Universidade de Lisboa
Faculdade de Ciências
Departamento de Química e Bioquímica

Pesquisa de módulos de genes associados à invasividade de *Streptococcus pneumoniae* usando semelhanças semânticas

João José dos Reis Malaquias

Dissertação
Mestrado em Bioquímica
Especialização em Bioquímica Médica

2014

Universidade de Lisboa

Faculdade de Ciências
Departamento de Química e Bioquímica

Pesquisa de módulos de genes associados à invasividade de *Streptococcus pneumoniae* usando semelhanças semânticas

João José dos Reis Malaquias

Dissertação
Mestrado em Bioquímica
Especialização em Bioquímica Médica
Orientador:
Prof. Dr. Francisco Rodrigues Pinto

2014

Agradecimentos:

Ao Francisco Pinto, meu orientador de tese, sem dúvida alguma. Pela paciência e disponibilidade que teve comigo durante todo o tempo de execução do trabalho da tese, ao iniciar-me numa área sobre a qual não tinha experiência.

Às pessoas do Centro de Química e Bioquímica, que sempre me proporcionaram um ambiente intelectualmente rico e estimulante, e sempre me ajudaram quando precisei.

À minha família, especialmente à minha mãe e meus irmãos.

A todos os que contribuíram de algum modo para a boa execução deste trabalho.

A todos bem-haja.

Resumo:

O *Streptococcus pneumoniae*, ou pneumococcus, é uma bactéria causadora de doenças em humanos,tais como pneumonia, meningite bacteriana e otite, entre outras. Apresenta uma elevada taxa de mortalidade em certas faixas etárias e regiões do mundo. Desde a década de 70 do séc. XX que têm sido desenvolvidas vacinas que conseguiram diminuir a taxa de infeção pelo pneumococcus. No entanto, além da progressiva resistência adquirida aos antibióticos e à vacinação, observou-se que a divisão dos serotipos em classes de estirpes causadoras de doença ou simplesmente colonizadoras não é estática. Algumas estirpes que causam doença num hospedeiro não causam noutro e vice-versa, além de clones com o mesmo serotipo apresentarem diferentes padrões de invasividade. Uma compreensão aprofundada dos mecanismos de invasividade do pneumococcus poderá fornecer novas vias para o combate às doenças provocadas por este microorganismo.

A capacidade do pneumococcus de causar doença é heterogénea entre as varias estirpes que fazem parte da espécie. Especificamente, o conteúdo génico inter-estirpes varia, sendo o reservatório do pan-genoma do pneumococcus não só as várias estirpes, como até outras espécies de *Streptococcus*, estando bem documentada a transferência horizontal de genes (HGT). Devido a esta heterogenidade, é possível observar co-presenças de genes específicos e a invasividade e deduzir potenciais associações "genes → doença".

Este trabalho está inserido num projeto que usa uma abordagem da biologia de sistemas, ao recorrer à construção de relações semânticas entre processos biológicos de genes de *S. pneumoniae*, com o objectivo de identificar módulos de genes funcionalmente coerentes associados à virulência. Se a função de um gene tem impacto na virulência do microrganismo, então os genes envolvidos em processos biológicos vizinhos em termos semânticos poderão ter uma influência similar, mesmo que indireta, ao afetar a atividade desse gene. A busca de módulos de genes é mais poderosa do que a identificação de associações de genes individuais com a virulência. A associação cumulativa do módulo pode ser significativa mesmo quando nenhum dos genes constituintes apresenta uma associação suficientemente forte *per se*.

A partir da lista de genes dos genomas de três estirpes de referência (G54, R6 e Tigr4), foi medida a semelhança semântica entre cada par possível de genes, usando para isso as anotações de processo biológico de acordo com a Gene Ontology (GO), um vocabulário controlado de termos usados para anotação de genes. Estas semelhanças foram então usadas para, computacionalmente, construir módulos de genes, que não são mais que conjuntos de genes com uma semelhança média elevada entre si. Usando dados de hibridação genómica comparativa por microarray (aCGH) referentes a 72 estirpes de pneumococcus, a contagem de presenças de genes de cada módulo em

estirpes consideradas invasivas foi comparada com a mesma contagem em estirpes consideradas colonizadoras. Aqueles módulos em que as estirpes consideradas invasivas tinham maior presença em relação às colonizadoras foram considerados associados à invasividade, sendo depois analisados os resultados, quanto à sua função biológica. Estes revelaram funções relacionadas com a síntese e a degradação proteicas, metabolismo e transporte de carbohidratos, e em menor magnitude módulos associados ao processamento da cápsula e a mecanismos relacionados com a expressão génica. Acreditamos que este método permite tirar partido das ontologias e das propriedades destas que permitem medir semelhanças semânticas, aplicando-as neste caso na descoberta de potenciais alvos terapêuticos.

Abstract:

Streptococcus pneumoniae, or pneumococcus, is a bacteria causing human disease, such as pneumonia, bacterial meningitis and otitis media, among others. It has a high rate of mortality in certain age groups and regions. Since the late 70th of XX century vaccines able to decrease the rate of pneumococcal infection have been developed. However, besides the progressive acquired resistance to antibiotics and vaccination, it was observed that the division of serotypes in classes causing disease or colonizing is not static. Some strains that cause disease in a host don't cause in another and vice versa, and clones with the same serotype have different patterns of gene expression. A thorough understanding of the mechanisms of invasiveness of the pneumococcus may provide new avenues for combating diseases caused by this organism.

The ability of the pneumococcus to cause disease is heterogeneous among the different strains belonging to the species. Specifically, the inter-strain gene content varies, being the reservoir of the pan-genome of pneumococcus not only the various strains, but also other Streptococcus species, being well documented horizontal gene transfer (HGT). Because of this heterogeneity, it is possible to observe the co-presence of the expression of specific genes and deduced invasiveness and potential "disease gene expression →" associations.

This work is inserted in a project using a systems biology approach, by using the construction of semantic relationships between biological processes of genes from S. pneumoniae in order to identify modules of functional coherent genes associated with virulence. If the function of a gene has an impact on virulence of the microorganism, so the semantically neighboring genes in terms of biological process may have a similar effect, even if indirectly, by affecting the activity of that gene. The search for gene modules is more powerful than the identification of associations of individual genes with virulence. The cumulative association of the module can be significant even when none of the constituent genes provides a strong enough association *per se*.

From the list of genes present in the genomes of three reference strains (G54, R6 and Tigr4), semantic similarity between each possible pair of genes was measured. These similarities were then used to computationally build gene modules which are no more than clusters of genes with a high average similarity. The module gene counts in strains considered invasive versus the same counts in strains considered colonizers were compared and those modules that were most present in invasive strains were considered to be associated with virulence, the results are then analyzed in terms of their biological processes. We believe that this method takes advantage of the properties of these ontologies and the capacity of measuring semantic similarity, applying them in the discovery of

potential therapeutic targets.

Índice:

Índice de figuras

Introdução:

O *Streptococcus pneumoniae* é uma bactéria gram-positiva que tem o seu nicho ecológico em humanos no trato respiratório superior. A relação ecológica da bactéria com a nossa espécie é, num quadro não patológico, de colonização assintomática [1]. No entanto, em indivíduos suscetíveis, tais como crianças, idosos e pessoas imunodeprimidas, essa relação pode-se modificar e a bactéria nesse caso provoca doenças, tais como pneumonia, meningite, otite, sinusite e sepsis [1]. O quadro clínico é variável, dependendo este de fatores como o serotipo da bactéria e a suscetibilidade do indivíduo doente, podendo ser fatal. Em 2010, cerca de 18% das mortes a nível mundial, em crianças com menos de 5 anos, foram devidas à pneumonia, sendo o pneumococcus o agente mais comum, responsável por esta doença.

Até antes da revolução causada pelos antibióticos, as infeções causadas por este agente, assim como por outros microorganismos, tinham alta taxa de mortalidade. Após esse período estas passaram a ser encaradas como triviais, dada a elevada taxa de sucesso dos tratamentos com recurso a antibióticos . Simultaneamente, avanços noutras áreas, tais como na vacinação, permitiram atuar na prevenção, aumentando assim o leque de abordagens possíveis no combate às doenças do foro bacteriano. No caso particular do pneumococcus, desde a década de 70 do séc. XX, foram aprovadas para uso vacinas tendo como agente imunogénico a cápsula polissacárida do pneumococcus (CPS) [1]. Estas vacinas conferem imunidade para aprox. 90% das estirpes causadoras de doenças. O principal problema desta classe de vacinas é o facto dos antigénios baseados na cápsula polissacárida serem fracos agentes imunogénicos em crianças com menos de 2 anos– o maior reservatório do pneumococcus em humanos – devido às reacções imunogénicas nesta faixa etária serem independentes das células T e não criarem memória imunológica a longo prazo. Novas classes de vacinas têm sido desenvolvidas, nomeadamente as vacinas conjugadas CPS/proteínas. Estas têm maior capacidade de estimular as células T, sendo aplicadas, por exemplo no Reino Unido, a crianças com menos de 5 anos, a indivíduos com mais de 65 aos e a outros grupos de risco, tais como indivíduos imunocomprometidos, diabéticos, doentes cardíacos crónicos. No entanto, esta classe de vacinas é dispendiosa, o que poderá ser um problema quanto à sua aplicação generalizada em países pobres.

A emergência de estirpes resistentes e multi-resistentes, a par da diminuição da taxa de descoberta de novos antibióticos pôs em evidência um problema que é a diminuição do arsenal de combate às infeções por microorganismos, incluindo o pneumococcus. Algumas das consequências são o aumento de estirpes resistentes a antibióticos de primeira linha e a emergência de estirpes

multi-resistentes.

Outro problema associado aos organismos patogénicos é o facto do nicho ecológico deixado livre pelas estirpes eventualmente eliminadas pela vacinação ser ocupado por estirpes serotipicamente diferentes. Este fenómeno é designado por *treatment strain replacement* ou *replacement effect* [2], [3]. Isto acontece devido ao facto dos microorganismos em geral terem reduzidos tempos inter-geracionais, um elevado número de descendência e mutabilidade elevada. Este fenómeno aumenta o risco potencial da emergência de patologias causadas por serotipos não incluídos na vacinação, e o pneumococcus não é exceção a este fenómeno. No estudo epidemiológico usado como fonte de informação para este trabalho, observou-se que o potencial de causar doença para os clones de alguns serotipos é variável, ou seja, alguns serotipos recolhidos em indivíduos doentes estão igualmente presentes em indivíduos portadores saudáveis, e vice-versa. [4]. Este facto reforça a hipótese da importância do papel da variabilidade genética no processo de invasividade. Existem já referências a casos de bactericémia causados por serotipos não incluídos nas classes de vacinas usadas [5]. Observou-se também que a pressão seletiva sobre as estirpes resistentes, em que o factor de pressão é o antibiótico, favorece a seleção de estirpes específicas, conforme o alvo do antibiótico[6]. Várias abordagens estão actualmente a ser estudadas, e a descoberta de novos potenciais alvos terapêuticos e até de alternativas aos antibióticos (*silencing* do RNA bacteriano [7], por exemplo) são potenciais ferramentas para que se mantenha a vantagem dos humanos sobre as doenças causadas por organismos patogénicos, vantagem essa suportada até agora pelos antibióticos e pela vacinação. Neste trabalho usaram-se semelhanças semânticas, que reflectem em princípio uma semelhança quanto à função biológica, como um agregador de entidades que, caso tenham uma associação positiva com a presença das estirpes consideradas invasivas, poderão ser importantes para o sucesso do processo patológico desencadeado pelo pneumococcus.

São vários os passos envolvidos no processo que culmina na patogénese do pneumococcus, sendo os fatores de virulência responsáveis pela evasão ao sistema imune do hospedeiro e também pelo processo inflamatório. A cápsula bacteriana desempenha um papel importante no processo de invasividade. Esta possui propriedades que permitem à bactéria resistir à fagocitose e contém os determinantes moleculares que definem o serotipo da bactéria em questão. Outros fatores de virulência incluem moléculas de ligação a componentes celulares e extracelulares do hospedeiro, componentes que dotam o pneumococcus de resistência ao stress oxidativo, à evasão do sistema imune de complemento do hospedeiro e péptidos antimicrobianos envolvidos na competição intra-específica pelo nicho ecológico (nasofaringe) [1].

O nicho ecológico do pneumococcus em humanos é a nasofaringe, onde este pode habitar

numa relação comensal com o hospedeiro. Os episódios patológicos acontecem quando o microorganismo coloniza outros locais do organismo do hospedeiro [8] ou quando o hospedeiro é susceptível.

Quanto à genética do pneumococcus, esta apresenta uma grande variabilidade, possuindo também capacidade de competência (incorporação de DNA exógeno). As estirpes comensais ou virulentas apresentam um genoma reduzido, na ordem dos 1,5 – 3 Mb, enquanto as estirpes de vida livre apresentam um genoma maior. Este facto provavelmente é devido à necessidade de economizar recursos não essenciais quando num organismo hospedeiro, mantendo no entanto a capacidade de adquirir DNA exógeno, facto verificado pela presença de operões que possibilitam a esses pneumococci a aquisição de DNA por *horizontal gene transfer* (HGT). Uma característica de algumas infecções crónicas provocadas por bactérias é a formação de um biofilme que protege as bactérias do sistema imunitário do hospedeiro. Adicionalmente, este biofilme, quando devido a infeções policlonais, cria um ambiente favorável às bactérias para que haja HGT, tendo tal facto sido verificado experimentalmente para *S. pneumoniae* [9].

As fontes da diversidade genética dos pneumococci são geralmente outros Streptococcus, nomeadamente S. infantis, S. oralis e S. mitis, partilhando estes entre si genes responsáveis pelos factores de virulência presentes em patologias humanas. Há estudos que apontam para a hipótese de que certos organismos possuam um pan-genoma distribuído num 'superset' genómico distribuído por várias estirpes e que aquele 'flui' por recombinação entre as várias estirpes com capacidade de recombinação (competência), designando-se este fenómeno por 'distributed-genome hypothesis (DGH). Observou-se que a dinâmica genética dos pneumoccoci se enquadra neste fenómeno [10]. Este facto foi verificado por análises moleculares filogenéticas, e é uma das hipóteses para explicar o facto do pan-genoma do pneumococcus ser tão extenso. Especificamente, analisando a origem dos genes dispensáveis, observou-se que *S. mitis* é o principal reservatório da variabilidade genética dos pneumoccoci [11]. Todos estes factos dotam os pneumococci de uma flexibilidade acrescida que aumentam as suas probabilidades de sobrevivência e evasão ao sistema imunitário do hospedeiro. Uma consequência resultante dos fenómenos relacionados com esta dinâmica de partilha génica inter-específica poderá ser o aumento da taxa de substituição génica de alelos e/ou genes. Alelos de genes não essenciais para a invasividade do pneumococcus que sejam usados em vacinas poderão ser mais rapidamente substituídos por outros que eventualmente poderão aumentar a taxa de sucesso da invasividade do pneumococcus. Num estudo epidemiológico [4], observou-se que, numa coleção de isolados recolhidos em indivíduos saudáveis portadores do pneumococcus e em indivíduos doentes, alguns dos serotipos recolhidos tanto em portadores como em doentes são os mesmos, o que indica que um mesmo serotipo em hospedeiros diferentes pode exibir características

de invasividade ou existir numa relação não patológica. Mais ainda, em alguns serotipos, observou-se que vários dos seus clones apresentavam capacidade variável para causar doença, o que indica que a variabilidade genética poderá ter um papel importante no surgimento de doença dentro do mesmo serotipo.

