Universidade de Lisboa
Faculdade de Ciências
Departamento de Química e Bioquímica

Search for coherent gene modules that predict *Streptococcus pneumoniae* strain invasiveness

Rui Ribeiro Catarino

Mestrado em Bioquímica
Especialidade em Bioquímica
Dissertação de Mestrado orientada por
Doutor Francisco Pinto

2012

Agradecimentos

Quero agradecer antes de mais aos meus pais, já que, no final de contas, são os grandes responsáveis pelo percurso que segui. Fico também grato pelo sacrifício pessoal que fizeram ao suportar o meu ensino superior e qualquer ocasional má disposição. Espero poder agora começar a repagar esse esforço, começando por pagar um jantar assim que isto tudo esteja acabado.

Quero também agradecer ao grupo de enzimologia pelo ambiente de investigação saudável que me proporcionou. Em particular, um enorme agradecimento ao professor Francisco Pinto, não só por me ter iniciado numa área da qual eu pouco conhecia, pela atenção que sempre me cedeu sem restrições, mas acima de tudo pela amizade e excelente disposição a que me habituou e que, confio, se irá prolongar. O ano que se avizinha é a minha melhor oportunidade para devolver o trabalho que foi investido em mim, mas, num prazo mais curto, um jantar parece-me uma boa forma de começar.

Por fim, quero agradecer ao Daniel Fonseca que além de amigo e companheiro em toda duração (e em todas as vertentes) da vida universitária, sempre me estimulou intelectualmente em todas as áreas do saber que hoje prezo. Também a ele convidaria para jantar, não soubera que já não se irá encontrar em Portugal a essa data. Posto isto, deixo o convite para me convidares para jantar quando te for visitar.

Abstract

Streptococcus pneumonia is a pathogenic bacterium responsible for several human diseases, such as pneumonia, meningitis and sepsis. Any pneumococcal disease is preceded by an asymptomatic colonization stage in the human nasopharynx. The transition from colonization to invasion is known to depend on both human and pathogen factors. This work aims to computationally identify pneumococcal genetic factors that influence the likelihood of invasion events.

For this purpose, we analyze microarray based comparative genomic hybridization data of 72 strains of pneumococcus. Each strain was classified as Invasive, Neutral or Colonizer according to a previous study that compared the frequencies with which strains were recovered from an asymptomatic carrier or from invasive disease episodes.

We propose to select genes that, individually or in a coordinated way, affect the frequency of invasion transitions among all colonization events, which we denominate as invasiveness. To detect coordinated sets of genes, we developed a method that uses networks of known interactions between genes to find gene modules that predict invasiveness. Each module is founded with a single gene and then grown with its closest neighbors in the network. Each module is then evaluated for its predictive power, statistical significance and robustness to data variability.

We tested the method with a network based on a distance score that integrates gene co-occurrence and co-invasiveness. Among others functions, the found modules implicate cell envelope, transport, sugar metabolism, osmotic response, aminoacid synthesis, spermidine synthesis and proteolysis functions in pneumococcal invasiveness.

Resumo

O *Streptococcus pneumoniae*, também chamado pneumococcus, é uma bactéria gram-positiva do subgrupo alfa-hemolítico do género *Streptococcus*. É um colonizador frequente do trato respiratório superior humano e embora possa ser encontrado em qualquer pessoa, tem maior prevalência em crianças e idosos. A colonização decorre tipicamente sem causar sintomas, mas pode por vezes culminar na invasão de outros tecidos e provocar doenças como pneumonia, meningite ou otite do ouvido médio. Sem tratamento, a infeção com pneumococcus tem uma taxa de mortalidade da ordem dos 30 por cento mas, atualmente, com o uso de antibióticos e vacinas, este número é muito mais reduzido. Contudo, a resistência a antibióticos tem vindo a ser reconhecida em pneumococcus e a vacinação, mais do que reduzir o número de doenças provocadas por pneumococcus, tem conduzido à substituição das estirpes que as originam. Por estes motivos, torna-se urgente entender o mecanismo de invasão e virulência do pneumococcus para que novas formas de combate a este patógeneo possam tomar forma.

