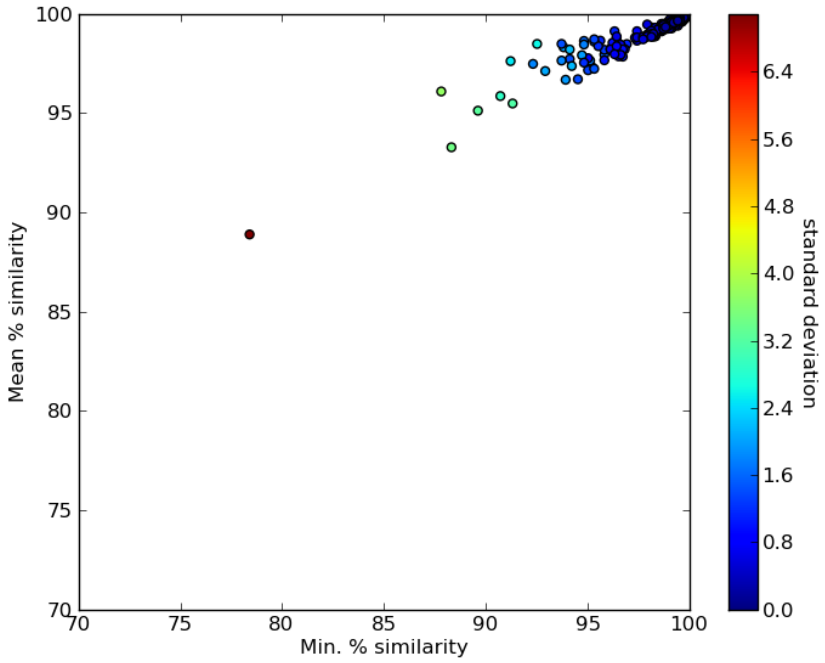

Figura 3.6: Gráfico obtido para uma análise do genoma *core*, utilizando 76 genomas de *S. pneumoniae*, considerado como parâmetros para obter cada CVAP 80% de identidade e 20% de diferença de tamanho máxima permitida.

consideravelmente com o número de genomas em análise. Para os 29 genomas do serótipo 1, por exemplo, o SCRAG demorou cerca de 8:09 horas a obter os resultados pretendidos, considerando uma análise do genoma core, 80% de identidade e 20% de diferença de tamanho. Para o serótipo 1 + serótipo 3 (60 genomas), o tempo de corrida aumenta para 29:53 horas para os mesmos parâmetros, cerca de 2.7 vezes mais que no caso anterior, apesar de número de genomas ter aumentado para pouco mais do dobro. Quando o ficheiro de resultados do BLAST já é existente para um dado conjunto de dados, o tempo de corrida reduz consideravelmente. Para o mesmo conjunto de dados de 29 genomas do serótipo 1, mas procedendo a uma análise do genoma acessório, para os mesmos parâmetros de identidade e diferença de tamanho, e desta vez lendo o ficheiro de resultados do BLAST já existente, o SCRAG demora apenas cerca de 1:24 horas a obter os resultados. Já quando se repete a análise para diferentes percentagens de diferença de tamanho, para uma mesma percentagem de identidade, a obtenção dos resultados é quase imediata, uma vez que se utilizam os ficheiros relativos aos CVAPs

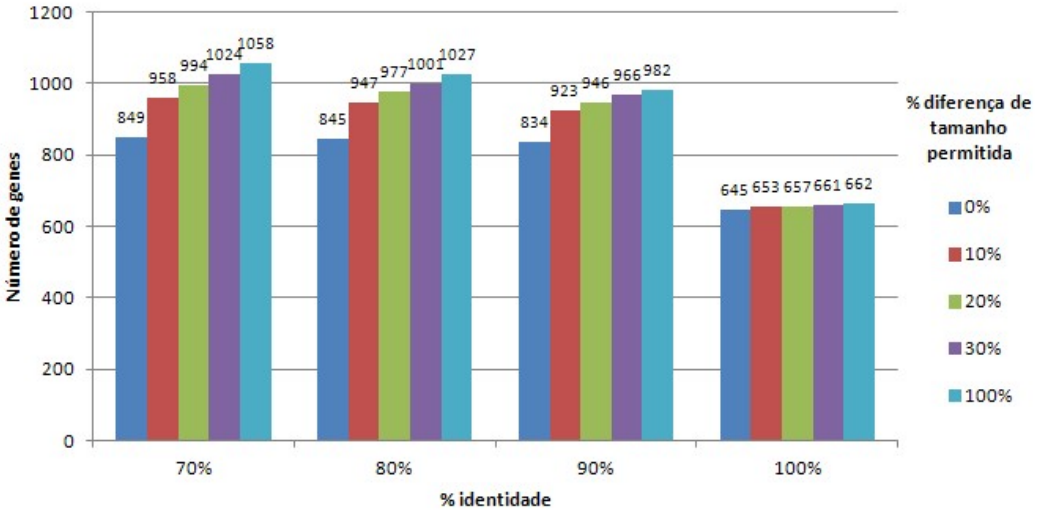

Figura 3.7: Resultados para o genoma acessório de *S. pneumoniae*, utilizando 76 genomas. Foram testados vários parâmetros de percentagem de identidade e diferença de tamanho permitida. Os resultados para a percentagem de diferença de tamanho de 100% correspondem a não ter efetuado o passo da exclusão por tamanho.

sem exclusão por tamanho. Neste caso o algoritmo demora apenas cerca 2 minutos, mesmo quando se consideraram os 76 genomas de *S. pneumoniae*.