A problemática da crescente resistência dos microorganismos em geral aos antibióticos disponíveis é uma área da ciência que, tal como provavelmente todas as outras, tem o potencial de tirar partido do conhecimento acrescido obtido a partir de informação pré-existente. No entanto, a premência da descoberta de novos potenciais alvos terapêuticos e a cada vez maior dificuldade na descoberta destes, no âmbito da resistência aos antibióticos, é um problema para o qual as ontologias se podem revelar uma ferramenta de valor acrescentado, nomeadamente na interpretação de resultados, devido às suas características. Outras áreas, como a farmacogenómica, em que a necessidade de obter resultados personalizados num intervalo de tempo que permita a viabilidade do tratamento, podem beneficiar também de ferramentas que façam uso das ontologias [12].

De modo a manipular com eficiência e tirar partido da cada vez maior quantidade de dados produzida pelas 'ómicas', por parte da comunidade científica, este trabalho visa tirar proveito da quantidade massiva de dados disponível e a sua crescente anotação. Para tal, implementou-se um método que tira partido da anotação dos produtos génicos, usando aquela para compará-los e medir a semelhança entre eles. A anotação usada para os produtos génicos considerados neste trabalho pertence a um conjunto que constitui um **vocabulário controlado** denominado **Gene Ontology** (GO), que é um conjunto de termos, tais como ”enzyme”, “transcription factor” ou “glicolysys pathway” que representam os mesmos termos usados na literatura das ciências biológicas e da saúde. O uso de termos numa linguagem natural visa criar um vocabulário que seja entendível por pessoas e passível de ser tratado computacionalmente, devido justamente ao vocabulário ser controlado, ou seja, este obedece a regras, sendo estas comuns a todos os sistemas digitais onde o vocabulário é usado e à aplicação por humanos. Este vocabulário, quando aplicado a uma base de dados representativa do genoma de um ou vários organismos, tal como a UniProt KnowledgeBase, dota cada entidade presente na base de dados (neste caso sequências representando produtos da expressão génica) de uma anotação adicional – uma ontologia – sendo a Gene Ontology também uma ontologia.

O conceito de ontologia é definido como uma caracterização objectiva e formal de determinado conceito ou entidade, usando como meios para tal um vocabulário comum, de modo que quando determinada entidade é caracterizada com recurso a uma ontologia, isso significa que a caracterização será feita com recurso a termos, tais como palavras, que pertencem ao vocabulário que constitui a ontologia. Diz-se que o vocabulário é comum porque, como a ontologia é aplicada a

uma determinada área do conhecimento, o contexto em que se inserem os termos do vocabulário dessa ontologia (criado especificamente para ela) está definido à partida. Este facto permite a reutilização da mesma ontologia para anotar várias coleções dentro da área de conhecimento para a qual foi criada. Uma ontologia proporciona um vocabulário comum que pode ser usado para caracterizar um domínio do conhecimento, anotando os as entidades pertencentes ao respetivo domínio do conhecimento com termos que transmitem informação sobre a natureza das respetivas entidades e também as relações entre elas. Esta anotação, se rigorosa, acrescenta utilidade às anotações que não apenas uma informação de teor enciclopédico ou uma maior facilidade de pesquisa (importantes por si só). No caso deste trabalho, implementou-se o conceito de **semelhança semântica**, que significa medir o quão semelhante são dois produtos génicos entre si, sendo tão mais semelhantes quanto mais termos GO (Gene Ontology) possuírem em comum.

Quando se comparam dois produtos génicos, ou simplesmente é necessário consultar a anotação de apenas um produto génico, os termos GO (Gene Ontology) permitem saber qual a sua função biológica, molecular, e a sua localização celular. Com base nesta informação, usou-se o conceito de **conteúdo informativo** (IC ou information content), que reflete a frequência com que determinado termo é usado na anotação da totalidade das entidades, num determinado ***corpus*** (o *corpus* neste caso é toda a base de dados Uniprot KnowledgeBase). Quanto mais entidades forem anotadas por um determinado termo dentro de um *corpus*, menor será o poder discriminatório desse termo específico dentro do *corpus* em análise e, como tal, menos informativo será. O inverso também é verdadeiro. Como o conteúdo informativo (IC) é um valor numérico (a frequência), pode-se usar esse valor para comparar dois termos entre si. Dois produtos génicos serão tão mais semelhantes semanticamente entre si. Se se compararem todos os termos de um organismo entre si, por exemplo numa matriz representativa do mapa combinatório de todos os produtos génicos, poderemos saber, para cada par de produtos génicos, o valor de semelhança entre ambos os elementos.

Este trabalho consistiu na construção de módulos coerentes de genes presentes em *S. pneumoniae* para posterior análise quanto à sua associação com a invasividade. Os módulos em si são conjuntos de 20 ou 50 sequências génicas, correspondentes a genes do pneumococcus, com semelhança funcional entre si. O cálculo da semelhança entre cada par de genes foi um trabalho prévio efectuado pelo BDXL, do departamento de informática da FCUL. Neste trabalho, o conjunto das semelhanças semânticas entre pares de genes foi usada por nós para construir módulos de genes funcionalmente coerentes.

Para testar a associação dos módulos à invasividade, estes foram comparados com os conteúdos génicos de várias estirpes de *S. pneumoniae* obtidas a partir de um estudo prévio, em que

foram estudadas 72 estirpes de S. pneumoniae isoladas em Portugal. 49 foram considerados invasivos e 23 simplesmente colonizadores, de acordo com dados epidemiológicos da população pneumocóccica portuguesa [4]. O genoma destes serotipos foi analisado por hibridação em microarrays. Estes representavam os genomas de três estirpes de referência: G54, R6 e Tigr4. Os resultados consistiram no mapa de presenças e ausências de cada ORF para cada uma das 72 estirpes testadas.

A associação dos módulos de genes à virulência consistiu em comparar o mapa de presenças e ausências de cada uma das 72 estirpes com cada um dos módulos. Para cada módulo, se o número de presenças de genes do módulo nas estirpes consideradas invasivas for superior ao número de presenças nas estirpes consideradas colonizadoras, o módulo é considerado como potencialmente associado à invasividade. A análise estatística usada seguiu o modelo de distribuição hipergeométrico, em que foi medido o nível de enriquecimento de presenças do módulo em estirpes invasivas versus colonizadoras. Nesse caso, cada um dos 20 ou 50 genes do módulo será posteriormente analisado quanto à sua função biológica.

Após a seleção dos módulos candidatos, foi necessário desenvolver e aplicar testes estatísticos aos mesmos, para calcular a probabilidade da presença diferencial em estirpes invasivas versus colonizadoras no módulo ser devida ao acaso ou efetivamente existir uma associação à invasividade.

Abordagem Computacional e ontologias:

O objetivo deste projeto consistiu na construção de módulos de genes com uma associação à invasividade, usando informação da presença diferencial de genes obtida através de hibridação genómica comparativa no âmbito de um estudo epidemiológico [4], anotando esses genes com recurso à Gene Ontology.

O projeto enquadra-se num esforço maior que consiste na criação de métodos que permitam obter uma associação entre motivos de rede e fenótipos virulentos. Este projeto tira partido da fusão entre as Tecnologias de Informação e as tecnologias de deteção molecular, que permitem a obtenção de informação em larga escala. Tal apenas é possível, num espaço de tempo razoável, se se possuir meios de automatizar tarefas e processar elevadas quantidades de informação. O exponencial crescimento tanto da informação obtida pelas "ómicas" (genómica, transcriptómica, proteómica, metabolómica, entre outros) como da obtenção de conhecimento acrescido usando como base essa informação deve-se, sem dúvida, à evolução das tecnologias de recolha de informação em larga escala, mas principalmente à capacidade crescente de poder computacional disponível e à eficácia dos algoritmos preditivos de sequência e função. Caso contrário não se conseguiria ter a quantidade, tanto de informação como principalmente da anotação desta, que se tem atualmente, graças, por exemplo, à crescente sofisticação dos algoritmos preditivos das funções génica, proteica e enzimática, entre outras.

Um dos fatores que pode ser limitante ao aproveitamento da informação disponível é a multiplicidade de regras de anotação díspares entre si, sem obedecer a uma norma. Este facto pode obrigar a um re-tratamento da informação, caso seja necessário o uso de informações anotadas de modos dispares, obrigando a trabalho adicional e gastos de tempo que podem ser evitados a montante, caso uma norma de classificação da informação seja usada por toda uma comunidade. Justamente com o objetivo de unificar o modo como a informação disponibilizada é anotada, surgiram várias iniciativas, sendo as ontologias (onde se inclui a Gene Ontology) uma delas.

O uso de uma linguagem comum para a anotação de funções bioquímicas ou biológicas proporciona à comunidade científica uma maior agilidade do tratamento da cada vez maior quantidade de informação que as "ómicas" produziram e continuam a produzir. Adiciona também confiança no uso de termos pertencentes ao universo da ontologia usada, já que o investigador, ao usar um termo em questão (numa pesquisa, por exemplo), saberá *a priori* que todas as entidades anotadas com esse termo serão consideradas. Simultaneamente, devido ao vocabulário ser normalizado, terá a segurança de nenhum resultado correspondente aos termos usados na pesquisa

ter sido excluído na pesquisa, caso se encontre anotado com termos pertencentes ao universo da ontologia usada. Uma característica importante das ontologias é o seu vocabulário (corpus) ser entendido tanto por humanos como por computadores, eliminando assim a necessidade de interfaces e agilizando assim o fluxo da informação entre os Sistemas Digitais dedicados às 'ómicas' e os utilizadores.

Gene Ontology:

A Gene Ontology (GO) é uma iniciativa que visa a implementação de um vocabulário comum que descreva genes e produtos da expressão génica para qualquer organismo de qualquer domínio da vida. Esta iniciativa, gerida pela Gene Ontology Consortium, é um trabalho colectivo de vários grupos, que contribuem com novas anotações, revêm e atualizam o vocabulário existente e propõem correções ao mesmo. A Gene Ontology faz parte de um esforço maior que tem em vista a criação de um vocabulário controlado que possa ser partilhado entre as diversas áreas da biologia e da medicina – a Open Biomedical Ontologies. O vocabulário GO visa classificar qualquer produto da expressão génica quanto a 3 características, ou classes: *função molecular*, *processo biológico* e *componente celular*, podendo a escolha das anotações de determinada entidade para as suas 3 classes ser resultado de trabalho experimental, deduzida através de algoritmos preditivos ou ambos. A anotação é feita com recurso ao vocabulário GO, que é o conjunto de todos os termos (palavras) disponíveis para uso na classificação e anotação, escolhendo-se, para cada uma das três classes, o ou os termos que, o mais especificamente possível, se adequam à entidade em questão. Os termos pertencentes ao vocabulário GO são em língua inglesa e são semelhantes aos usados na literatura científica.

Formalmente, o vocabulário GO é organizado na forma de um **grafo acíclico dirigido.** Este tipo de grafo toma a forma aproximada de uma árvore, em que junto à raíz estão as 3 classes **independentes** pelas quais qualquer entidade pode ser classificada (*função molecular*, *processo biológico* e *componente celular*). Estas 3 classes incluem todos os termos que constituem o vocabulário GO, estruturados numa hierarquia em que os termos mais gerais, e por consequência menos específicos, estão perto da raíz, e subordinado a cada termo existe um ou mais termos "filhos", mais específicos que o termo parental. Quanto mais informação se possuir sobre determinada entidade candidata a ser classificada com recurso a termos GO, maior é a hipótese de esta ser classificada com termos mais específicos e, consequentemente mais informativos. Como existem três grafos independentes, expressando cada um deles as relações hierárquicas entre os

termos de cada uma das classes de anotação disponíveis (função molecular, processo biológico e componente celular) e como cada entidade pode ser anotada pelos termos das 3 classes, cada entidade pode estar representada nos três grafos GO.

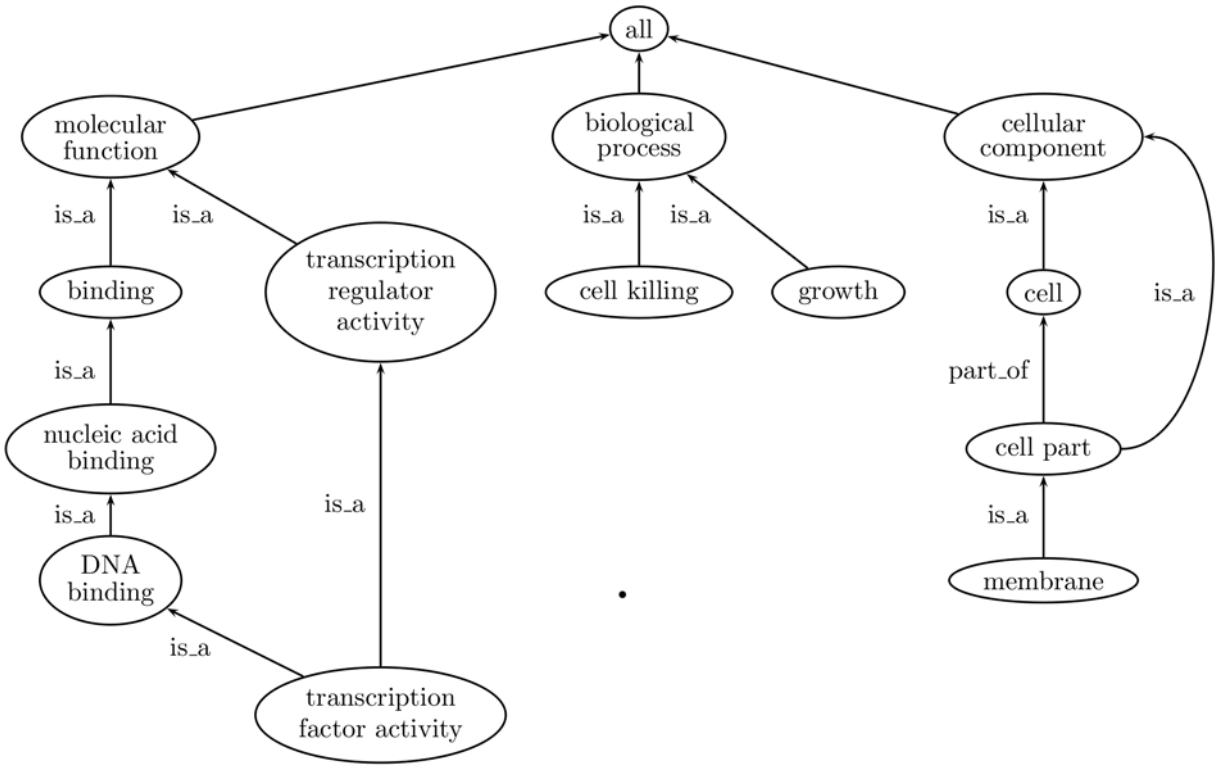

Figura 1: Secção do grafo GO representando os sub-grafos correspondentes às três classes de anotações GO existentes (molecular function, biological process e cellular component), assim como as relações hierárquicas entre os termos. Estas são sempre designadas com termos tais como 'is a' ou 'part of'', de modo a proporcionar, de um modo o mais natural possível de entender, as relações entre os termos. A relação entre as entidades 'transcription factor activity', 'cell part' e os seus respetivos termos parentais reflete a forma das relações GO como um grafo acíclico dirigido e não como uma árvore, em que nesta todos os termos-filho apenas têm um termo parental.