Como em muitos outros organismos que habitam meios de composição pouco variável, na maioria patógenos, o pneumococcus tem um genoma reduzido. O genoma apresenta grande plasticidade, variando cerca de 10 por cento entre estirpes e contem apenas 60 a 80 por cento de genes mantidos em todas as estirpes. A totalidade dos genes do pneumococcus, o pangenoma, é consideravelmente mais vasto que o genoma de qualquer estirpe e juntamente com a capacidade de trocar genes entre a própria espécie, ou por vezes com espécies próximas, confere a esta bactéria uma grande adaptabilidade e resposta rápida a mudanças no seu meio ambiente. A transferência de genes horizontal é de facto uma idiossincrasia do pneumococcus e é, por vezes, acompanhada pela indução de morte de células da mesma espécie para que estas libertem DNA. Este fenómeno, conhecido como fratricídio, acontece quando a célula entra num estado de competência, também chamado estado X. O segundo nome foi proposto por ser mais abrangente, evitando que o estado fosse apenas associado à competência. Neste estado, o perfil de transcrição da bactéria é globalmente alterado e além de expressar genes que promovem a competência, expressa também bacteriocinas tóxicas para as células vizinhas e proteínas que protegem a própria célula dessas bacteriocinas. A facilidade de incorporação de DNA

de outras células contribui significativamente para a sobrevivência da bactéria. A resistência à penicilina, por exemplo, é conferida por genes que foram adquiridos de uma espécie próxima, o *Streptococcus Mitis*.

A invasividade e virulência do pneumococcus varia de estirpe para estirpe e é função do conteúdo génico. A bactéria está especialmente adaptada para colonizar, visto passar a maior parte do tempo na nasofaringe e que o principal meio de transmissão ocorre por aerossol e quase exclusivamente durante a colonização. Embora não exista consenso sobre o motivo desta adaptação, é consensual que algumas estirpes são mais aptas para a invasão de outros tecidos e, consequentemente, causar doença. Entre os determinantes de virulência, o mais estudado é a cápsula polisacarídica ou serótipo. São conhecidos mais de 90 serótipos que diferem em estrutura e composição, mas apenas pouco mais de vinte estão associados a doença. A cápsula é um dos mais importantes mecanismos de defesa contra o sistema imunitário humano, já que, além de cobrir grande parte dos epítopos que seriam facilmente reconhecíveis, ainda inibe o sistema do complemento. Vários outros determinantes têm vindo a ser identificados mas o contexto genético tem sido descurado. Alguns dos genes associados com virulência numa estirpe, foram associados com colonização noutra, evidenciando a relevância das interações entre genes. A noção de que a invasividade pode ser conferida por interação entre genes complexifica tanto a busca de determinantes, como os próprios determinantes.

É possível identificar determinantes de invasividade procurando diferenças entre grupos de estirpes invasivas e grupos de estirpes colonizadoras. Estas diferenças podem ocorrer em diferentes níveis como o conteúdo génico ou a sua expressão. Dada a grande variabilidade do genoma do pneumococcus, é expectável encontrar determinantes de invasividade ao nível do conteúdo génico. Estas diferenças podem ser detetadas em larga escala por ensaios de microarrays de Hibridação Genómica Comparativa. É importante notar que esta abordagem é observacional e que, portanto, os resultados permitem apenas estabelecer correlações e não relações de causa efeito. Em contrapartida, permite observar múltiplas interações com diferentes backgrounds genéticos e a interação entre diferentes determinantes. Desta maneira, esta abordagem encaixa-se no paradigma da biologia de sistemas, visto estudar não só os genes individualmente, mas antes em interação com os demais.

A procura de determinantes que distingam estirpes invasivas de estirpes colonizadoras é um problema de classificação, uma área da aprendizagem supervisionada. Existem já muitos algoritmos desenhados para resolver este tipo de problema. Tipicamente, o sucesso destes algoritmos é avaliado pela sua capacidade de classificar corretamente as estirpes a partir dos seus genótipos. Entre outros, algoritmos como as redes neuronais são conhecidos por uma elevada exatidão de classificação. No entanto, o foco deste trabalho não é a exatidão de classificação mas antes a compreensão dos mecanismos que conduzem à invasividade. Grande parte dos algoritmos existentes resultam num conjunto de regras difíceis de interpretar e ainda mais de traduzir para um nível biológico, em especial se considerarmos que as estirpes invasivas podem ser um grupo heterogéneo com diferentes mecanismos de invasividade. Por este motivo, surgiu a necessidade de desenhar um novo algoritmo que foque primordialmente identificar determinantes de invasividade.