3.5 Discussão e conclusões

Utilizando o SCRAG foi possível obter o genoma *core* e o genoma acessório para estirpes de *S. pneumoniae*, conforme referido acima. Os resultados vão de encontro ao esperado, obtendo-se menos CVAPs quanto mais restritivos são os parâmetros, sendo que que os resultados obtidos dependem portanto dos parâmetros considerados. No entanto, o método utilizado demonstra ser também bastante restritivo, pois somando o número de genes *core* com o número de genes acessórios obtêm-se menos genes do que os que são encontrados normalmente em cada uma das estirpes (cerca de 2000). Este facto pode ser devido à ocorrência de parálogos.

É necessário também ter em conta que a definição de genoma *core* (conjunto de genes encontrados em todos os genomas) e de genoma acessório (conjunto de genes que não são encontrados em todos os genomas) [6, 8] é bastante restritiva, se considerarmos que um "mesmo gene"se refere a uma

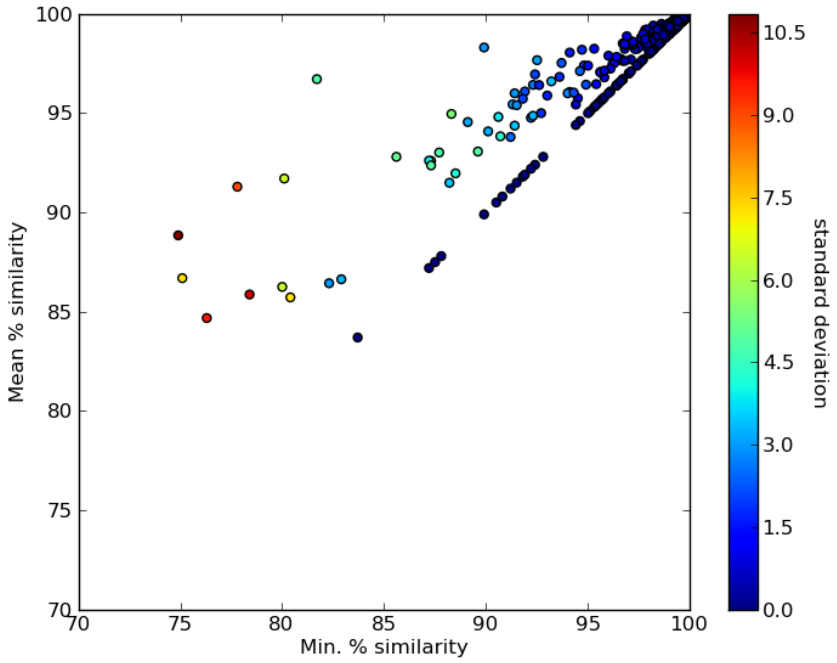

Figura 3.8: Gráfico obtido para uma análise do genoma *core*, utilizando 76 genomas de *S. pneumoniae*, considerado como parâmetros para obter cada CVAP 80% de identidade e 20% de diferença de tamanho máxima permitida.

sequência exatamente idêntica a outra. Desta forma, é importante considerar qual o grau de semelhança adequado quando se procede à identificação dos CVAPs, ou seja, quais os parâmetros adequados para identificar alelos que codificam para um mesmo *locus*, isto é, que representam "o mesmo"gene (embora possa não ser totalmente idêntico). Observando os resultados obtidos, é possível perceber que para percentagens de diferença de tamanho superiores a 30%, a diferença no número de CVAPs encontrados não vai ser muita, relativamente a percentagens de 30% ou mesmo 20%, uma vez que quando se ignora estas diferenças de tamanho (100%) o número de genes encontrados não aumenta muito, para as percentagens de identidade testadas. Já para as diferentes percentagens de identidade consideradas, a grande diferença reside nos 100% (o BLAST gera alinhamentos com correspondência perfeita), relativamente às percentagens de 90%, 80% e mesmo 70%, cujos valores se revelam bastante aproximados entre si (para as mesmas percentagens de diferença de tamanho). De notar que durante o processo de obtenção dos CVAPs, é também verificado o Blast Score Ratio (BSR) do alinhamento,

e excluídos os resultados em que BSR é inferior a 0.6 – valor que corresponde a cerca de 80% de similaridade. Este aspeto só por si evita a ocorrência de CVAPs com grandes diferenças a nível da identidade e similaridade, e é uma possível explicação para os valores não aumentarem grandemente quando consideradas percentagens de 70% de identidade ou inferiores.

Relativamente ao genoma *core* e genoma acessório, para ambos os conjuntos de dados se obtiveram mais genes acessórios do que genes *core*, o que vai de encontro aos resultados obtidos em estudos realizados anteriormente [6, 7, 8]. Seria esperado que à medida que fossem adicionados mais e mais genomas à análise, mais genes acessórios fossem encontrados, uma vez que *S. pneumoniae* apresenta um pangenoma aberto [8]. Seria também de esperar que o tamanho do pangenoma (genoma *core* e genoma acessório) também aumentasse, à medida que o genoma acessório aumenta. No entanto, tal facto não se verifica, uma vez que, apesar de o genoma acessório aumentar, o genoma *core* diminui bastante para os 76 genomas, relativamente à análise com apenas 25 genomas. Isto poderá ser devido aos parâmetros e restrições consideradas na análise com o SCRAG. O SCRAG (Strict CoRe and Accessory Genome), como o nome indica, permite-nos obter o genoma *core* e o genoma acessório estritos, efetuando múltiplas verificações aos resultados obtidos pelo BLAST, e filtrando os mesmos de uma forma bastante rigorosa, o que torna também o método bastante robusto. No entanto, como é em primeira instância verificado o número de alinhamentos obtidos para cada sequência na base de dados (sendo logo excluídos os casos em que há mais alinhamentos para a sequência de interrogação do que o número de genomas em análise), e depois verificados os vários parâmetros, é possível que muitos CVAPs sejam descartados devido a fenómenos como duplicação de genes (são considerados apenas os CVAPs em que todas as sequência são procedentes de estirpes diferentes) ou obtenção de alinhamentos com pontuação mais baixa, que originam percentagens de identidade ou valores de BSR inferiores aos valores considerados, mesmo que outros alinhamentos para a mesma sequência de interrogação pudessem passar na verificação destes parâmetros. Deste modo, quantos mais genomas são considerados, mais genes *core* são perdidos, e apesar de haver um aumento no número de genes acessórios, muitos são também perdidos por não passarem na verificação de todos os parâmetros: é possível observar que o número de genes acessórios não aumenta tanto quanto seria esperado. Deste modo, o SCRAG será mais adequando para realizar análises mas estritas e robustas, evitando a ocorrência de "falsos positivos", ou seja, evitando a identificação e classificação de genes como *core* ou acessórios quando não passam em algum dos parâmetros considerados, sendo que por outro lado os CVAPs obtidos podem ser considerados com bastante certeza como *core* ou acessórios.