Classificação de produtos génicos:

A Gene Ontology, embora possa ser consultada, não proporciona informação biológica *per se*. É a aplicação dos seus termos na anotação das entidades biológicas que a torna uma ferramenta. Quando determinada entidade é anotada (neste trabalho as entidades são genes), o que acontece é a associação dos termos que melhor a classificam. A classificação é feita, como dito anteriormente, em três classes (*função molecular*, *processo biológico* e *componente celular)*, sendo esta feita automática ou manualmente. Cada entidade poderá ser anotada com um ou mais termos de cada

uma das três classes que melhor descrevam a entidade. Cada termo terá uma posição no grafo da respetiva classe, conforme o seu nível de especificidade. No entanto, o que se descreveu até agora não foi a implementação da Gene Ontology a algo concreto. Apenas se explicou a sua arquitetura. A Gene Ontology é aplicada geralmente a um conjunto de produtos génicos. Embora nada impeça a anotação de um qualquer produto génico isolado com termos GO, o valor acrescentado do uso das ontologias observa-se quando este é aplicado a bases de dados com alguma dimensão. Uma delas, que foi a usada neste trabalho, é a UniProt KnowledgeBase (UniProtKB). Esta base de dados contém sequências de produtos génicos, nomeadamente sequências de aminoácidos resultantes da tradução de um transcrito. No caso específico da UniProtKB, esta contém sequências de diversos organismos, incluíndo do pneumococcus. O facto de se usar como fonte de dados uma base de dados de grande dimensão permite obter certas informações, tais como o valor do conteúdo informativo, uma informação essencial para a execução deste trabalho. O **conteúdo informativo, ou Information content (IC)**, é essencial para o cálculo das semelhanças semânticas e será explicado em detalhe em seguida. Cada entidade disponibilizada pela UniProtKB, caso tenha anotação GO, possibilita a visualização de cada termo na árvore GO, onde se pode consultar a posição de cada um dos termos que anotam a entidade em questão no grafo GO, para cada classe de anotação.

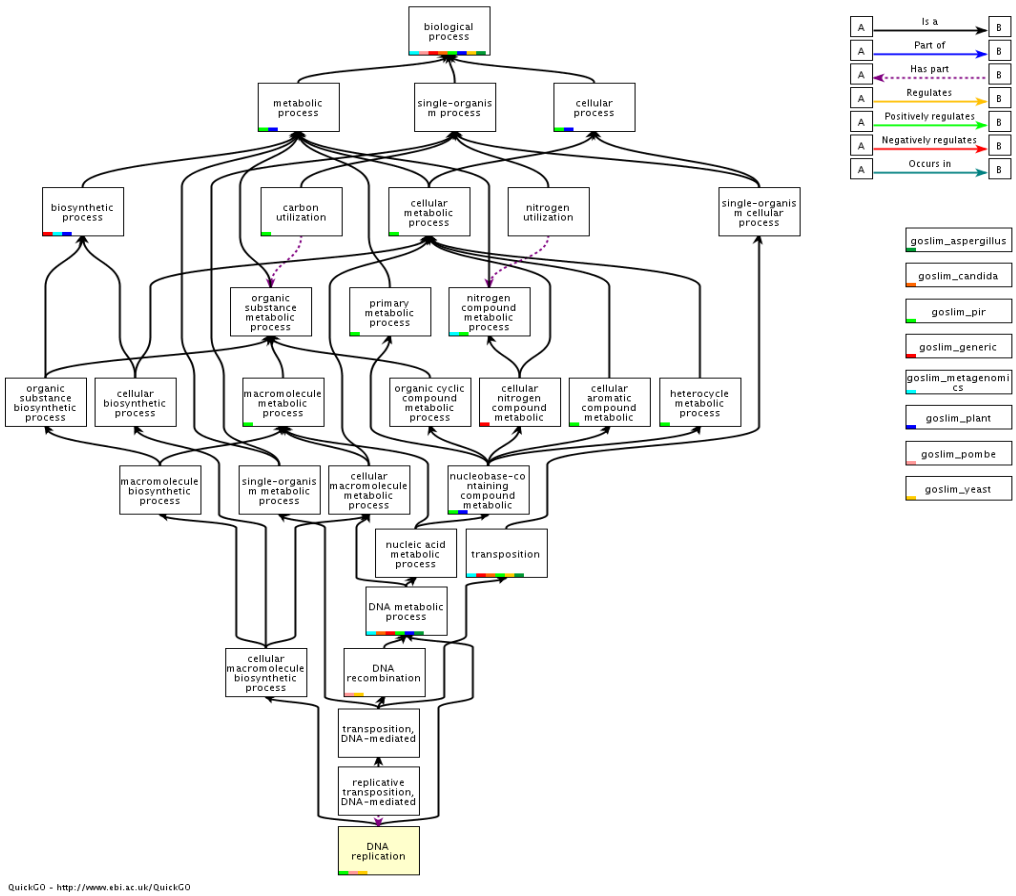

Figura 2: Diagrama onde se pode observar a posição na árvore GO do termo GO:0006260, que corresponde ao processo biológico "DNA replication". Note-se que a árvore mostra a posição de um termo da classe "processo biológico". caso a sequência que este termo GO anota tenha também anotações pertencentes às classes "função molecular" e "componente celular", esses termos estarão representados nas árvores correspondentes às classes respectivas e não nesta.

Se determinada entidade a ser classificada for, por exemplo, uma proteína de função desconhecida, com localização celular precisamente conhecida, a localização relativa dos termos de cada uma das três classes de anotações nos respectivos grafos GO, em relação à raiz da árvore, será topologicamente bastante díspar. A sua classificação no grafo *componente celular* poderá fazer uso de termos bastante particulares, tais como 'alpha-1,6-mannosyltransferase complex'. Por outro lado, a sua classificação no grafo *processo biológico* pode não existir, devido à falta de informação, e a função molecular poderá ser apenas 'proteína', caso não exista informação sobre a sua função molecular. Visualmente, na árvore das relações hierárquicas entre os termos GO, o termo alpha-1,6-mannosyltransferase complex, que é parte de 'Golgi cisterna', que 'é' um 'Golgi cisterna',

O modo como esta relação hierárquica é estruturada é, topologicamente, um **grafo acíclico dirigido** (directed acyclic graph), já que a organização das entidades GO anotadas é direcional, ou seja: os termos mais gerais (função molecular, processo biológico e componente celular) são os nós (ou vértices) aos quais todas as anotações GO existentes estão ligadas, embora algumas entidades possam não possuir anotação no total das 3 classes, pelos motivos anteriormente expostos**.** Assim, uma biomolécula que esteja envolvida na apoptose, por exemplo, terá como termo parental "morte celular", e este terá como termo parental "processos biológicos". A posição que a entidade classificada irá ocupar no grafo dependerá do detalhe com que é possível classificá-la. Assim, se para determinada proteína se souber apenas que é uma proteína membranar, não se poderá classificá-la como "transmembranar", "integral" ou "de superfície", pois não se dispõe de informação suficiente para tal. Nesse caso, o termo "membranar" ocupará uma determinada posição na hierarquia, herdando também todos os termos mais gerais que também a qualificam, tais como "proteína de membrana" e "proteína". Caso posteriormente se obtenha informação adicional que permita classificar com maior detalhe a entidade em questão e existam termos GO que reflitam com rigor o novo conhecimento, a nível de anotação, bastará actualizar a mesma. Neste trabalho os genes considerados foram descritos tendo em conta apenas a classificação GO associada ao processo biológico.

Semelhanças semânticas entre entidades anotadas:

Uma técnica que tira partido da anotação de entidades biológicas com recurso a uma ontologia biomédica é a determinação das **semelhanças semânticas** entre duas entidades anotadas. As anotações com base em ontologias dão a possibilidade de comparar duas entidades entre si, e a comparação dos termos que anotam cada uma permitem medir quão semelhantes estas são. Graças ao conceito de **conteúdo informativo** (IC), é possível quantificar a semelhança porque, no contexto de uma base de dados de anotação de determinado organismo ou de um conjunto de produtos génicos de vários organismos (tal como a base de dados UniProt KnowledgeBase), o IC de cada termo GO que anota determinada entidade deriva da sua presença relativa (do seu uso relativo) dentro do conjunto de entidades que estamos a comparar, podendo o conjunto ser, por exemplo, uma base de dados de produtos génicos de vários organismos, tal como a UniProt KnowledgeBase.

O IC, intrínseco a cada termo usado na anotação (já que é sempre possível em qualquer classificação calcular a presença relativa de um qualquer termo usado, que pode ir de 0 a 1), é usado

para o cálculo das semelhanças semânticas.

Existem vários métodos para calcular as distâncias semânticas entre entidades classificadas por ontologias. Estes são divididos em dois grupos: Comparações baseadas em nós (vértices) e comparações baseadas em links. Como neste trabalho se usou o primeiro método, apenas este será descrito.

Na comparação baseada em nós é usado o conceito de **conteúdo informativo** (information content, ou IC). O conteúdo informativo pode ser quantificado pela seguinte fórmula:

$IC = -\log p(c)$

em que *p(c)* é a probabilidade de determinado termo (c) do vocabulário GO ser usado na anotação de qualquer entidade presente no corpus (no conjunto de todas as entidades anotadas consideradas). A probabilidade de anotação é estimada pela frequência relativa do termo em questão ao total dos termos usados na anotação das entidades do corpus.[13] [14]. Quanto maior for a probabilidade da ocorrência de um termo particular no corpus, menos informativo esse será. Se $IC = -\log p(c)$, para um termo que anote todas as entidades, a sua probabilidade de anotação é 1 e como tal $-\log p(1) = 0$

Outra propriedade do conteúdo informativo (IC) é que, mesmo quando uma entidade tem termos-filhos e estes hipoteticamente não são usados na anotação de entidades de um determinado corpus, isso não significa que o termo, mais geral que os seus termos-filhos, seja menos informativo, ao nível do seu IC, já que, embora os termos-filhos o sejam justamente por possuírem uma anotação mais específica que o termo parental, o conteúdo informativo é igual a -log p(c) ou seja, reflete a presença relativa de um determinado termo e não o nível de conhecimento científico que o termo significa *per se*. Este pormenor visa clarificar que quando não se dispõe de conhecimento suficiente sobre determinado gene, este poderá ser anotado com termos relativamente gerais, que eventualmente possuem termos-filhos mais específicos, no grafo GO, mas que não podem ser usados para classificar a entidade em questão devido a não se saber qual seria o mais adequado, ou até se algum deles o seria, e mesmo assim o valor de IC poder ser elevado. Por exemplo, se numa determinada base de dados existisse apenas um gene anotado como "transposition", e num momento posterior o conhecimento científico permitisse anotar o mesmo gene como "regulation of transposition, RNA-mediated", igualmente mantendo-se aquele como a única anotada com este termo na base de dados, o IC associado a ambos os termos, na base de dados em questão, teria o mesmo valor, independentemente de um termo possuir mais rigor classificativo que o outro. Independentemente de um termo transmitir mais informação que o outro,

ambos, nas condições acima conceptualizadas, possuem o mesmo poder discriminatório. No cálculo do IC, é a presença relativa de determinado termo na base de dados que é considerada.

Método Resnik:

A semelhança semântica entre dois termos é um conceito que apenas faz sentido quando se comparam duas entidades biológicas entre si com recurso aos termos que as anotam. Não é um "valor próprio" pertencente a cada termo GO porque intrínseco à definição de semelhança está associada uma operação de comparação.

O método Resnik, usado para o cálculo das distâncias semânticas, usa também o conceito de *information content*, mas não é afectado pelas potenciais diferenças de presenças de entidades intermédias entre cada termo e o ancestral comum mais informativo (MICA), usando como valor de *information content* o valor do MICA, de acordo com a seguinte fórmula:

$sim_{Resnik}(c1,c2)=IC(c_{MICA})$

O método de Resnik permite obter o IC do termo comum mais informativo entre duas entidades biológicas calculando as distâncias semânticas entre os valores de cada par de elementos da combinação entre os termos que anotam cada entidade biológica do par.
Um exemplo para a escolha do MICA usando o método de Resnik pode ser visualizado na figura abaixo.

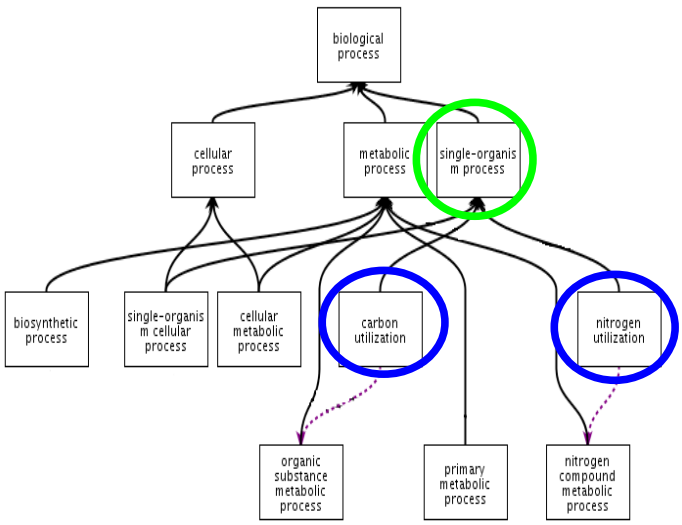

Figura 3: Representação de uma secção do grafo acíclico direcionado dos termos da classe "processo biológico". Os termos "carbon utilization" e "nitrogen utilization" (a azul) têm como termo parental comum (MICA) o termo "single organism process" (a verde), sendo o valor de semelhança semântica dos dois termos igual ao IC do termo parental, ou seja, ao logaritmo da frequência relativa da presença do termo parental na base de dados a que as entidades anotadas pertencem.

Aborgagem pairwise:

O cálculo das distâncias semânticas entre os conjuntos de termos pertencentes às entidades biológicas do par a comparar (pares de genes) pode ser feita com recurso a várias abordagens. Dependendo da abordagem, o valor da semelhança semântica entre duas entidades biológicas poderá ser diferente. Isto deve-se ao facto de cada elemento do par de entidades biológicas ser normalmente anotado por vários termos. A distância semântica é sempre comparada medindo a mesma entre cada termo de uma entidade com todos os termos da outra. Neste trabalho usámos a abordagem "pairwise". Esta abordagem tem duas variantes:

- A técnica *all pairs* ("todos os pares") considera a semelhança entre cada combinação possível dos termos das duas entidades biológicas a comparar, sendo o valor final da semelhança semântica a média, a soma ou o máximo dos valores de semelhança parciais, correspondendo cada um deste todas as semelhanças entre cada par de termos.
- A técnica *best pairs* calcula igualmente a semelhança semântica entre os termos da combinação de cada conjunto associado a cada entidade biológica, mas considera apenas a combinação com o maior ou maiores valores de MICA (most informative common ancestor).