A procura de determinantes que tenham em conta a interação de genes constitui um problema computacional acrescido. A busca de múltiplos genes, módulos de genes, que constituam um determinante transforma-se num problema combinatorial em que o número de possibilidades aumenta exponencialmente com o número de genes. Para evitar uma busca exaustiva de todas as combinações, o algoritmo usa informação sobre interações entre os genes que podem ser de cariz metabólico, regulatório, físico, entre outros, mas que podem ser facilmente descritas num formato comum – as redes. As redes têm a vantagem de expressarem facilmente padrões de interações complexos e de serem manipuláveis e pesquisáveis computacionalmente.

Os dados usados neste trabalho resultam de um estudo de microarray de Hibridação Genómica Comparativa com 72 estirpes que usou como controlos as estirpes Tigr4, G54 e R6. Estas estirpes foram previamente classificadas como invasivas, neutras ou colonizadoras, de acordo com a frequência com que foram identificadas em indivíduos saudáveis ou em indivíduos portadores de doença. A presença ou ausência dos genes nas estirpes foi organizado numa matriz denominada matriz de presença génica. As estirpes neutras não foram incluídas na matriz por terem um cariz incerto. A classificação de uma estirpe como neutra pode dever-se tanto a motivos biológicos como à insuficiência de poder estatístico para a classificar como invasiva ou colonizadora.

Não foi usada uma rede de interações de genes mas sim uma matriz de distância que avalia a coocorrência e a coinvasidade. A coocorrência é um parâmetro que avalia a frequência com que dois genes estão presentes individualmente comparativamente com a frequência com que estão presentes em conjunto. A coinvasidade é um parâmetro que avalia a semelhança de associação de cada um dos genes com a invasividade. Esta associação é medida usando um teste estatístico de Fisher. Juntos, estes parâmetros asseguram que dois genes com uma baixa distância são genes que coocorrem frequentemente e que têm uma associação com a invasividade semelhante. A matriz de distâncias é usada para criar módulos de genes que serão depois avaliados. Os módulos são criados a partir de um gene semente, ao qual são gradualmente adicionados mais genes. O gene adicionado é sempre o gene com menor distância ao gene semente.

Os módulos de genes são inicialmente avaliados quanto à sua presença dos seus genes em estirpes invasivas e colonizadoras através de um teste de *runs*. Este teste avalia se a distribuição das presenças pelas classes de estirpes é significativa ou se pode ser considerada aleatória, caso em que o módulo é abandonado. De seguida é definido um número de genes, limite, acima do qual o módulo é considerado presente numa estirpe. Este limite é definido de forma a que o módulo esteja presente exclusivamente em estirpes invasivas. Se tal limite não existir o módulo é abandonado. Caso tenha sido possível estabelecer um limite, é avaliada a significância do mesmo. Para tal é usado um teste unilateral que calcula a probabilidade do limite ter sido fixado com um valor tão ou mais baixo. Caso o limite não tenha significância estatística de 0.05 o módulo é abandonado.

Dado o método de formação dos módulos, é possível que nem todos os genes contribuam para a associação do módulo com a invasividade. Para eliminar essas situações é avaliada a associação individual de cada gene com as estirpes em que o módulo está presente usando um teste de Fisher. Os genes que não estiverem associados são eliminados do módulo. Após a remoção de genes o limite é recalculado e a sua significância é reavaliada. Terminado este passo, é selecionado apenas um módulo de entre os módulos criados a partir do mesmo gene semente. O módulo selecionado é aquele que for constituído pelo maior número de genes. Por fim realizou-se uma correção para testes múltiplos que estabeleceu a taxa de descobertas falsas em 5 por cento. Este passo eliminou todos os módulos com menos de 24 genes.