Verifica-se ainda que os CVAPs obtidos para o genoma acessório apresentam uma maior dispersão a nível de similaridade do que os CVAPs relativos ao genoma *core*, de acordo com os gráficos de pontos gerados. Isto pode mais uma vez ser explicado pelo caráter restritivo do método utilizado, sendo que quanto mais alinhamentos obtidos para uma mesma sequência de interrogação, mais difícil será todos eles passarem na verificação dos parâmetros. Assim, será provável que os CVAPs classificados como *core* apresentem sequências bastante idênticas entre si, o que se reflete numa menor dispersão dos pontos no gráfico. Já para o genoma acessório, será provável que os CVAPs com menos sequências se encontrem na zona superior direita do gráfico, enquanto que os CVAPs maiores podem representar pontos mais dispersos, aproximando-se da parte inferior esquerda do gráfico. Também é importante notar que quando o gráfico é gerado, são consideradas as percentagens de similaridade obtidas através da matriz de distâncias obtida com o ClustalW, que por sua vez é gerada através do alinhamento múltiplo realizado com o MUSCLE, e também que são utilizadas nesta fase sequências de ADN. Tendo em conta a redundância do código genético (codões diferentes podem codificar para aminoácidos iguais), e que são recuperadas as sequências completas (BLAST só retorna a parte que alinhou e os valores de identidade e pontuação utilizados inicialmente, aquando da filtração dos resultados, só se referem a estas porções), bem como as diferenças entre identidade e similaridade, é de esperar que os valores representados no gráfico não correspondam exatamente aos valores de identidade calculados na primeira fase da análise. Isto explica também a ocorrência de alguns pontos em percentagem de similaridade abaixo da percentagem de identidade considerada na realização da análise.

Assim, podemos considerar o SCRAG um método conservador para a obtenção do genoma *core* e acessório de uma espécie bacteriana, sendo que os resultados obtidos podem ter muitas aplicações para outros estudos, como a análise das funções atribuídas ao genes *core* e aos genes acessórios, geração de árvores filogenéticas, ou mesmo a comparação de serótipos de uma espécie bacteriana – tendo este último caso sido testado e apresentado no capítulo 4.

Capítulo 4

Comparação dos serótipos 1 e 3 de *S. pneumoniae*

4.1 Método e objetivos

Neste capítulo pretende-se efetuar uma comparação do número de genes partilhados – considerando tanto genes *core* como genes acessórios – entre estirpes de *Streptococcus pneumoniae* do serótipo 1, do serótipo 3 e de outros serótipos, utilizando o total dos 76 genomas analisados no capítulo anterior. É possível assim utilizar o SCRAG para obter o genoma *core* e o genoma acessório para estes três grupos (serótipo 1, serótipo 3 e outros serótipos) e para as combinações possíveis entre si, e desta forma gerar um diagrama de Venn onde se obtêm as intersecções entre os conjuntos de dados, conforme se pode observar na figura 4.1.

No total, para a análise pretendida, precisamos de obter o genoma *core* e o genoma acessório para os três grupos (serótipo 1, serótipo 3, outros serótipos), para cada par de conjuntos (serótipo 1 + serótipo 3, serótipo 1 + outros serótipos, serótipo 3 + outros serótipos) e para os três grupos (genoma *core* da espécie, quando são utilizados os genes *core*). Desta forma, e conforme explicado na figura 4.1, pretende-se obter os CVAPs que estão em:

- Todos os 76 genomas de todos os serótipos (serotipo 1 + serótipo 3 + outros serótipos);
- Serótipo 1 (29 genomas) apenas;
- Serótipo 3 (31 genomas) apenas;
- Outros serótipos (16 genomas) apenas;

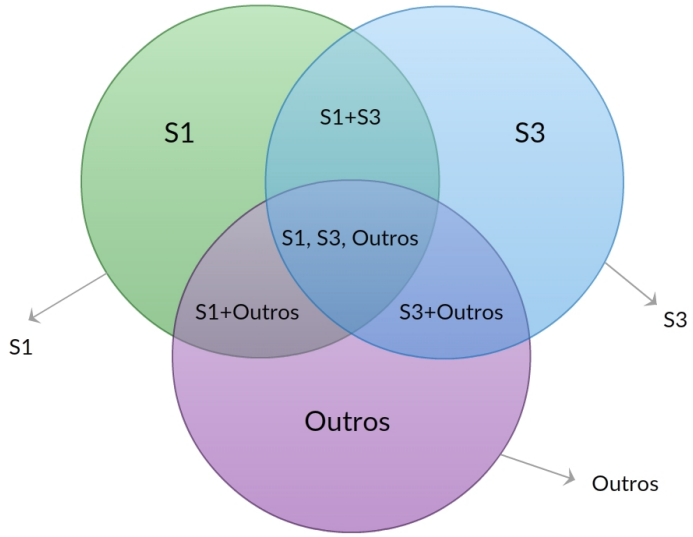

Figura 4.1: Diagrama de Venn explicativo das comparações genómicas efetuadas entre serótipos de *S. pneumoniae*. Cada círculo representa um dos conjuntos de dados em análise. As intersecções representam o que é comum aos conjuntos de dados visados. "S1"= serótipo 1, "S3"= serótipo 3, "Outros"= outros serótipos.