Metodologia

Descrição da metodologia:

Este trabalho pode ser dividido em várias fases:

- Recolher as distâncias semânticas a usar, já que apenas uma parte delas representa as ORF's presentes nos microarrays usados para obter o mapa das presenças génicas, e os recursos computacionais necessários para carregar todas as distâncias semânticas previamente calculadas são bastante exigentes, sendo um desperdício de recursos e de tempo de computação.
- A segunda fase consistiu em criar os módulos de genes. Para tal, usaram-se valores de semelhança entre os membros do conjunto de genes usado. Antes de criar os módulos, excluiram-se todos os genes com potencial de confundimento, por terem uma identificação ambígua no microarray.
- A terceira fase consistiu em aplicar o tratamento estatístico com vista a avaliar se a presença enriquecida de cada módulo nos genomas de estirpes invasivas seria devido ao acaso ou especificamente associada à virulência.
- Na quarta fase, os módulos considerados estatisticamente válidos foram sujeitos a um agrupamento pelo método de hierarchical clustering, de modo a criar grupos de módulos com semelhanças ao nível do conjunto de genes que incluem. Cada grupo foi classificado quanto à sua função biológica, usando a anotação GO e, considerando os seus elementos (produtos génicos) propuseram-se hipóteses quanto ao seu eventual papel na virulência do pneumococcus.

Todo o trabalho foi efetuado usando a linguagem de programação Python versão 2.7, exceto a aplicação do método de "hierarchical clustering", em que foi usada a linguagem Matlab® versão R2010b.

Fonte dos dados genómicos e epidemiológicos:

A Hibridação Genómica Comparativa(CGH) em microarrays compara as estirpes em análise, quanto à presença/ausência de genes. Para tal, usa um conjunto pré-definido de sequências génicas, correspondentes a oligómeros de DNA presentes no microarray, representando cada uma um

determinado gene. Neste caso, cada spot contém várias cópias de oligómeros representativos do genoma completo de três estirpes de *S. pneumoniae (G54, R6 e Tigr4)*.

Os dados usados neste estudo foram obtidos de um estudo epidemiológico com vista a analisar o genoma de amostras de pneumococci isoladas de diversos fluidos biológicos estéreis (sangue, fluidos pleural e cérebro-espinal) com origem em adultos e crianças até 18 anos, recolhidos entre 2001 e 2003. Do total de isolados, foram selecionadas 72 estirpes e posteriormente analisadas com recurso a microarrays representativos do genoma total de 3 estirpes: G54, R6 e Tigr4 [4]. Cada microarray contém 3620 spot's, correspondendo cada oligómero do spot a um gene de uma, duas ou três estirpes representadas no microarray. Os resultados desse trabalho foram disponibilizados em formato digital pelo Instituto de Medicina Molecular da Universidade de Lisboa. Estes incluem a informação sobre a presença/ausência de cada um dos 3620 spot's nas 72 estirpes analisadas por CGH, na forma de uma matriz. Cada linha corresponde a uma estirpe de *S. pneumoniae* e cada coluna a um spot. Para cada coordenada correspondente ao conjunto spot/estirpe, assinala-se a presença ou a ausência de hibridação no respectivo spot, para a estirpe correspondente, com um 1 ou 0, respetivamente.

Cada uma das 72 estirpes foi classificada, de acordo com um estudo epidemiológico anterior [4], como tendo um comportamento invasivo, neutro ou colonizador. Neste trabalho foram apenas utilizados os dados referentes a estirpes colonizadoras e invasivas.

Não foram incluídos no estudo os genes que apesar de presentes no microarray, não tinham uma identificação única, ou seja, cuja sequência representada no microarray correspondia a mais do que um gene. Também foram retirados do estudo os genes que foram detectados em todas as estirpes analizadas por hibridação genómica comparativa e os que nunca foram detectados. Obtiveram-se 778 genes que cumprem todos estes requisitos.

Semelhanças semânticas:

Para este trabalho foi essencial a informação das semelhanças semânticas entre as anotações GO (de processo biológico) correspondentes aos genes presentes nos microarrays. Para cada gene presente no microarray foi procurada a referência UniProt correspondente na base de dados UniProt KnowledgeBase. Nesta base de dados foram também identificados os termos GO, do ramo *biological process* (BP), associados a cada gene. Alguns genes representados no microarray não tinham referência UniProt, ou não tinham anotações GO BP, e consequentemente, não foram utilizados neste trabalho. Para cada par possível de referências UniProt, foi definida a semelhança semântica como sendo a semelhança máxima pela fórmula de Resnik entre dois termos GO, um de

cada termo UniProt. Estes cálculos foram realizados com a ferramenta ProteInOn (http://xldb.fc.ul.pt/biotools/beta/proteinon/).

Construção dos módulos:

A obtenção dos módulos associados à invasividade constitui o objectivo principal deste trabalho. A metodologia consistiu então em desenvolver um algoritmo que agregasse os genes em grupos – os módulos – usando como critério a máxima semelhança funcional entre eles. Criaram-se módulos compostos por 20 e 50 elementos. A dimensão real dos módulos de virulência é desconhecida. Como a procura é realizada com grupos de genes envolvidos em processos biológicos semelhantes, os números 20 e 50 tentam captar processos biológicos que envolvam um número menor ou maior de produtos génicos.

O método de criação do módulo consiste em, a partir de um gene "semente" inicial, ir acrescentando genes selecionados, um a um, com base na semelhança semântica média entre o gene selecionado e os genes que já fazem parte do módulo. Para cada adição de um novo gene ao módulo, todos os genes que não fazem parte do módulo são testados quanto à sua semelhança semântica média com os genes do módulo em construção. É selecionado o gene que tem uma maior semelhança média com os genes já presentes no módulo. Este processo repete-se até cada módulo ter 20 ou 50 elementos na sua constituição.

Todos os 778 genes foram usados como "semente" inical, gerando 778 módulos de 20 genes e 778 módulos de 50 genes.

Análise estatística da associação dos módulos com o comportamento invasivo ou colonizador:

Após a obtenção dos módulos foi efectuada a contagem das presenças dos genes de cada módulo em cada estirpe, discriminada por classe (invasiva ou colonizadora). Das 72 estirpes com dados de hibridação genómica comparativa, 31 foram anteriormente classificadas como invasivas e 16 como colonizadoras. A contagem de genes de cada módulo apenas foi efectuada para estas 47 estirpes. Após a contagem, para cada módulo dispomos de 31 números de genes do módulo presentes em estirpes invasivas e 16 números de genes do módulo presentes em estirpes colonizadoras. O objectivo principal deste trabalho é encontrar módulos que apresentem significativamente maiores contagens de genes num grupo de estripes face ao outro. Ou seja, procuramos dois tipos de módulos: uns cujos genes tendem a estar presentes em simultâneo apenas

em estirpes invasivas, e outros cujos genes tendem a estar presentes em simultâneo apenas em estirpes colonizadoras.

Para detectar se as diferenças no número de genes do módulo presentes entre estirpes invasivas e colonizadoras são estatisticamente significativas ou se podem ser explicadas com a variabilidade natural de presença de genes de estirpe para estirpe, foi aplicado um teste hipergeométrico com *threshold* variável. Este teste tinha sido aplicado com sucesso num trabalho semelhante realizado anteriormente [15]. Resumidamente consiste em variar um *threshold* que define o número de genes do módulo que é necessário estar presente numa estirpe para o módulo ser considerado 'activo'. Para cada *threshold*, o teste hipergeométrico [16] devolve um valor p. Este corresponde à probabilidade de encontrar ao acaso um conjunto de estirpes com uma diferença igual ou maior do que a observada entre as proporções de estirpes com módulos "activos" entre as estirpes invasivas e as colonizadoras. Consideremos um módulo de 20 genes hipotético, em que os números de presenças nas 31 estirpes invasivas são:

20, 20, 20, 19, 19, 19, 19, 19, 19, 19, 18, 18, 18, 18, 18, 17, 17, 17, 17, 17, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 16, 15

e nas 16 estirpes colonizadoras são:

17, 17, 17, 15, 15, 15, 13, 13, 13, 13, 10, 10, 10, 6, 6, 5

Se o *threshold* for 17, existem 20 estirpes invasivas com o módulo "activo e 3 estirpes colonizadoras com o módulo "activo". A proporção 20/31 é superior a 3/16, o que sugere que o facto de uma estirpe ter o módulo 'activo' pode contribuir para o comportamento invasivo. O teste hipergeométrico permite definir se a diferença entre as proporções é significativa. A execução do teste produz um valor p que corresponde à probabilidade de encontrar uma diferença de proporções igual ou maior à que foi observada supondo que a presença do módulo é independente do comportamento da estirpe. Se o valor p for muito pequeno, podemos inferir que a presença do módulo está significativamente associada ao comportamento invasivo das estirpes. Neste caso o valor p para o threshold 17 é 0.00329. O teste é repetido para todos os thresholds possíveis, sendo o valor p final do módulo o menor valor p encontrado. Quando o menor valor p for inferior a 0.05, consideramos que o módulo está associado ao comportamento invasivo (como no exemplo dado) ou ao comportamento colonizador (quando a diferença de proporções é no sentido contrário).

Agrupamento dos módulos significativos e caracterização:

Dado o modo como foram criados os módulos, é expectável encontrar módulos parcialmente redundantes. Os módulos em que os respetivos genes 'semente' possuem elevada semelhança semântica entre si tendem a partilhar elementos. De forma a tentar evitar esta redundância, foi realizado um agrupamento hierárquico dos módulos significativamente associados à invasividade (ou à colonização). Entre cada dois módulos foi definida uma medida de distância que consiste na fração de genes que diferem entre os dois módulos. Estas distâncias foram utilizadas por um algoritmo de agrupamento hierárquico (*hierarquical clustering*) com ligação média (*average linkage*) para gerar um dendrograma. Os dendrogramas gerados foram inspecionados visualmente para selecionar o número de grupos de módulos que maximizava a separação entre grupos e a homogeneidade dentro de cada grupo. Após a definição dos grupos de módulos, foram organizadas listas com os termos GO de processo biológico associados aos genes pertencentes aos módulos do grupo. Foram considerados mais relevantes os termos com maior frequência, quer por serem comuns a vários genes dentro do mesmo módulo, quer pela maior repetição dos genes anotados com esse termo em vários módulos do mesmo grupo. Em princípio os termos que aparecem em mais genes do módulo e que surgem num maior número de módulos do grupo serão os responsáveis pela significância estatística da associação dos módulos ao comportamento invasivo (ou ao comportamento colonizador).

Resultados e discussão:

Na pesquisa de módulos associados à invasividade, foram encontrados 66 módulos de 20 genes e 58 módulos de 50 genes com associação significativa. Não foi encontrado nenhum módulo, quer de 20 ou de 50 genes, associados a um comportamento colonizador.

Após o agrupamento dos módulos com 50 elementos associados à invasividade, o dendrograma respectivo foi dividido em 5 grupos. O dendrograma representativo do agrupamento dos módulos de 20 elementos foi dividido em 6 grupos.

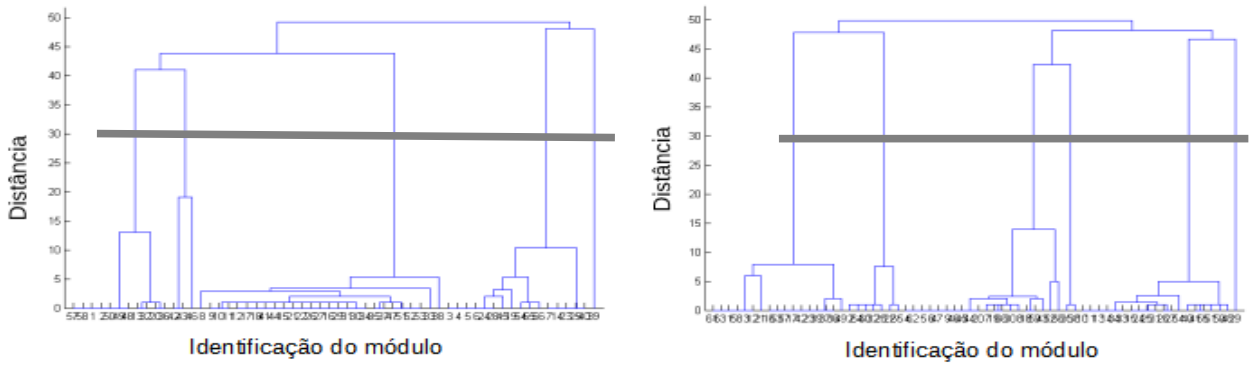

Figura 4: Dendrogramas resultantes da aplicação do método de 'hierarchical clustering' aos módulos de 50 e 20 elementos, usando a abordagem 'top-down'. A linha horizontal que atravessa cada dendrograma representa a zona de corte escolhida, sendo o número de grupos correspondente ao número de ramos atravessados pela linha. O eixo Y representa a percentagem de genes diferentes entre os módulos, para o mesmo nível; À esquerda: dendrograma representativo dos módulos com 50 elementos. À direita: dendrograma representativo dos módulos com 20 elementos.

Após a obtenção dos grupos, estes foram verificados manualmente de modo a fazer corresponder aos genes de cada um deles as respetivas anotações GO. Para tal, consultou-se a base de dados UniProtKB, de modo a, a partir da referência UniProt associada a cada gene, obter a ou as anotações GO que o classificam. As anotações GO foram então recolhidas e associadas ao respetivo gene e, para cada grupo, criou-se uma lista com todos os termos GO que anotam cada um dos genes. Para cada grupo, os termos foram ordenados por ordem decrescente de presença no grupo, sendo que o termo mais presente é o termo que anota mais genes no grupo em questão e o termo menos presente é o que anota menos genes no grupo. Posteriormente, e para cada grupo, pesquisou-se a literatura com o objetivo de encontrar informação sobre processos biológicos associados à invasividade que sejam coerentes com o conjunto dos processos biológicos descritos pelos respetivos termos GO.

Os processos biológicos associados aos grupos encontrados são predominantemente relacionados com processos de síntese e degradação proteica, metabolismo e transporte de carbohidratos, processos associados ao processamento da cápsula bacteriana, e expressão génica, tanto nos módulos de 20 elementos como nos com 50 elementos.