De todo este processo resultaram 26 módulos significantes pelos padrões estatísticos exigidos e que estão presentes exclusivamente nas estirpes invasivas. Embora os módulos sejam distintos, existe grande sobreposição entre eles. É possível observar submódulos que surgem repetidos em vários módulos e que eram possivelmente módulos por si, tendo sido eliminados pela correção por testes múltiplos.

Para cada módulo, observou-se que a presença dos seus genes está correlacionada com o rácio de probabilidade da invasividade das estirpes. Esta correlação observa-se mesmo para as estirpes neutras, ainda que estas não tenham sido usadas como input no algoritmo. Embora as classes invasiva e colonizadora tenham sido usadas pelo algoritmo, os dados dos seus rácios de probabilidade de invasividade não foram. Em conjunto, os módulos usam um total de 111 genes e, usados em conjunto, é possível encontrar uma correlação semelhante. A correlação dos módulos, individualmente e em conjunto, com os rácios de probabilidade de invasividade e com as estirpes neutras é um resultado positivo que suporta a relevância e autenticidade destes módulos como determinantes de invasividade.

Os módulos são robustos contra pequenas alterações na matriz de presença de genes. A experiência de microarray a partir da qual os dados foram originados tem um erro inerente e esta alta robustez confere confiança na autenticidade dos resultados do algoritmo, mostrando que dificilmente são consequência de erros do microarray. A existência de um limite para definir presença de módulos, por oposição à exigência de presença de todos os genes em simultâneo, pode ser uma fonte de robustez contra perturbações nos perfis de presença dos genes.

Não foi encontrado enriquecimento de funções entre os genes selecionados pelo algoritmo nem entre os módulos. O enriquecimento das funções foi avaliado usando a anotação do JCVI. Apesar de não se ter verificado enriquecimento funcional usando a anotação da base de dados do JCVI, alguns genes têm claramente relações funcionais. O nrdD codifica um ribozima que é ativado pelo nrdG. Os genes Argh e ArgG codificam enzimas que catalisam reações sequenciais que constituem uma via alternativa da síntese da arginina. O enzima manitol-1-fosfato desidrogenase (mTLD) utiliza como substrato o manitol-1-fosfato, que é o produto do transporte de manitol pelo sistema PTS (MTLA e mtlF). O RuvB tem a sua atividade como estimulador de recombinação facilitada pela presença da proteína de ligação de DNA de cadeia simples ssb. Um transportador ABC requer a presença de vários componentes que

foram selecionados pelo algoritmo, tais como módulos de ligação ao ATP (ou NBDs) e permeases transmembrananares. A ação da aquaporina Z (aqpZ) tem levantado dúvidas na comunidade científica, já que a sua ação parece conduzir ao acumular de pressão de turgescência celular excessiva. O canal mecanosensível largo (MsCl) proporciona uma resposta eficaz para a pressão de turgescência e pode ser um contrapartida biológica da aqpZ. Poliaminas, como a espermidina e norespermidina, têm sido relatadas como possíveis substitutos da colina e são, por conseguinte, intervenientes importantes na estrutura da parede celular e possivelmente na ligação a proteínas que se ligam a colina.

A maioria dos genes selecionados foi previamente associada com a invasão ou tem alguma conexão plausível com os mecanismos de invasão. Proteínas da cápsula e proteínas que ligam colina desempenham um papel importante na proteção contra as defesas do hospedeiro. São importantes na inibição da ação do sistema imunitário, nomeadamente pela remoção das proteínas do complemento, ou pela ligação ao fator H, que é um inibidor do complemento. Vários elementos genéticos móveis foram identificados dentro ou perto do locus dos genes da cápsula e tem sido relatado o impacto destes elementos na regulação da transcrição de vários genes desse locus.