- Serótipo 1 + serótipo 3 (60 genomas), mas não em outros serótipos;
- Serótipo 1 + outros serótipos (45 genomas), mas não no serótipo 3;
- Serótipo 3 + outros serótipos (47 genomas), mas não no serótipo 1.

Efetuaram-se as comparações dos CVAPs presentes em cada conjunto de dados utilizando um *script* desenvolvido para o efeito, designado por "compare_sero.py". Este *script* lê ficheiros de dois ou três diretórios e compara o seu conteúdo. Uma vez que pretendemos comparar CVAPs com tamanhos diferentes (quanto mais genomas no conjunto de dados considerado, mais sequências têm os respetivos CVAPs, quando se trata de genes *core*), é necessário verificar se os CVAPs mais pequenos, com menos sequências, estão contidos nos CVAPs maiores, ou seja, se todas as sequências de um CVAP (o mais pequeno) de um conjunto de dados são as mesmas que estão no outro CVAP (o maior) do outro conjunto de dados. Se isto se verificar, os CVAPs não são exclusivos do conjunto de dados menor (menos genomas, menos sequências por CVAP, para os genes *core*), pelo que os que são exclusivos são os que só estão no conjunto de dados menor. Por exemplo, para obter os genes comuns ao serótipo 1 e ao serótipo 3, mas que não estão nos

outros serótipos, considerando os genes *core*, utiliza-se o conjunto de genes do serótipo 1 + serótipo 3 (CVAPs menores) e o conjunto de genes *core* dos 76 genomas (CVAPs maiores), e verifica-se quais os genes que estão no serótipo 1 e serótipo 3, mas não estão no total dos 76 genomas (ou seja, que estão também em outros serótipos). Deste modo, é possível comparar todos os conjuntos de dados em análise e obter as relações genómicas entre eles.

Também é possível fazer estas comparações utilizando o genoma acessório de cada um dos conjuntos de dados – ou seja, os CVAPs que não contém sequências de todos os genomas desse conjunto de dados – mas nesse caso é necessário ter em conta que os CVAPs não apresentam um tamanho fixo, e como tal não será correto comparar apenas um conjunto de dados relativo a menos genomas, e que para os genes *core* apresenta também CVAPs mais pequenos, com um conjunto de dados relativo a mais genomas, uma vez que neste caso este último poderá ter CVAPs mais pequenos que os do conjunto de dados relativo ao menor número de genomas. Assim, é necessário verificar qual o CVAP mais pequeno e só então proceder à comparação, verificando se o CVAP menor está contido no maior.

Abaixo são apresentados os resultados obtidos para esta análise.

4.2 Resultados

Utilizando o SCRAG, obteve-se o genoma *core* e o genoma acessório para os conjuntos de dados referidos acima, considerando como parâmetros de percentagem de identidade e percentagem de diferença de tamanho 80% e 20%, respetivamente, uma vez que, dos parâmetros testados anteriormente, parecem ser aqueles em que se obtém os resultados mais intermédios, não sendo demasiado restritivos nem demasiado permissivos. O número de CVAPs encontrados para os vários conjuntos de dados são apresentados na tabela 4.1.

4.2.1 Genoma *core*

Utilizando o *script* desenvolvido para o efeito foi possível obter o número de CVAPs específico de cada conjunto ou par de conjuntos de dados, ou seja, os CVAPs que não estão no(s) restante(s) conjuntos. Os valores obtidos para o genoma *core* são apresentados na figura 4.2. A explicação do diagrama é apresentada na figura 4.1.

Observando a figura 4.2 pode concluir-se que são encontrados mais genes *core* comuns ao serótipo 3 e outros serótipos, do que comuns ao serótipo 1 e a qualquer dos outros grupos. Isto parece indicar que o serótipo 1 é geneticamente mais divergente dos restantes, ao passo que o serótipo 3 é

Conjunto de dados	Número de genomas	Genoma *core*	Genoma acessório
S1	29	327	536
S3	31	707	437
Outros	16	670	797
S1 + S3	60	260	671
S1 + Outros	45	252	933
S3 + Outros	47	586	873
S1 + S3 + Outros (total)	76	226	977

Tabela 4.1: Número de CVAPs obtidos para o genoma *core* e genoma acessório dos conjuntos de dados em análise, considerando 80% de identidade e 20% de diferença de tamanho entre sequências do mesmo CVAP, e número de genomas em cada conjunto de dados. "S1"= serótipo 1, "S3"= serótipo 3, "Outros"= outros serótipos.

bastante idêntico geneticamente ao grupo dos outros serótipos. No entanto, relativamente ao número de genes *core*, o serótipo 1 também é aquele que apresenta menos CVAPs – cerca de 38% do total de genes (pangenoma) – enquanto que o serótipo 3 é o que apresenta mais CVAPs – cerca de 62% do total. Já o grupo de genomas relativo a "outros serótipos"apresenta cerca de 46% de genes *core*.

Assim, e tendo em conta o número de CVAPs *core* encontrados no total para cada um dos conjuntos de dados, podemos perceber que 13.0% dos CVAPs do serótipo 1 são específicos desse serótipo, bem como 13.1% do serótipo 3 e 10.0% de outros serótipos, aproximadamente. Quanto aos 37 CVAPs comuns ao serótipo 1 e 3, não encontrados em outros serótipos, estes representam 11.3% dos CVAPs *core* do serótipo 1, 5.2% do serótipo 3 e 14.2% do total dos genes *core* comuns ao serótipo 1 e serótipo 3. Os 28 genes partilhados em exclusivo entre o serótipo 1 e outros serótipos, por sua vez, correspondem a cerca de 8.6% do genoma *core* do serótipo 1, 4.2% do genoma *core* de outros serótipos e 11.1% dos genes *core* em comum. Já a relação entre o serótipo 3 e outros serótipos (361 CVAPs exclusivos destes dois grupos) situa-se em cerca de 51.0% e 53.9% dos genes *core* do serótipo 3 e de outros serótipos, respetivamente, e 61.6% do total dos genes *core* comuns aos dois grupos. Avaliando estas percentagens, é possível perceber que a relação entre o serótipo 3 e outros serótipos é de facto a mais forte, ao passo que serótipo 1 e outros serótipos são os grupos que apresentam menos em comum.