Módulos com 50 elementos:

1º grupo – síntese proteica; parede celular:

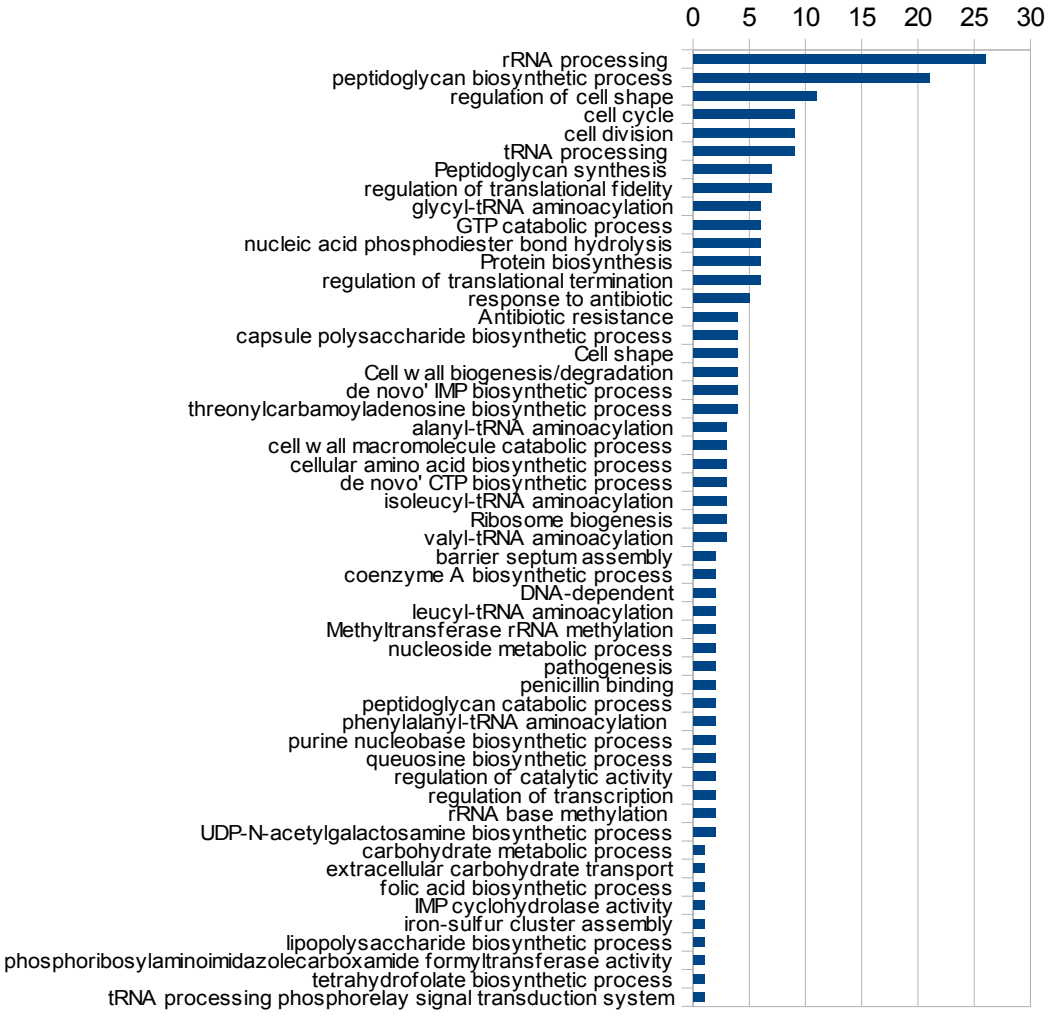

Figura 5: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Este grupo tem predominantemente genes associados à síntese proteica e a processos biológicos relacionados com a parede celular. O papel essencial do processo de síntese proteica e, mais precisamente, o do ribossoma torna-o num alvo terapêutico privilegiado. Em particular, os

ribossomas procariotas são estruturalmente diferentes dos seus homólogos eucariotas, o que permite minimizar efeitos secundários no paciente devidos ao uso terapêutico dos antibióticos. Os ribossomas possuem ainda motivos mais complexos que os tornam alvos mais específicos que outros tipos de RNA celulares menos estruturados [17]. Existem várias famílias de antibióticos que visam precisamente sub-unidades ribossomais, e algumas delas apenas interagem com o rRNA [18].

Nos streptococci, assim como em outras bactérias, a parede celular tem um papel estrutural e de proteção, sendo também um alvo terapêutico para várias famílias de antibióticos β-lactâmicos, das quais a penicilina é o exemplo mais conhecido. Noutra perspetiva, estudos imunológicos revelaram que a proteína pneumocócica de superfície A (PspA) está presente em todas as estirpes de S. pneumoniae até agora conhecidas [19], estando esta localizada na parede celular [20]. Esta proteína está envolvida na proteção da bactéria contra o sistema de complemento do hospedeiro. Existem também evidências de que a secreção e a distribuição de fatores de virulência pelos pneumococci estão inseridos num processo altamente organizado, em que um microdomínio focal localizado no citoplasma designado como ExPortal [21]- coincide espacialmente com o local da síntese *de novo* do peptidoglicano [22], [23] . Um maior conhecimento sobre a interdependência destes dois processos poderá ser um potencial fonte de alvos terapêuticos.

2º grupo - biossíntese:

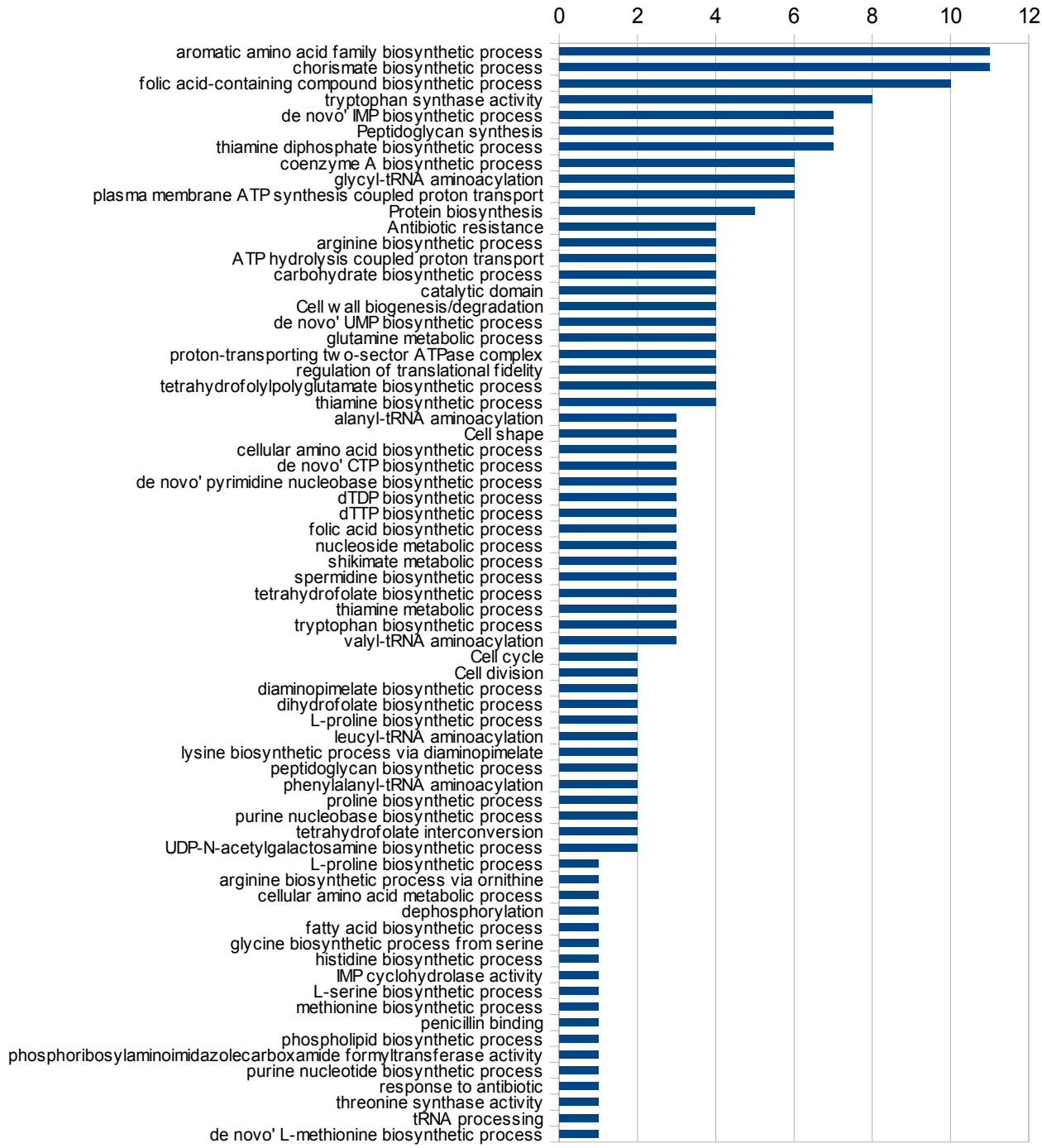

Figura 6: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

As entidades contidas neste grupo evidenciam propriedades associadas à biossíntese, tendo-se encontrado um maior número de anotações ligadas à via biossintética do chiquimato. Esta via

metabólica liga o metabolismo dos carbohidratos à biossíntese dos compostos aromáticos [24]. Esta via é reconhecida como um alvo terapêutico, por se encontrar presente em bactérias, fungos, plantas e em certas classes de parasitas, mas não em humanos [25]. Uma estirpe de *S. suis* mutante para um operão que o torna auxotrófico para os aminoácidos aromáticos revelou-se imunogénica quando usada como base para a criação de uma vacina contra a estirpe WT, mas avirulenta. Além disso, a mutação faz com que a bactéria fenotipicamente não tenha cápsula [26]. O IMP é usado para a síntese das purinas e, especificamente em certas estirpes de Streptococci, é um percursor dos nucleótidos de guanina. Observou-se em modelos de murganho uma diminuição da virulência, usando como agente infeccioso uma estirpe mutante de *S. suis* com knock-out para o gene da inosina 5-MP desidrogenase. Adicionalmente, este mutante, *in vitro*, apresentava um pH alterado e uma taxa de crescimento reduzida em relação à estirpe WT [27].

O acetil CoA é um co-fator envolvido na síntese e oxidação dos ácidos gordos. Em *S. pneumoniae*, esta molécula desempenha um papel essencial na formação da parede celular, estando envolvido na via da síntese do UDP-N-acetilglucosamina, um substrato para as vias sintéticas do lípido-A e do peptidoglicano [28].

A timina difosfato é um co-fator envolvido em vários processos metabólicos. Na família Streptococcus (por exemplo *S. sanguis*), observou-se o uso da timina como co-fator do piruvato oxidase [29].

3º grupo – degradação proteica:

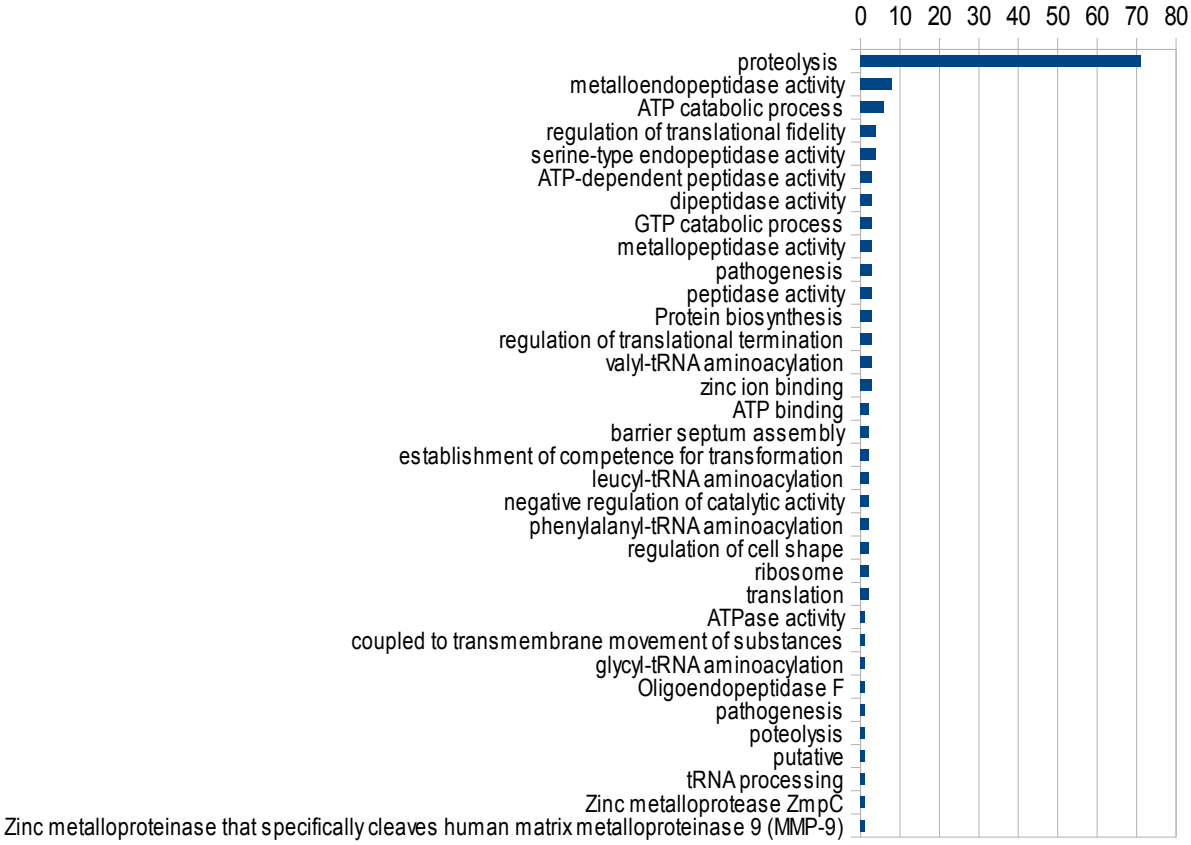

Figura 7: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Este grupo contém entidades anotadas com termos relacionados com a degradação proteica. As metalopeptidades são uma classe de proteínas com atividade enzimática, presentes na superfície de várias bactérias do género *Streptococcus*. Existem evidências de que estas proteínas desempenham um papel na invasividade dos *streptococci*. Num estudo onde se efetuou a análise genética de 218 isolados, representativos de 35 serotipos de pneumococcus, recolhidos de doentes com estados patológicos causados pela bactéria, todos eles possuíam genes codificantes das metalopeptidades zmpB e igA [30]. Uma porção variável dos isolados possuíam outros dois tipos destas proteases. Observou-se, usando um modelo de colonização em culturas de células epiteliais respiratórias que a protease IgA1, com especificidade para as imunoglobulinas humanas A1, cliva esta imunoglobulina na sua região variável, expondo assim uma região catiónica do anticorpo que servirá de promotor para a adesão do pneumococcus ao tecido epitelial, sugerindo a observação que

a região catiónica dos anticorpos clivados pela peptidase IgA1 poderá neutralizar a carga catiónica da cápsula do pneumcoccus, promovendo assim a invasividade [31].