A invasão de novos tecidos requer uma adaptação rápida a um ambiente novo, tanto às suas propriedades físicas como à disponibilidade de nutrientes. Foram selecionados genes de resposta a mudanças da pressão osmótica que parecem mais dirigidos a uma resposta rápida a grandes alterações da pressão do que à regulação fina da pressão e são, portanto, de particular interesse na adaptação a novos meios. Genes de resposta anaeróbica como o nrdD e o seu ativador, nrdG, dificilmente são funcionais na nasofaringe, uma vez que são estritamente anaeróbicos. No interior do organismo humano contudo, a concentração de oxigénio é reduzida, uma vez que este está quase sempre ligado a moléculas biológicas como a hemoglobina. Nestas circunstâncias o nrdD pode ser crucial para manter as funções dos enzimas aeróbios equivalentes. A capacidade de utilizar diferentes fontes de energia e de carbono é de extrema importância para a invasividade de uma estirpe. O elevado número de transportadores de açúcar está relacionado com a capacidade das estirpes invasivas sobreviverem em meios de variadas composições. Na mesma lógica, alguns genes foram selecionados que codificam para enzimas do metabolismo de diferentes açúcares, aumentando

também a adaptabilidade da estirpe a diferentes meios. Genes de proteólise estão provavelmente relacionadas com as necessidades nutricionais de aminoácidos.

A síntese de proteínas é um processo constante em todas as bactérias e exige uma disponibilidade permanente de aminoácidos e tRNA. Foram selecionados genes de síntese de aminoácidos que proporcionam vias alternativas para a síntese de aminoácidos, utilizando substratos alternativos. O algoritmo também selecionou genes ligados à síntese e ligação de tRNA ao aminoácido correspondente. Estes enzimas não foram caracterizados em *Streptococcus pneumoniae* e é difícil prever a sua influência na síntese proteica.

Por fim, a grande heterogeneidade dos genomas do pneumococcus advém da sua capacidade de recombinação. Alguns dos genes selecionados pelo algoritmo promovem a heterogeneidade do genoma, aumentando a recombinação com o DNA extracelular. Entre os genes selecionados é promovida a internalização de DNA, a sua estabilização e a recombinação com DNA não homólogo. O estado de competência do pneumococcus é acompanhado por uma apetência para induzir a apoptose em células vizinhas, aumentando a concentração de fragmentos de DNA no meio. Várias bacteriocinas foram associadas por este trabalho à invasividade, bem como genes que inibem a apoptose da própria célula. Esses genes dão à célula uma vantagem natural na competição com outros colonizadores.

Em suma, alcançou-se o objetivo pretendido de encontrar determinantes de invasividade. Estes determinantes são fruto de um estudo observacional e é portanto de notar que a relação que têm com a invasividade é apenas de correlação. Para determinar o impacto que estes módulos de genes têm na invasividade é necessário realizar estudos laboratoriais que averiguem em maior detalhe a função biológica dos genes e a sua relação com os mecanismos de invasão.

Index

Figure index

Table index

Streptococcus pneumoniae

Streptococcus pneumoniae, also known as pneumococcus, is a human commensal gram-positive bacterium. Although commonly found in the human upper respiratory tract in asymptomatic carriers, pneumococcus can spread to other organs and cause diseases such as pneumonia, meningitis or otitis media[1,2]. Attempting to reduce its impact on human life, scientists have studied pneumococcus for more than a century and this effort resulted in ground-breaking discoveries in major areas of life sciences [3,4].

Pneumococcus was first isolated in 1881 by Sternberg in the United States and on the same year by Pasteur in France [3-5]. Both classified it as a diplococcus. It would later come to note that the diplococcus form was a consequence of the liquid medium where it was isolated. The availability of new biochemical and molecular techniques for taxonomic identification led to the classification of this bacterium within the genus *Streptococcus*, as *Streptococcus pneumoniae*.

The association with pneumonia was not immediate. It was only in 1896 that Weichselbaum would come to settle the argument between Friedlander's and Albert Fraenkel's laboratories over pneumonia etiology. This argument is of special importance as it is the first time Christian Gram's stain was used. Interestingly, it was created not to distinguish bacteria, as it is used today, but to facilitate the visualization of pneumococcal bacteria in histologic sections of the lung [3,5].