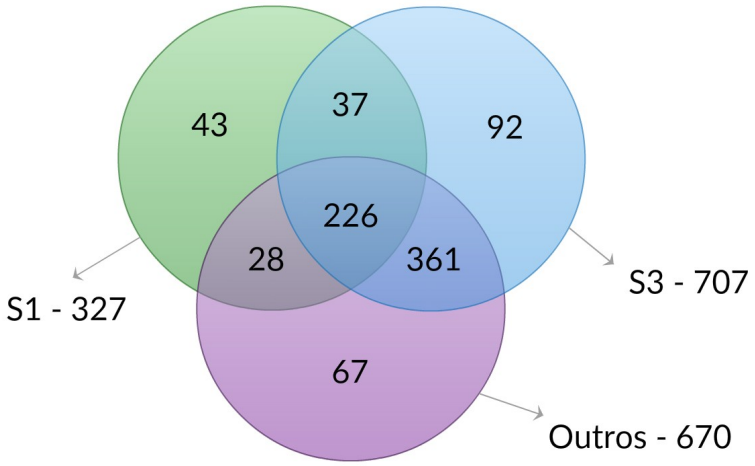

Figura 4.2: Relação entre conjuntos de dados utilizando genes *core*: serótipo 1, serótipo 3 e outros serótipos. Os números indicam o número de CVAPs encontrados para cada conjunto (cada um dos círculos), os CVAPs específicos de cada conjunto ou par de conjuntos e os CVAPs comuns aos três conjuntos de dados (região central), conforme explicado na figura 4.1. "S1"= serótipo 1, "S3"= serótipo 3, "Outros"= outros serótipos.

4.2.2 Genoma acessório

Utilizando o mesmo método, obteve-se mais uma vez o número de CVAPs específico de cada conjunto ou par de conjuntos de dados, mas desta vez relativamente ao genoma acessório. Os valores obtidos são apresentados na figura 4.3 e a explicação da mesma é apresentada na figura 4.1.

É importante ter em conta que os CVAPs encontrados relativos ao genoma acessório representam genes que são encontrados em todos os genomas do conjunto de dados em análise, podendo conter entre uma a N menos uma sequências, sendo N o número de genomas. Por exemplo, para o conjunto de dados do serótipo 1 + serótipo 3 (60 genomas), encontraram-se 671 CVAPs referentes a genes acessórios, estando presentes no mínimo em um dos genomas, e no máximo em 59 (ou seja tendo entre uma a 59 sequências; CVAPs com 60 sequências representam os genes *core*).

Tendo presente este aspeto, podemos então observar que o serótipo 1 e o serótipo 3 partilham um elevado número de genes acessórios (140) que não se encontram em outros serótipos – 20.9% do total de CVAPs partilhados entre os dois grupos. Este mesmo valor diz respeito a 26.1% dos CVAPs encontrados para o serótipo 1 e 32.0% dos CVAPs encontrados para o serótipo 3. Entre o serótipo 3 e outros serótipos, são encontrados 109 CVAPs exclusivos,

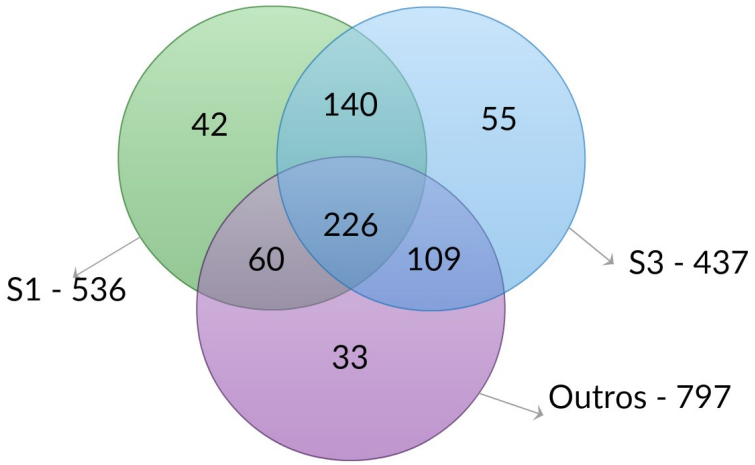

Figura 4.3: Relação entre conjuntos de dados utilizando genes acessórios: serótipo 1, serótipo 3 e outros serótipos. Os números indicam o número de CVAPs encontrados para cada conjunto (cada um dos círculos), os CVAPs específicos de cada conjunto ou par de conjuntos e os CVAPs comuns aos três conjuntos de dados (região central), conforme explicado na figura 4.1. "S1"= serótipo 1, "S3"= serótipo 3, "Outros"= outros serótipos.

12.5% do valor total, que representam 24.9% dos genes acessórios do serótipo 3 e 13.7% dos genes acessórios de outros serótipos. Já o serótipo 1 e outros serótipos são os grupos que apresentam a relação mais fraca, partilhando apenas 60 CVAPs que não são comuns ao serótipo 3 – apenas cerca de 6.4% do total de CVAPs do conjunto de dados – constituindo 11.2% do genoma acessório do serótipo 1 e 7.5% do genoma acessório de outros serótipos.