4º grupo – metabolismo de carbohidratos e aminoácidos:

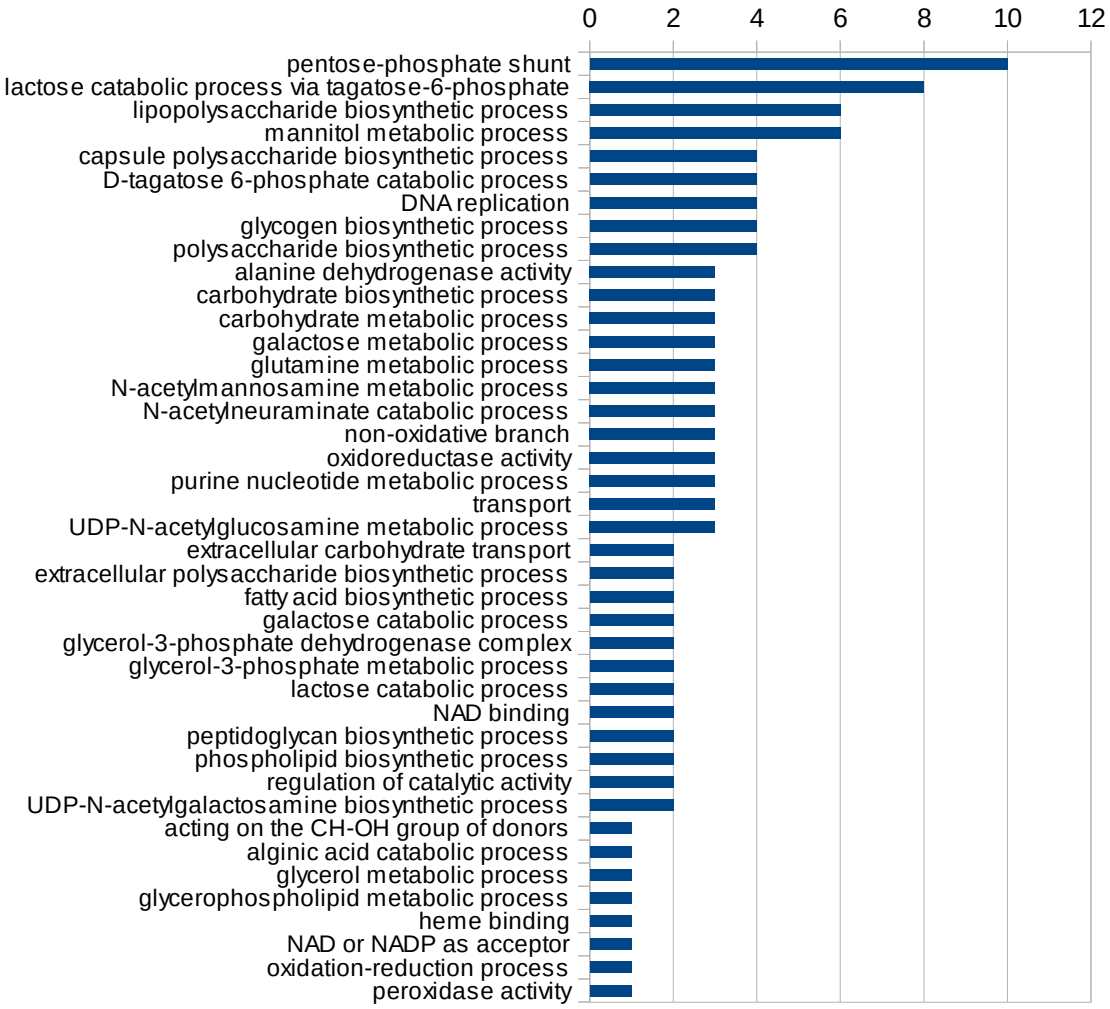

Figura 8: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Este grupo contém entidades relacionadas com o metabolismo de carbohidratos e aminoácidos. A via dos fosfatos de pentose é uma via alternativa à via da glicólise, estando a oxidação dos carbohidratos nesta via acoplados à síntese de NADPH, tornando-a numa das principais fontes de equivalentes redutores.

Os lipopolissacáridos estão envolvidos na colonização do pneumococcus. O manitol faz parte dos

polissacáridos da cápsula, um agente *major* de virulência em *S. pneumoniae*. Neste, o d-manitol foi encontrado nos polissacáridos da cápsula, em vários serotipos, tendo sido propostos os genes mnp1 e mnp2, localizados no cluster génico CPS (capsule polysaccharide). Noutros Streptococci, o manitol é conhecido por possibilitar um maior rendimento da via glicolítica, especificamente em *S. mutans*, em condições anaeróbias [32].

5º grupo – transporte de carbohidratos:

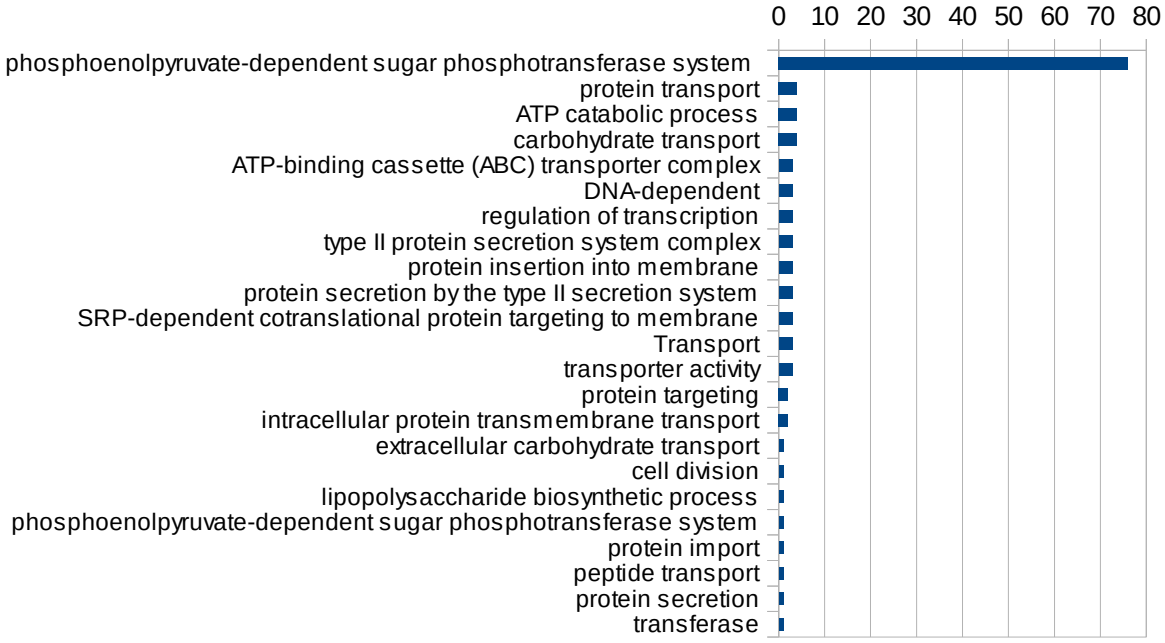

Figura 9: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Os carbohidratos são a fonte exclusiva de carbono do *S. pneumoniae*, representando os sistemas de transporte destes 30% do total dos sistemas de transporte do *S. pneumoniae*.
O sistema de fosfotransferase, em bactérias Gram-positivas, incluindo o género Streptococcus, está envolvido em vários processos biológicos, tais como a quimiotaxia, a transcrição génica e a repressão catabólica [33]. Observou-se que mutantes para certos transportadores de açúcares apresentam virulência atenuada in vivo [34] [35].

Módulos com 20 elementos

1º grupo - via do ácido fólico:

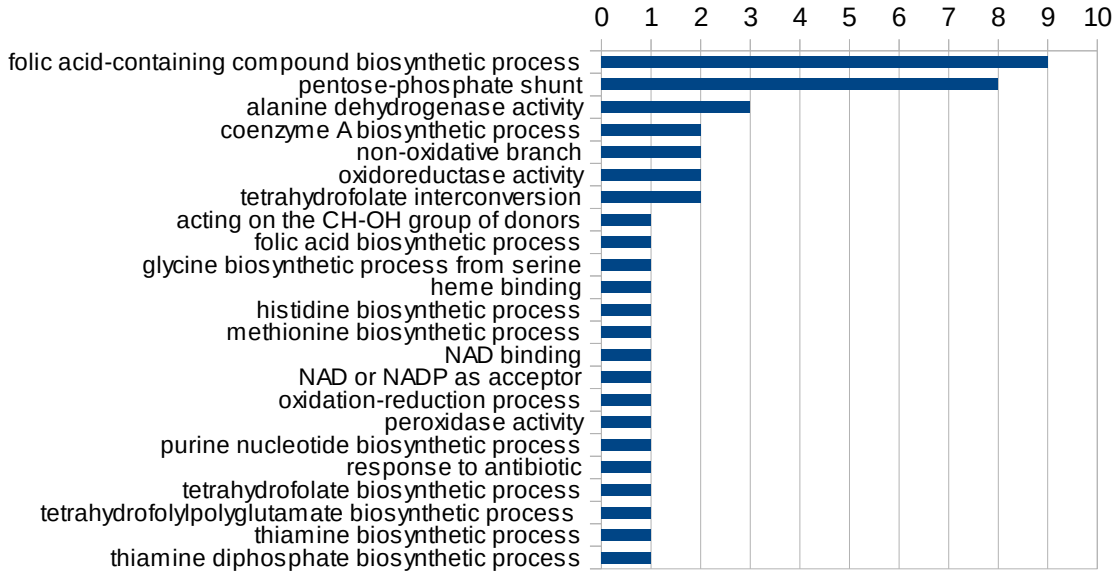

Figura 10: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Os derivados do folato desempenham um papel essencial como cofatores na síntese de de ácidos nucleicos e aminoácidos em todas as células de todos os domínios da vida [36]. As interferências na síntese e uso dos folatos e seus derivados tem sido um dos alvos terapêuticos estudado no combate a infeções bacterianas, sendo as sulfonamidas, inibidores competitivos do dihidropteroato sintetase, um exemplo. O folato desempenha também um papel na taxa de crescimento populacional do pneumococcus. Foi observado em bibliotecas de *S. pneumoniae* mutantes que o dihidrofolato/ folipoliglutamato sintase, codificado pelo gene *folC,* desempenha um papel na fase de crescimento bacteriana ao manter os níveis das cadeias de poliglutamil-folato, que vão sendo depletadas durante a fase 'log' do crescimento bacteriano em meios com baixas taxas de CO_2 [37].

2º grupo - processamento da cápsula:

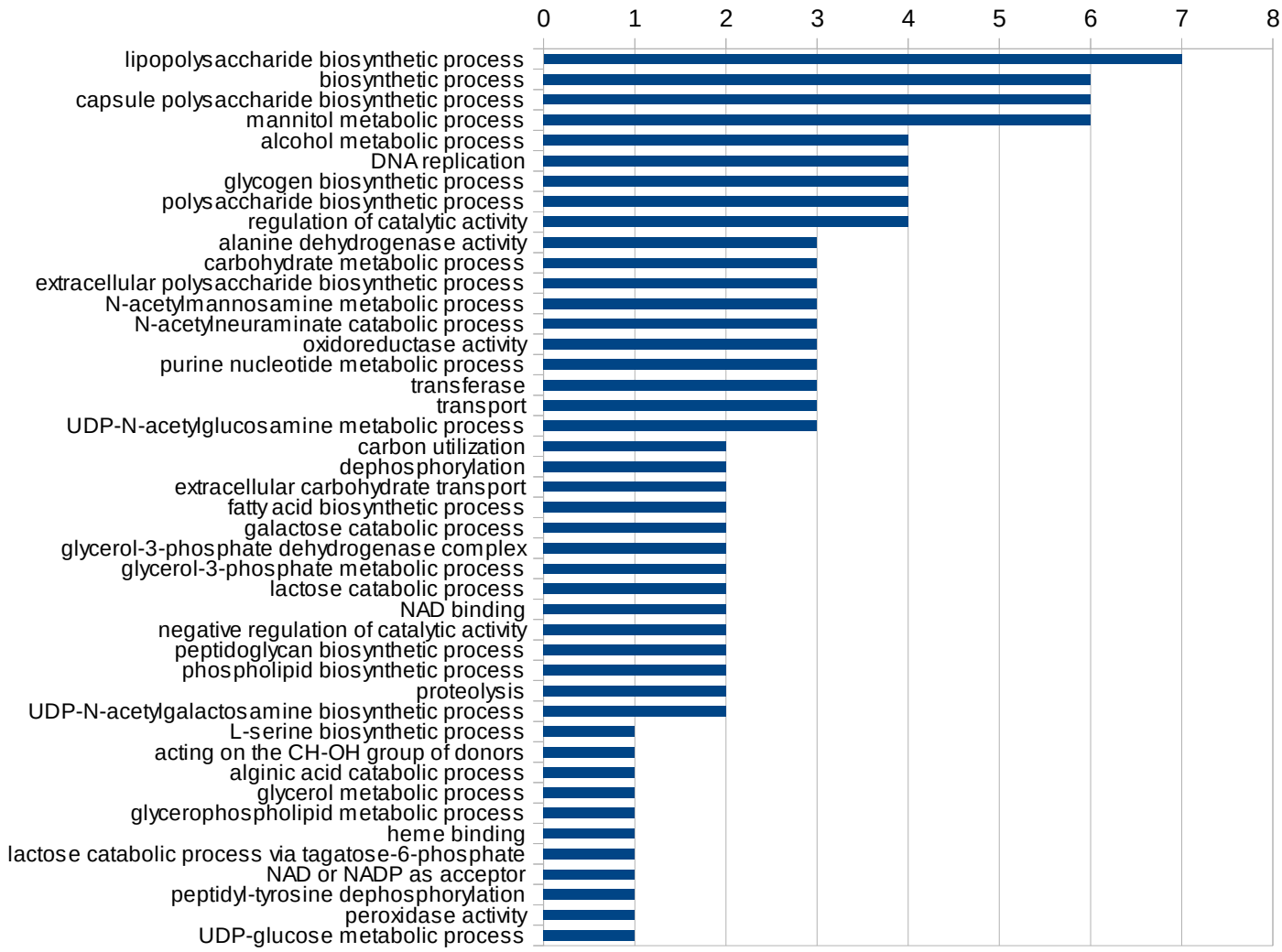

Figura 11: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

A existência de cápsula em *S. pneumoniae* é uma condição *sine qua non* para a invasividade, tendo sido observada a co-relação entre a sua presença e/ou o seu serotipo com a capacidade da respectiva estirpe de causar doença. O pool do total de alelos responsáveis pelos diferentes serotipos da cápsula polissacárida pneumocócica é de cerca de 450 kb, o equivalente a 25% do genoma individual médio pneumocócico . Existem mais de 90 serotipos capsulares diferentes de pneumococcus. Desde logo, a existência de cápsula evita a eliminação da bactéria pelo muco do hospedeiro, facilitando o acesso desta à zona epitelial [38]. a sua existência inibe os processos de fagocitose por neutrófilos e o processo de lise celular promovido pelo sistema de complemento,

nomeadamente o processo de deposição do complexo C3b/iC3B na membrana celular da bactéria [39]. O pneumococcus consegue evitar o processo de fagocitose, por este ser dependente da ação, a montante, do sistema de complemento [40][41].

3º grupo – dúbio:

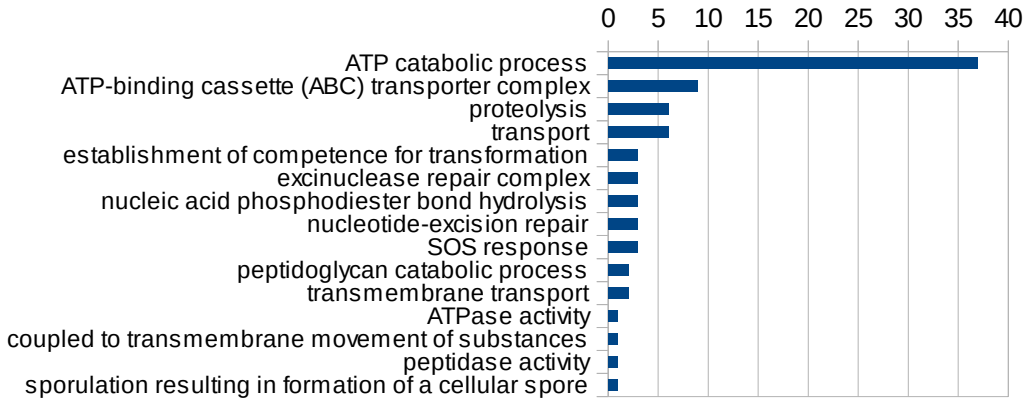

Figura 12: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Os elementos deste conjunto de processos biológicos não permitem caracterizar o grupo como associado a um processo específico, como se pode ver pela presença de genes relacionados com o metabolismo, com o transporte e outros processos celulares relacionados com a expressão génica, a competência e a adaptação.