In 1904, Sir William Osler said about pneumococcus: "In the Mortality Bills, pneumonia is an easy second, to tuberculosis; indeed in many cities the death-rate is now higher and it has become, to use the phrase of Bunyan 'the captain of the men of death." [3,4,6]. In that time, pneumococcal infection had a case fatality rate of 30 to 35% and there was no successful therapy [3,7]. However, several discoveries would be made in the same decade resulting in serotherapy, reducing the case fatality rate to ~20% [3]. Neufeld and Haendel identified pneumococcal types 1 and 2. Klemperers recognized the protective value of antiserum against infections with homologous organism and Neufeld discovered the lytic effect of bile on pneumococci, facilitating diagnosis. Serotherapy

specificity to types of pneumococcus (serotypes) pushed the identification (typing) of pneumococcus [3,4].

In 1925, another contribution to immunology resulted from the study of pneumococcus. Dochez and Avery discovered that the capsular composition, which is the main immunogenic substance of pneumococcus, was not proteinaceous in nature but in fact a polyschariде, becoming the first non-proteic antigen identified [3,4].

The identification of DNA as the genetic vehicle is pointed by many as the single greatest impact in biology to come from the study of bacteria [4]. It starts in the early 1920s, when Griffith took advantage of the morphological differences between encapsulated (smooth) and unencapsulated (rough) pneumococcal strains and their discrepancies in virulence proficiency to discover the "transforming principle". His experience consisted in injecting a mix of live rough (non-virulent) and dead smooth (virulent) pneumococcus in healthy mice, resulting in the death of the animals and the recovery of live smooth bacteria expressing the same capsular serotype as the dead strains [3,4].

Avery, McLeod and McCarty, recreated and extended this experiment in vitro and in 1944 identified the "transforming principle" as DNA. They took advantage of other pneumococcal property, its natural competent state to incorporate smooth DNA into rough strains. Previously unencapsulated strains acquired a capsule, becoming, in McCarty's words, "sugar-coated bacteria" [3,4]. Unfortunately their work was received with scepticism and wasn't even quoted when Watson and Crick published their seminal work, 9 years later [4].

Also in 1944, Tillet et. al published their paper on penicillin and its incredible efficacy in recovering patients with pneumococcal pneumonia (reducing case fatality rate to 5 to 8%) [8]. The disease once feared was now regarded as casual and typing sera, which was a byproduct of therapeutic sera ceased [3,4]. The second half of the century was marked by a shift in the scientific community opinion towards microbiology: one Nobel laureate asked, "Who cares anymore [about bacteria]?" and the US Surgeon General stated, "The war against infectious diseases has been won" [4]. Not even reports in the 1960s of penicillin-acquired resistance changed the view that pneumococcus was an overthrown bacteria and it was nearly abandoned from scientific research [3,4].

Today, the generalized use of penicillin and subsequent antibiotics resulted in growing reports of resistant strains and even multi resistant strains [9,10]. From all the antibiotics available, only vancomicine remains to select resistant mutations [3,4]. Moreover, antibiotics unspecificity to serotypes and even to organism disturbs the ecological equilibrium of the microflora. The resident flora inhibits colonization by *S. pneumoniae*, *H. influenzae*, *S. aureus*, and *M. catarrhalis* [1]. Also, *S. pneumoniae* itself can interfere with the growth of *S. aureus* and this effect has been attributed to pneumococcal hydrogen peroxide [1]. Nevertheless, antibiotics importance is not to be underappreciated and will surely continue to be the major therapy form against pneumococcus. New drugs will need to be designed continuously as pneumococcal start to be resistant and therefore this approach doesn't seem to bring forward a definitive strike on pneumococcus influence in human health.

An alternative approach to fight pneumococcal infection is through widespread immunization with vaccines [2]. Vaccines against pneumococcus exist since 1940s but, at the time, the newly discovered penicillin overshadowed them and forced the withdrawal from the market [3]. Austrian and Gold denounced the decreasing effectiveness of case management procedures and started to redesign the polyvalent polysaccharide pneumococcal capsule vaccines [11]. Data from trials with vaccines have showed decrease in infections with the targeted serotypes and carriage itself decreased [2]. In particular, infections in younger hosts have diminished. Vaccination has a clear impact on transmission dynamics as the colonization success rate decreases in half within hosts immune to the particular serotype [3]. Nevertheless, recent data shows that the selective pressure against some serotypes has allowed for others to strive resulting in a phenomenon called Serotype Replacement [12]. Infection diseases caused by typically less frequent strains evidence that widespread immunization isn't leading to the predicted results.