Quanto aos genes exclusivos de cada grupo, encontram-se 42 para o serótipo 1 (7.8%), 55 para o serótipo 3 (12.6%) e 33 para outros serótipos (4.1%). Pode assim verificar-se que o serótipo 3 é o que apresenta um genoma acessório exclusivo maior, e o grupo relativo a outros serótipos é o que apresenta o genoma acessório exclusivo menor. Verifica-se ainda que o serótipo 1 apresenta 62% de genes acessórios, o serótipo 3 38%, e outros serótipos 54%.

4.3 Discussão e conclusões

Observando os resultados obtidos podemos verificar primeiramente, relativamente ao número de genes *core* e acessórios encontrados para cada um dos conjuntos de dados analisados com o SCRAG, que para todos eles se

encontram mais genes acessórios do que genes *core*, com exceção do serótipo 3, que apresenta mais genes *core* (707, cerca de 62% do total de CVAPs encontrados para este grupo). Isto parece indicar que o serótipo 3 é mais conservado, mantendo muitos genes *core*, essenciais ao seu funcionamento e uma menor percentagem de genes acessórios.

Já o serótipo 1 apresenta apenas cerca de 38% de genes *core* e 62% de genes acessórios, sendo o que apresenta a maior percentagem de genes acessórios, o que indica um maior número de genes não presentes em todas as estirpes desse serótipo, podendo estar associados a fenómenos de adaptação de determinadas estirpes do mesmo. O serótipo 1 é também, dos três grupos considerados, o que apresenta menos genes no total (*core* + acessórios) – o que vai de encontro a estudos anteriores, que indicam uma limitada diversidade genética para este serótipo, bem como elevada similaridade intraserótipo [5, 15]. De notar que para o serótipo 1 foram identificadas três linhagens [14], mas os genomas deste serótipo obtidos pela Unidade de Microbiologia Molecular e Infeção pertencem todos à mesma linhagem, pelo que se fossem utilizados apenas estes genomas o número de genes *core* encontrados seria muito superior. No entanto, ao adicionar à análise os genomas obtidos da base de dados do GenBank, estamos provavelmente a considerar todas as linhagens, e assim são obtidos menos genes *core*, o que parece indicar que as linhagens do serótipo 1 terão divergido à mais tempo, e como tal apresentam um genoma *core* diferente entre si.

O grupo de outros serótipos é o que apresenta mais genes no total, e também mais genes acessórios – o que seria esperado, uma vez que estão representados vários serótipos, e como tal haverá mais variedade genética. De referir que os outros serótipos utilizados a que foi possível ter acesso ao genoma completo são apenas nove (serótipos 2, 4, 5, 6B, 11A, 14, 19A, 19F, 23F), não representando a totalidade dos serótipos de *Streotococcus pneumoniae*, uma vez que existem mais de 90 tipos diferentes [5, 13, 14]. No entanto, os serótipos que foram utilizados são os que se encontram disponíveis, por serem os mais relevantes clinicamente.

Também se pode observar mais uma vez que o número de genes acessórios aumenta com o número de genomas, ao passo que o número de genes *core* vai diminuindo quantos mais genomas são analisados, considerando os conjuntos de dados que são subconjuntos de outros. Assim, para os subconjuntos "S1 + S3", "S1 + Outros", "S3 + Outros", obtém-se mais genes *core* e menos genes acessórios do que para o total dos 76 genomas. Também para os subconjuntos do serótipo 1, serótipo 3 e "outros serótipos"se obtêm mais genes *core* e menos genes acessórios do que os respetivos pares de subconjuntos, conforme se pode observar na tabela 4.1.

Relativamente ao número de genes *core* e acessórios específicos de cada

conjunto de dados em análise, verificou-se que o serótipo 3 é o que apresenta um maior número de CVAPs, tanto *core* como acessórios, correspondendo também a percentagens mais elevadas (13.1% e 12.6%, respetivamente) de genes específicos daquele serótipo. No entanto, o serótipo 1, apesar de apresentar menos genes *core* específicos, estes correspondem a uma percentagem semelhante de genes *core* no serótipo 3 (13.0%), embora apresente uma menor percentagem de genes acessórios exclusivos (7.8%). É importante ter em conta que os genes *core* de um dos grupos em análise não representam genes *core* da espécie, mas sim genes acessórios. Deste modo, os genes *core* de um dos serótipos, sobretudo os que são específicos desse serótipo, podem representar funções importantes na sua adaptação, sobrevivência e diferenciação relativamente aos outros serótipos, enquanto que os genes acessórios são genes que não são encontrados em todas as estirpes desse serótipo, e como tal não serão essenciais a esse serótipo, mas poderão estar relacionados com fenómenos de adaptação específicos de determinadas estirpes. Assim, serótipo 1 e serótipo 3 apresentam percentagens semelhantes de genes *core* exclusivos dessas estirpes, apesar de o serótipo 1 apresentar muito menos genes *core* no total, ou seja, genes que estão presentes em todas as estirpes desse serótipo, apesar de estarem também presentes em todas as estirpes do serótipo 3 e de "outros serótipos". Já o grupo de "outros serótipos"é o que apresenta menor percentagem tanto de genes *core* como acessórios exclusivos, embora seja o que apresenta mais genes no total, o que seria de esperar, uma vez que é constituído por vários serótipos, logo apresentará maior variabilidade genética, que se traduz num maior conjunto de genes, mas menos genes específicos daquele conjunto de dados.