4º grupo – carbohidratos:

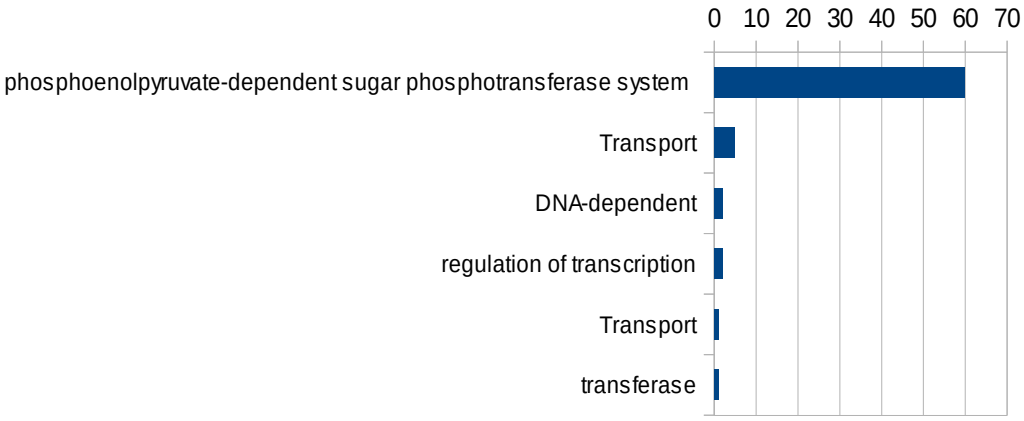

Figura 13: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

Os carbohidratos são uma fonte de energia essencial para o crescimento e divisão celular dos streptococci. A estirpe R6 de *S. pneumoniae*, por exemplo, consegue metabolizar várias famílias de carbohidratos, incluíndo monossacáridos (glucose, manose, frutose e galactose), dissacáridos (sacarose, lactose, trealose, maltose e celobiose) e trissacáridos, tais como a rafinose e oligossacáridos (inulina). Em *S. pneumoniae*, estudos genómicos identificaram 21 sistemas de fosfotransferase, 7 transportadores "ATP Binding Cassette" de uptake de carbohidratos, um sistema simporte Na^+/soluto e uma permease [42].

O sistema de fosfotransferase dependente do fosfoenolpiruvato, ou "phosphoenolpyruvate-dependent phosphotransferase system" (PTS) cataliza a incorporação e a fosforilação de carbohidratos para e no citoplasma das bactérias, promovendo assim um gradiente de concentração dos carbohidratos dentro da célula, devido à fosforilação daquele [43]. O sistema possui três enzimas essenciais para o seu funcionamento: O enzima EI, que recebe o grupo fosforil do fosfoenolpiruvato, o enzima Hpr (Heat-stable Protein), que recebe o fosfato do enzima I, e o complexo enzimático membranar EII, que desempenha a função de permease e que pode ser constituído de 1 a 4 domínios proteicos, podendo ou não estarem covalentemente ligados. Estes catalizam a fosforilação do carbohidrato ao atravessar a permease [44]. Desde a descoberta deste sistema, em 1964 [45], que se descobriram outras funções desempenhadas pelo sistems PTS, predominantemente envolvidas na regulação transcricional e na regulação de mecanismos metabólicos [46], tais como a regulação de enzimas do catabolismo de carbohidratos, de permeases de açúcares e da adenilato ciclase.

5º grupo – tradução:

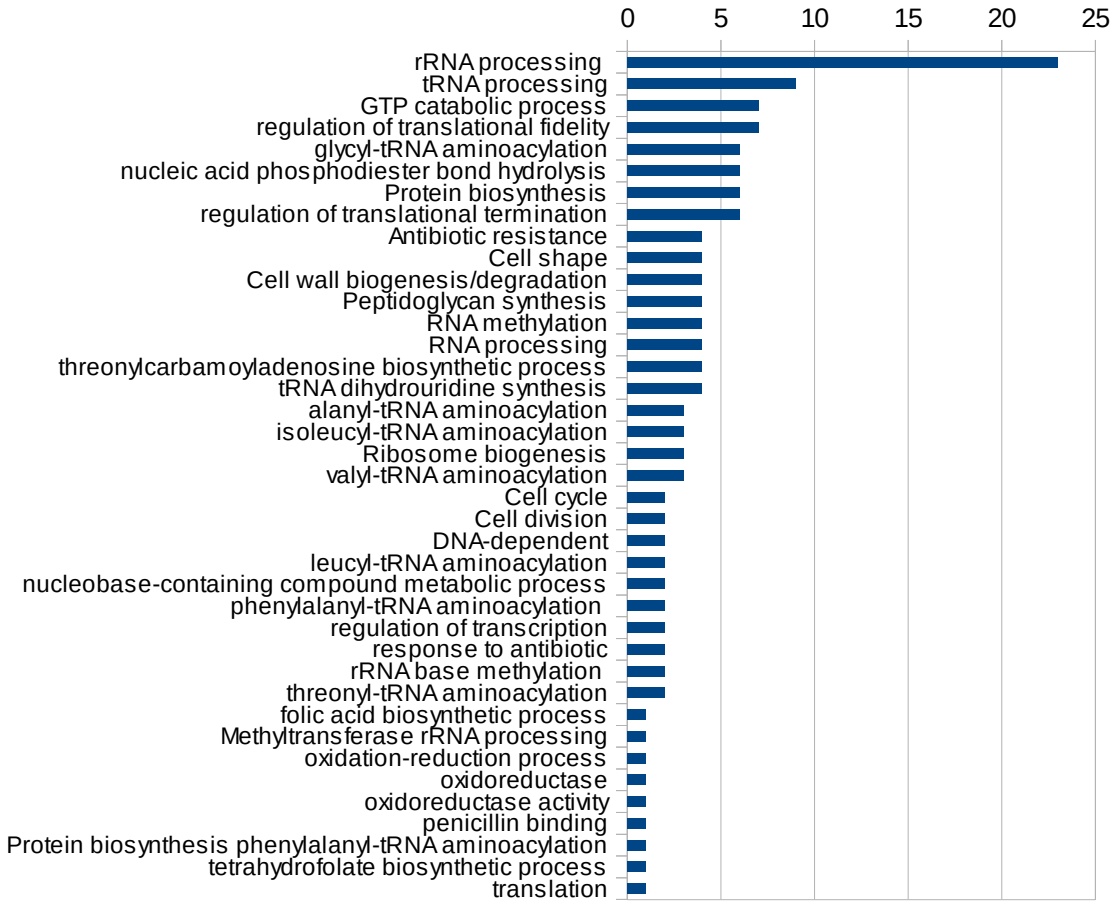

Figura 14: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

A tradução é o processo em que a informação contida no mRNA é usada para sintetizar polipéptidos, por via dos ribossomas e de acordo com o código genético de cada organismo. Pode ser o último passo da via da expressão génica, ou os polipéptidos poderão ainda ser alvo de modificações pós-traducionais.

Sendo as proteínas ubíquas quanto à sua localização celular e quanto à sua intervenção em todo o espectro de funções biológicas, estas, e especialmente os sistemas a montante que as sintetizam, são alvos terapêuticos importantes, porque afetando o sistema de tradução, o processo a jusante de síntese proteica também é afetado [47].

Os sistema de tradução procariota tem diferenças em relação ao sistema de tradução eucariota, nomeadamente por este último envolver mais componentes proteicos e o processo ser ligeiramente diferente [48]. Estas diferenças podem ser exploradas na busca de novos alvos terapêuticos, já que

as diferenças entre os dois sistemas aumentam a probabilidade de que moléculas com potencial terapêutico e afinidade para componentes ribossomais e/ou envolvidos no processo de tradução dos procariotas não tenham a mesma afinidade para componentes ribossomais eucariotas, diminuindo assim potenciais efeitos secundários no paciente.

6º grupo – reparação do DNA / adaptação:

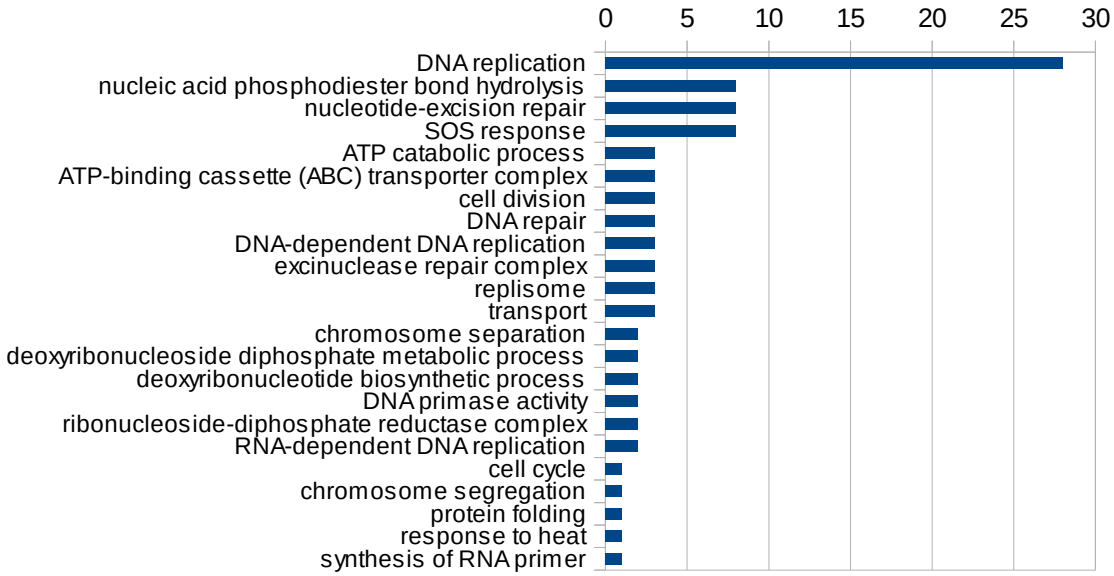

Figura 15: Número de presenças de termos GO no grupo. O eixo y representa o número de presenças de cada termo no grupo.

A capacidade de adaptação ao meio é uma necessidade para os microorganismos, incluíndo o pneumococcus. A sua taxa de mutabilidade das bactérias é favorecida pela sua competência natural e pela distribuição do seu pan-genoma intra e inter espécie. Vários mecanismos de variabilidade genética são conhecidos, nomeadamente rearranjos e aquisição exógena de DNA: o mecanismo de Horizontal Gene Transfer, ou HGT. Este mecanismo permite a incorporação de DNA exógeno, na forma de cadeia simples, e posterior integração, com taxa de sucesso variável e tolerância a regiões não homólogas de até cerca de 10% [49]. Uma das consequências da adaptação pneumocócica foi a emergência de estirpes com resistência aos antibióticos, incluindo a família de antibióticos dos beta-lactâmicos (especificamente, proteínas de ligação à penicilina com baixa afinidade para os beta-lactâmicos). Existem evidências que a recombinação inter-específica será responsável pela evolução dos factores de virulência e da resistência aos antibióticos [50]. Observou-se, em algumas estirpes

resistentes à penicilina, que os próprios genes que proporcionam resistência à penicilina são mosaicos, com partes da sua sequência que correspondem a sequências do gene homólogo noutras espécies de streptococci. Foi também observada uma diversidade considerável entre essas mesmas variedades alélicas, existindo informação que, numa coleção de 45 estirpes Pen^R, foram detetados 18 mosaicos diferentes do gene pbp2x e 16 mosaicos diferentes do gene pbp2b [51]. Os dadores prováveis de sequências com impacto na evolução dos genes de resistência à penicilina (pbp) serão *S. oralis* e *S. mitis* [51][52], tendo já sido demonstrado experimentalmente a transformação dos genes $pbp2^a$ e pbp1b do pneumococcus com sequências originárias de *S. mitis*, e também a aquisição de resistência à optocina por parte de estirpes de pneumococcus, sendo o dador *S. oralis* [53].

Num estudo recente, comparou-se a expressão génica de fatores de virulência entre estirpes de *S. suis* com acesso a diferentes tipos de carbohidratos. Observou-se que a presença de carbohidratos (alfa-glucano) que não apenas a glucose, normalmente presentes na orofaringe do hospedeiro de *S. suis* (suínos jovens) estava correlacionada com a expressão de 19 factores de virulência de *S. suis*, envolvidos na adesão e invasão das células epiteliais do hospedeiro. Adicicionalmente, observou-se, também nas mesmas condições, que a toxina suilisina é expressa numa ordem de grandeza superior nas estirpes com acesso a pululano e a baixos níveis de glucose [54]. *S. suis* tem como hospedeiro preferencial os suínos mas está a emergir como um patogéneo em humanos [55], [56].

Conclusão:

O trabalho desenvolvido no decurso desta tese teve como objetivo identificar vias tendencialmente presentes em estirpes de *S. pneumoniae* com maior potencial de invasividade. A abordagem usada foi criar entidades – módulos de genes – que se caracterizam pelos seus constituintes terem uma semelhança semântica média elevada. O modo de medir a semelhança foi comparar os termos que classificam cada gene, termos esses que pertencem a uma ontologia. Quanto mais semelhantes forem os termos entre os genes, mais semelhante será a sua função, em princípio. Quanto mais genes pertencentes a estirpes consideradas invasivas estiverem presentes num módulo, maior será a probabilidade dos genes desse módulo estarem associados à invasividade.

Foram usados vários critérios de modo a obter um conjunto de genes que sejam informativos. Para tal, os dados foram sujeitos a processos de exclusão, de modo a obtermos um conjunto o mais informativo possível e excluir informação dúbia. Acreditamos que a aplicação dos testes estatísticos permitiu excluir as presenças de genes pertencentes a estirpes invasivas nos módulos devido ao acaso.

Pela análise dos grupos obtidos por "hierarquical clustering", observa-se que os processos biológicos anotados com recurso ao vocabulário GO descrevem processos que, com base na literatura, são processos importantes que co-ocorrem com a invasividade e também refletem acontecimentos que estão ligados ao sucesso da invasividade. Acreditamos, como tal, que a construção de módulos – neste trabalho o foco é a associação à invasividade - com recurso às semelhanças semânticas é uma técnica promissora para estudar sistemas e processos complexos. Um dos problemas é a dificuldade em eliminar os genes falsos positivos, sendo uma das causas a anotação pouco rigorosa das entidades biológicas em análise, que poderá ser minimizado com a melhoria da qualidade das anotações, à medida que novos conhecimentos forem surgindo e as anotações automáticas forem sendo substituídas por anotações baseadas em informação mais rigorosa e/ou que os algoritmos preditivos da função dos respetivos genes forem evoluindo.

Este método tem, como tal, a vantagem acrescida de usar como fonte de informação entidades que estão em permanência sujeitas a revisão pela comunidade científica. Construções de módulos com base em informação anterior e menos rigorosa poderão ser reconstruídos sem grande esforço, à medida que nova informação sobre as entidades biológicas vai surgindo, ou até em caso de descobertas que afetem de forma significativa o conhecimento sobre as entidades em estudo. Como

tanto o processo de cálculo como o acesso às anotações nas bases de dados são automatizados (embora neste trabalho não tenha sido implementado nenhum processo de aquisição automático de anotações, tendo estas sido obtidas manualmente), os investigadores ganham mais tempo para a investigação e discussão científicas e dispendem menos com as aplicações informáticas associadas à obtenção dos módulos.

Referências:

[1] A. Kadioglu, J. N. Weiser, J. C. Paton, and P. W. Andrew, "The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease.," *Nat. Rev. Microbiol.*, vol. 6, no. 4, pp. 288–301, Apr. 2008.

[2] M. Martcheva, B. M. Bolker, and R. D. Holt, "Vaccine-induced pathogen strain replacement: what are the mechanisms?," *J. R. Soc. Interface*, vol. 5, no. 18, pp. 3–13, Jan. 2008.

[3] J. Mehtälä, M. Antonio, M. S. Kaltoft, K. L. O'Brien, and K. Auranen, "Competition between Streptococcus pneumoniae strains: implications for vaccine-induced replacement in colonization and disease," *Epidemiology*, vol. 24, no. 4, pp. 522–529, 2013.

[4] R. Sá-Leão, F. Pinto, S. Aguiar, S. Nunes, J. a Carriço, N. Frazão, N. Gonçalves-Sousa, J. Melo-Cristino, H. de Lencastre, and M. Ramirez, "Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype.," *J. Clin. Microbiol.*, vol. 49, no. 4, pp. 1369–75, Apr. 2011.

[5] A. P. Steenhoff, S. S. Shah, A. J. Ratner, S. M. Patil, and K. L. Mcgowan, "Emergence of Vaccine-Related Pneumococcal Serotypes as a Cause of Bacteremia," vol. 19104, pp. 907–914, 2006.

[6] F. Baquero, "Antibiotic consumption and resistance selection in Streptococcus pneumoniae," *J. Antimicrob. Chemother.*, vol. 50, no. 90003, pp. 27–38, Dec. 2002.

[7] L. Good and J. E. M. Stach, "Synthetic RNA silencing in bacteria - antimicrobial discovery and resistance breaking.," *Front. Microbiol.*, vol. 2, no. September, p. 185, Jan. 2011.

[8] J. P. Lynch, G. G. Zhanel, and D. Ph, "Streptococcus pneumoniae : Epidemiology , Risk Factors , and Strategies for Prevention," 2000.

[9] G. D. Ehrlich, A. Ahmed, J. Earl, N. L. Hiller, J. W. Costerton, P. Stoodley, J. C. Post, P. Demeo, and F. Z. Hu, "The Distributed Genome Hypothesis as a Rubric for Understanding Evolution in situ During Chronic Bacterial Biofilm Infectious Processes," *FEMS Immunol Med Microbiol.*, vol. 59, no. 3, pp. 269–279, 2011.

[10] N. L. Hiller, B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N. E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S. D. Bentley, J. C. Post, G. D. Ehrlich, and F. Z. Hu, "Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome.," *J. Bacteriol.*, vol. 189, no. 22, pp. 8186–95, Nov. 2007.

[11] C. Donati, N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V Angiuoli, M. Oggioni, J. C. D. Hotopp, F. Z. Hu, D. R. Riley, A. Covacci, T. J. Mitchell, S. D. Bentley, M. Kilian, G. D. Ehrlich, R. Rappuoli, E. R. Moxon, and V. Masignani, "Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species," *Genome Biol.*, vol. 11, no. 10, p. R107, 2010.

[12] A. Coulet and M. Smaïl-Tabbone, "Suggested ontology for pharmacogenomics (SO-Pharm): modular construction and preliminary testing," *Move to Meaningful Internet Syst. 2006 OTM 2006 Work.*, vol. 4277, pp. 648–657, 2006.

[13] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies.," *PLoS Comput. Biol.*, vol. 5, no. 7, p. e1000443, Jul. 2009.

[14] P. Resnik, S. M. Laboratories, and T. E. Drive, "Using information content to evaluate semantic

similarity in a taxonomy," vol. 1, 1995.

[15] R. Catarino, "Search for coherent gene modules that predict Streptococcus pneumoniae strain invasiveness," Faculdade de ciências da Universidade de Lisboa, 2012.

[16] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?," *Bioinformatics*, vol. 23, no. 4, pp. 401–7, Feb. 2007.

[17] A. Yonath, "Antibiotics targeting ribosomes: resistance, selectivity, synergism and cellular regulation.," *Annu. Rev. Biochem.*, vol. 74, pp. 649–79, Jan. 2005.

[18] L. S. McCoy, Y. Xie, and Y. Tor, "Antibiotics that target protein synthesis.," *Wiley Interdiscip. Rev. RNA*, vol. 2, no. 2, pp. 209–32, 2011.

[19] M. J. Crain, W. D. Waltman, J. S. Turner, J. Yother, D. F. Talkington, L. S. McDaniel, B. M. Gray, and D. E. Briles, "Pneumococcal surface protein A (PspA) is serologically highly variable and is expressed by all clinically important capsular serotypes of Streptococcus pneumoniae.," *Infect. Immun.*, vol. 58, no. 10, pp. 3293–9, Oct. 1990.

[20] B. Y. L. S. Mcdaniel, G. Scott, J. F. Kearney, and D. E. Briles, "Monoclonal antibodies against protease-sensitive pneumococcal antigens can protect mice from fatal infection with Streptococcus pneumoniae," vol. 160, no. August, pp. 386–397, 1984.

[21] J. Rosch and M. Caparon, "A microdomain for protein secretion in Gram-positive bacteria.," *Science*, vol. 304, no. 5676, pp. 1513–5, Jun. 2004.

[22] P. Hu, Z. Bian, M. Fan, M. Huang, and P. Zhang, "Sec translocase and sortase A are colocalised in a locus in the cytoplasmic membrane of Streptococcus mutans.," *Arch. Oral Biol.*, vol. 53, no. 2, pp. 150–4, Feb. 2008.

[23] A. Raz and V. a Fischetti, "Sortase A localizes to distinct foci on the Streptococcus pyogenes membrane.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 47, pp. 18549–54, Nov. 2008.

[24] A. Rev, P. Physiol, P. Mol, B. Downloaded, K. M. Herrmann, and L. M. Weaver, "The shikimate pathway," pp. 473–503, 1999.

[25] J. Rohr, "Shikimic Acid. Metabolism and Metabolites. Von E. Haslam. Wiley, Chichester, 1993. 387 S., geb. 75.00 £. – ISBN 0-471-93999-4," *Angew. Chemie*, vol. 107, no. 5, p. 653, 1995.

[26] N. Fittipaldi, B. D. Amours, S. Lacouture, and M. Gottschalk, "Potential use of an unencapsulated and aromatic amino acid-auxotrophic Streptococcus suis mutant as a live attenuated vaccine in swine," vol. 25, pp. 3524–3535, 2007.

[27] X. Zhang, K. He, Z. Duan, J. Zhou, Z. Yu, Y. Ni, and C. Lu, "Microbial Pathogenesis Identification and characterization of inosine 5-monophosphate dehydrogenase in Streptococcus suis type 2," *Microb. Pathog.*, vol. 47, no. 5, pp. 267–273, 2009.

[28] C. R. H. Raetz, *Escherichia coli and Salmonella: Cellular and Molecular Biology*. 1996, pp. 1035–1063.

[29] F. M. Letters and E. Fem, "Pyruvate oxidase activity dependent on thiamine pyrophosphate , flavin adenine dinucleotide and orthophosphate in Streptococcus sanguis," vol. 25, pp. 53–56, 1985.

[30] R. Camilli, E. Pettini, M. Del Grosso, G. Pozzi, A. Pantosti, and M. R. Oggioni, "Zinc metalloproteinase genes in clinical isolates of Streptococcus pneumoniae : association of the full array with a clonal cluster comprising serotypes 8 and 11A Printed in Great Britain," pp. 313–321, 2006.

[31] J. N. Weiser, D. Bae, C. Fasching, R. W. Scamurra, A. J. Ratner, and E. N. Janoff, “Antibody-enhanced pneumococcal adherence requires IgA1 protease,” vol. 100, no. 7, pp. 4215–4220, 2003.

[32] J. Carlsson, “A numerical taxonomic study of human oral streptococci.,” *Odontol. Revy*, vol. 19, no. 2, p. 137, 1968.

[33] C. Vadeboncoeur and M. Pelletier, “The phosphoenolpyruvate: sugar phosphotransferase system of oral streptococci and its role in the control of sugar metabolism,” *FEMS Microbiol. Rev.*, vol. 19, pp. 187–207, 1997.

[34] R. Iyer and A. Camilli, “Sucrose metabolism contributes to in vivo fitness of Streptococcus pneumoniae.,” *Mol. Microbiol.*, vol. 66, no. 1, pp. 1–13, Oct. 2007.

[35] A. Embry, E. Hinojosa, and C. J. Orihuela, “Regions of Diversity 8, 9 and 13 contribute to Streptococcus pneumoniae virulence.,” *BMC Microbiol.*, vol. 7, p. 80, Jan. 2007.

[36] Blakley et al., “Folates and pterins,” vol. 1, 1984.

[37] P. Burghout, A. Zomer, C. E. van der Gaast-de Jongh, E. M. Janssen-Megens, K.-J. Françoijs, H. G. Stunnenberg, and P. W. M. Hermans, “Streptococcus pneumoniae folate biosynthesis responds to environmental CO2 levels.,” *J. Bacteriol.*, vol. 195, no. 7, pp. 1573–82, Apr. 2013.

[38] A. L. Nelson, A. M. Roche, J. M. Gould, A. J. Ratner, J. N. Weiser, M. Clearance, and K. Chim, “Capsule Enhances Pneumococcal Colonization by Limiting Mucus-Mediated Clearance Capsule Enhances Pneumococcal Colonization by Limiting,” 2007.

[39] C. Hyams, E. Camberlein, J. M. Cohen, K. Bax, and J. S. Brown, “The Streptococcus pneumoniae capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms.,” *Infect. Immun.*, vol. 78, no. 2, pp. 704–15, Feb. 2010.

[40] J. Yuste, A. Sen, L. Truedsson, G. Jönsson, L.-S. Tay, C. Hyams, H. E. Baxendale, F. Goldblatt, M. Botto, and J. S. Brown, “Impaired opsonization with C3b and phagocytosis of Streptococcus pneumoniae in sera from subjects with defects in the classical complement pathway.,” *Infect. Immun.*, vol. 76, no. 8, pp. 3761–70, Aug. 2008.

[41] J. Yuste, M. Botto, J. C. Paton, D. W. Holden, and J. S. Brown, “Additive Inhibition of Complement Deposition by Pneumolysin and PspA Facilitates Streptococcus pneumoniae Septicemia,” *J. Immunol.*, vol. 175, no. 3, pp. 1813–1819, Jul. 2005.

[42] A. Bidossi, L. Mulas, F. Decorosi, L. Colomba, S. Ricci, G. Pozzi, J. Deutscher, C. Viti, and M. R. Oggioni, “A functional genomics approach to establish the complement of carbohydrate transporters in Streptococcus pneumoniae.,” *PLoS One*, vol. 7, no. 3, p. e33320, Jan. 2012.

[43] S. Roseman, “The Transport of Carbohydrates by a Bacterial Phosphotransferase System Components of the Phosphotransferase System,” *J. Gen. Physiol.*, vol. 54, no. 1, pp. 138–184, 1969.

[44] M. H. Saier and J. Reizer, “Proposed Uniform Nomenclature for the Proteins and Protein Domains of the Bacterial Phosphoenolpyruvate : Sugar Phosphotransferase System,” *J. Bacteriol.*, vol. 174, no. 5, 1992.

[45] W. Kundig, S. Ghosh, and S. Roseman, “Phosphate bound to histidine in a protein as an intermediate in a novel phospho-transferase system,” … *Sci. United States …*, pp. 1067–1074, 1964.

[46] M. H. Saier, “The bacterial phosphotransferase system: structure, function, regulation and evolution.,” *J. Mol. Microbiol. Biotechnol.*, vol. 3, no. 3, pp. 325–7, Jul. 2001.

[47] T. Lambert, “Antibiotics that affect the ribosome.,” *Rev. Sci. Tech. (International Off. …*, vol. 31, no. 1, pp. 57–64, 2012.

[48] S. L. Berg JM, Tymoczko JL, “Eukaryotic Protein Synthesis Differs from Prokaryotic Protein Synthesis Primarily in Translation Initiation.,” in *Biochemistry. 5th edition*, 2002, p. http://www.ncbi.nlm.nih.gov/books/NBK22531/.

[49] O. Humbert, M. Prudhomme, R. Hakenbeck, C. G. Dowson, and J. P. Claverys, “Homeologous recombination and mismatch repair during transformation in Streptococcus pneumoniae: saturation of the Hex mismatch repair system.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 20, pp. 9052–6, Sep. 1995.

[50] K. Poulsen, J. Reinholdt, C. Jespersgaard, T. A. Brown, M. Hauge, M. Kilian, K. Poulsen, J. Reinholdt, C. Jespersgaard, and K. I. T. Boye, “A Comprehensive Genetic Study of Streptococcal Immunoglobulin A1 Proteases : Evidence for Recombination within and between Species A Comprehensive Genetic Study of Streptococcal Immunoglobulin A1 Proteases : Evidence for Recombination within and between S,” *Infect. Immun.*, vol. 66, no. 1, pp. 181–190, 1998.

[51] C. G. Dowson, T. J. Coffey, and B. G. Spratt, “Origin and molecular epidemiology of penicillin-binding-protein-mediated resistance to beta-lactam antibiotics.,” *Trends Microbiol.*, vol. 2, pp. 361–366, 1994.

[52] R. Hakenbeck, T. Grebe, D. Zähner, and J. B. Stock, “beta-lactam resistance in Streptococcus pneumoniae: penicillin-binding proteins and non-penicillin-binding proteins.,” *Mol. Microbiol.*, vol. 33, no. 4, pp. 673–8, Aug. 1999.

[53] A. Fenoll, R. Muñoz, E. Garcia, and A. G. de la Campa, “Molecular basis of the optochin-sensitive phenotype of pneumococcus: characterization of the genes encoding the F0 complex of the Streptococcus pneumoniaeStreptococcus oralis H+-ATPases,” *Mol. Microbiol.*, vol. 12, no. 4, pp. 587–598, 1994.

[54] M. L. Ferrando, P. van Baarlen, G. Orrù, R. Piga, R. S. Bongers, M. Wels, A. De Greeff, H. E. Smith, and J. M. Wells, “Carbohydrate availability regulates virulence gene expression in Streptococcus suis.,” *PLoS One*, vol. 9, no. 3, p. e89334, Jan. 2014.

[55] Z. R. Lun, Q. P. Wang, X. G. Chen, A. X. Li, and X. Q. Zhu, “Streptococcus suis: an emerging zoonotic pathogen,” *Lancet Infectious Diseases*, vol. 7. pp. 201–209, 2007.

[56] H. F. L. Wertheim, H. D. T. Nghia, W. Taylor, and C. Schultsz, “Streptococcus suis: an emerging human pathogen.,” *Clin. Infect. Dis.*, vol. 48, pp. 617–625, 2009.

Anexos:

Foi usado o seguinte software:

- Enthought Canopy version 1.3.1 (64 bit)
    - python 2.7
    - NumPy
    - ScyPy
- Matlab versão 2010b
- OpenOffice

No CD anexado à dissertação estão os ficheiros com o código usado para o cálculo dos módulos, para a aplicação dos testes estatísticos e para a aplicação do método de "hierarchical clustering" e criação do dendrograma. O CD contém um ficheiro "readme.txt", com a descrição do tipo e função de cada ficheiro.