Quanto aos genes partilhados em específico por dois dos grupos em análise, verificou-se que o serótipo 3 apresenta uma relação mais forte com o grupo de outros serótipos, relativamente aos genes *core* (61.6%), enquanto que o serótipo 1 e o grupo de outros serótipos são os que partilham uma menor percentagem de genes *core* exclusivos (11.1%), e o serótipo 1 e serótipo 3 partilham apenas cerca de 14.2% de genes *core* que não estão presentes em outros serótipos. Já o número de genes acessórios exclusivos situa-se nos 20.9% para o serótipo 1 e serótipo 3, 12.5% para o serótipo 3 e outros serótipos e apenas 6.4% para o serótipo 1 e outros serótipos. Assim, conclui-se que o serótipo 1 e outros serótipos apresentam a relação mais fraca, tanto a nível de genes *core* como de genes acessórios exclusivos. Quanto ao serótipo 3 e outros serótipos, estes apresentam a relação mais forte relativamente aos genes *core* exclusivos, mas para os genes acessórios a relação mais forte verifica-se entre serótipo 1 e serótipo 3. No entanto, a percentagem de genes *core* partilhados entre o serótipo 3 e outros serótipos, não encontrados no serótipo 1, é a mais elevada de todas, correspondendo a mais de metade de todos os genes

core partilhados entre serótipo 3 e "outros serótipos", e também mais de metade dos genes *core* do serótipo 3 e dos genes *core* de "outros serótipos". É assim possível concluir que o serótipo 1 será o grupo mais divergente geneticamente em relação aos outros serótipos, enquanto que o serótipo 3 parece ter bastante em comum com o grupo de outros serótipos. O serótipo 1 e o serótipo 3 parecem partilhar uma maior proporção de genes acessórios do que genes *core*, sendo mesmo os grupos que apresentam a relação mais forte relativamente aos genes acessórios. Assim sendo, estes dois grupos parecem apresentar algumas características comuns, sendo que há uma grande proporção de genes partilhados que não estão no total das 60 estirpes dos dois serótipos.

De uma forma geral, é possível concluir que o serótipo 1 diverge bastante dos restante serótipos, embora possa ter algumas semelhanças com o serótipo 3. Já o serótipo 3 parece ser geneticamente bastante semelhante aos restantes. Assim, o serótipo 1 parece ter sofrido mais modificações, relativamente a um ancestral comum, ao passo que o serótipo 3 se terá mantido pouco alterado.

Capítulo 5

Conclusões e trabalho futuro

5.1 Análise e conclusões

O SCRAG é uma ferramenta de obtenção do genoma *core* e genoma acessório estritos. Desta forma, tem apenas em conta apenas os CVAPs encontrados para o genoma *core* ou para o genoma acessório que cumprem um conjunto de parâmetros, como referido anteriormente, sendo portanto bastante restritivo, e eliminando todos os CVAPs que suscitem algumas dúvidas por não cumprir algum dos parâmetros. No entanto, é ainda possível ao utilizador escolher, além do tipo de análise (genoma *core* ou genoma acessório), as percentagens de identidade e de diferença de tamanho, sendo que percentagens de identidade maiores e percentagens de diferença de tamanho menores se mostram mais restritivas e precisas, obtendo-se um menor número de CVAPs. Desta forma, o utilizador tem a possibilade de determinar o quão rigorosa vai ser a análise, consoante o que seja pretendido.

Nos conjuntos de dados testados, obtiveram-se mais genes acessórios do que genes *core*, sendo que o número de genes *core* será menor e o número de genes acessórios será maior quantos mais são os genomas considerados. No entanto, tendo em conta as restrições consideradas na análise dos resultados do BLAST impostas pelo SCRAG, não se verificou um aumento no total de CVAPs encontrados (*core* + acessórios), tendo o número de genes *core* encontrados diminuído bastante e o número de genes acessórios não tendo aumentado tanto como seria esperado, uma vez que muitos CVAPs terão sido assim excluídos pelo não cumprimento de todos os parâmetros considerados.

Utilizando o SCRAG, foi possível proceder a uma comparação de três grupos distintos de estirpes de *S. pneumoniae*: serótipo 1, serótipo 3 e "outros serótipos". Obteve-se o genoma *core* e o genoma acessório (considerando 80% de identidade e 20% de diferença de tamanho) para cada um destes

grupos e para cada par de grupos, e utilizando também os dados relativos à análise com 76 genomas, obtidos anteriormente, procedeu-se a comparações de forma a determinar os genes únicos ou partilhados entre os diferentes grupos de genomas referidos. Pelos resultados obtidos, foi possível perceber que o serótipo 1 é o grupo que apresenta menos genes *core*, bem como menos genes no total (*core* + acessórios), apresentando pouca diversidade genética, conforme demonstrado previamente [5, 15]. O serótipo 1 é também aquele que se mostrou mais diferente dos restantes serótipos, ao passo que o serótipo 3 parece apresentar bastantes semelhanças com os outros serótipos, existindo uma elevada percentagem de genes partilhados entre estes dois grupos que não são encontrados no serótipo 1 (cerca de 62% dos genes *core* e 12.5% dos genes acessórios). No entanto, e considerando percentagens, o serótipo 1 e o serótipo 3 apresentam ambos cerca de 13% de genes *core* exclusivos de cada um desses grupos, partilhando também entre si em exclusivo (não se encontrado no grupo "outros serótipos") cerca de 14% dos genes *core* comuns, e 21% do total dos genes acessórios partilhados entre os dois grupos – a relação mais forte relativamente aos genes acessórios. O serótipo 3 apresenta também uma elevada percentagem de genes *core* (62%), contrariamente aos restantes grupos, o que parece revelar uma grande similaridade entre estirpes do mesmo serótipo. No entanto, é também o grupo com uma maior percentagem de genes acessórios exclusivos. Já o grupo referente a outros serótipos é aquele em que parece haver maior diversidade genética, como seria de esperar, uma vez que apresenta o maior número de CVAPs encontrados, bem como o maior número de genes acessórios, mas as menores percentagens de genes exclusivos desse grupo.

5.2 Trabalho futuro

Para completar o trabalho aqui apresentado, seria pertinente proceder a uma identificação das funções atribuídas a determinados genes ou grupos de genes, como a distinção em termos funcionais do genoma *core* e do genoma acessório, e ainda dos genes característicos do serótipo 1 e do serótipo 3, bem como os genes que partilham em comum e os que partilham com outros serótipos. Identificar as funcionalidades atribuída aos dados aqui apresentados seria assim importante para compreender melhor uma espécie bacteriana, neste caso, *Streptoccoccus pneumoniae* e respetivos serótipos 1 e 3.

Também seria importante desenvolver mais *scripts* de análise aos resultados obtidos, como um *script* para gerar gráficos ilustrativos dos resultados de forma automática (para além dos gráficos de pontos), ou para facilitar a análise de dados mais complexos, como um *script* que permita obter subcon-

juntos de resultados do BLAST, quando se permite realizar análises múltiplas de subconjuntos de um mesmo conjunto de dados, evitando assim repetir o passo de interrogação da base de dados do BLAST.

Devido a limitações temporais não foram efetuadas determinadas verificações que sustentam e explicitam os resultados obtidos e demonstram a robustez do método, como a repetição das análises utilizando sequências de ácidos nucleicos em vez de aminoácidos, e proceder ao alinhamento múltiplo e geração dos gráficos sem obter as sequências de ácidos nucleicos, para comparar com os resultados obtidos. Também será importante testar os mesmos conjuntos de dados utilizando o SCRAG e outros *software* que procedem a análises similares e realizar a respetiva análise e comparação dos resultados, de forma a testar a fiabilidade dos resultados obtidos com o SCRAG.

Bibliografia

[1] N. Tong, "Priority medicines for europe and the world - a public health approach to innovation - update on 2004 background paper." `http://www.who.int/medicines/areas/priority_medicines/BP6_22Pneumo.pdf`, Maio 2013.

[2] H. Tettelin et al, "Complete genome sequence of a virulent isolate of Streptococcus pneumoniae," *Science*, Julho 2001.

[3] "Immunization, vaccines and biologicals - pneumococcal disease." `http://www.who.int/immunization/topics/pneumococcal_disease/en/`, Setembro 2014.

[4] M. J. Struelens et al, "Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems," *Clinical Microbiology and Infection*, Agosto 1996.

[5] C. Donati et al, "Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species," *Genome Biology*, Outubro 2010.

[6] H. Tettelin, D. Riley, C. Cattuto, and D. Medini, "Comparative genomics: the bacterial pan-genome," *Current Opinion in Microbiology*, Outubro 2008.

[7] D. Field, G. Wilson, and C. van der Gast, "How do we compare hundreds of bacterial genomes?," *Current Opinion in Microbiology*, Agosto 2006.

[8] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, "The microbial pan-genome," *Current Opinion in Genetics and Development*, Setembro 2005.

[9] T. Lefébure and M. J. Stanhope, "Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition," *Genome Biology*, Maio 2007.

[10] A. Pertsemlidis and J. W. F. III, "Having a blast with bioinformatics (and avoiding blastphemy)," *Genome Biology*, Setembro 2001.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, Maio 1990.

[12] R. She, J. S.-C. Chu, K. Wang, J. Pei, and N. Chen, "genBlastA: Enabling BLAST to identify homologous gene sequences," *Genome Research*, Setembro 2008.

[13] P. Martens, S. W. Worm, B. Lundgren, H. B. Konradsen, and T. Benfield, "Serotype-specific mortality from invasive Streptococcus pneumoniae disease revisited," *BMC Infectious Diseases*, Junho 2004.

[14] N. D. Ritchie, T. J. Mitchell, and T. J. Evans, "What is different about serotype 1 pneumococci?," *Future Microbiology*, Janeiro 2012.

[15] T. M. Williams, N. J. Loman, C. Ebruke, D. M. Musher, R. A. Adegbola, M. J. Pallen, G. M. Weinstock, and M. Antonio, "Genome analysis of a highly virulent serotype 1 strain of streptococcus pneumoniae from west africa," *PLOS One*, Outubro 2012.

[16] J. Ahl, N. Littorin, A. Forsgren, I. Odenholt, F. Resman, and K. Riesbeck, "High incidence of septic shock caused by Streptococcus pneumoniae serotype 3 - a retrospective epidemiological study," *BMC Infectious Diseases*, Outubro 2013.

[17] D. A. Rasko, G. S. Myers, and J. Ravel, "Visualization of comparative genomic analyses by BLAST score ratio," *BMC Bioinformatics*, Janeiro 2005.

[18] J. W. Sahl, J. G. Caporaso, D. A. Rasko, and P. Keim, "The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes," *PeerJ*, Abril 2014.

[19] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, Fevereiro 2004.

[20] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, Março 2010.

[21] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzne, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *Journal of Computational Biology*, Maio 2012.

[22] T. M. Williams, N. J. Loman, C. Ebruke, D. M. Musher, R. A. Adegbola, M. J. Pallen, G. M. Weinstock, and M. Antonio, "Genome analysis of a highly virulent serotype 1 strain of Streptococcus pneumoniae from west africa," *PLoS One*, Outubro 2012.

[23] Z. Sui, W. Zhou, K. Yao, L. Liu, G. Zhang, Y. Yang, and J. Feng, "Complete genome sequence of Streptococcus pneumoniae strain A026, a clinical multidrug-resistant isolate carrying tn2010," *Genome Announcements*, Dezembro 2013.

[24] G. Lia, F. Z. Hub, X. Yangc, Y. Cuic, J. Yangd, F. Que, G. F. Gaof, and J.-R. Zhang, "Complete genome sequence of Streptococcus pneumoniae strain ST556, a multidrug-resistant isolate from an otitis media patient," *Journal of Bacteriology*, Março 2012.

[25] The Human Microbiome Jumpstart Reference Strains Consortium, "A catalog of reference genomes from the human microbiome," *Science*, Maio 2010.