Pneumococcal genome dynamics

As many organisms living in controlled environments, usually pathogens, pneumococcus has a reduced genome [13,14]. Still, much plasticity exists within the

pneumococcus genome, with up to 10% variation between strains, and up to 5% of repeated sequences[13]. An average genome has around 2 to 2.2 Mbp (a typical strain as TIGR4 has 2160837 bp) [15], and an average of 60 to 80% of each sequence is conserved by all strains [16]. In Donatti et al, the size of the total S. pneumoniae gene pool accessible to the species, or pan-genome, was calculated using two different static descriptive methodologies, namely the finite supragenome model and the power law regression model [13]. While the first estimated the pan-genome to have 3000 to 5000 genes, the second estimated that the pan-genome is open, having a potentially infinite number of genes. It is conceivable that an open pan-genome, together with a mechanism to spread genes through unrelated strains, guarantees a quick and economical response to fluctuating environments [17,18]. In fact, Horizontal Gene Transference (HGT) is an idiosyncrasy in pneumococcus and it is sometimes preceded by fratricide, a phenomena where pneumococcus cells induce death on other pneumococcal cells [19]. The dying cells release their DNA content, which may then be acquired by the killer cells. Competence in pneumococcus is induced in response to detection by a two-component signalling system of a peptide pheromone (Csp) secreted by the bacterium [20]. In microarray analyses, a large number of genes whose expression is altered by CSP signalling were detected, but among 124 genes, only 23 are clearly necessary for transformation [21,22]. This complex shift in protein expression is not fully understood but plays a major role in pneumococcal adaptability and response to antibiotics [20]. For example, genetic variation resulting in resistance to penicillin is due to acquisition of fragments of the genes encoding penicillin-binding proteins from *S Mitis*, reducing the affinity of these proteins for the drug [9,23].

Pneumococcal invasiveness determinants

Being a commensal bacterium, pneumococcus is highly adapted to human nasopharynx environment and transmission between hosts occurs essentially during carriage, by aerosol droplets[14,24]. It is therefore puzzling why pneumococcal is so proficient in causing infection. Some theories point that the main traits favouring virulence were acquired primarily to improve colonisation

and that infection is a by-product of effective colonisation[14,24]. Supporting this theory is the finding of many necessary virulence factors in the core genome [13]. However, tissue specific virulence factors have also been reported, suggesting a real adaptation to invasiveness [14]. One explanation regards infection as a means to efficiently eliminate competitor bacteria and other portraits infection as stimuli of cough and mucus secretion as a mean to improve transmission [24]. While this remains an open question, it is undeniable that different pneumococcal strains show different aptitude for invasive behaviour. An inverse correlation between time of carriage and infection rate has been found and in most cases infection occurs shortly after colonization evidencing that some strains are especially virulent.

Search for invasiveness determinants or virulence factors is an widespread goal. Chief among virulence factor is the polysaccharide capsule or serotype [12,25]. There are at least 93 serotypes differing in structure and composition but infection is typically caused by only ~20 [26]. The capsule is responsible for resisting complement-mediated opsonophagocytosis and providing camouflage to the highly immunogenic epitopes in cell [27]. Capsular polysaccharide is highly negatively charged and sterically inhibits the interaction between phagocytic complement proteins with receptors fixed to pneumococci [6,7].

Beside the capsule many other virulence factors have been identified. Special attention should be given to Signature-tagged mutagenesis (STM) studies, negative genetic screens that allow a complex mixture of mutagenized strains to be screened simultaneously in a host for attenuated mutants [28-30]. Comparison of the three STM studies determined that the majority of loci identified were hit by only one study [14]. Although this may reflect differences in methodology, it is likely that the use of three different strains of pneumococcus—G54 (serotype 19F); strain 0100993 (serotype 3); and T4 (serotype 4) —contributed to the discrepancy observed between the three STM studies. This disparity suggests that strain-dependent variations may influence single gene knock out impact on virulence. On a different study, an island of genes (RD5) was linked to invasive phenotype in serotype 6A, but also with non-invasive phenotype in serotype 14 [26]. It has been proposed that virulence factors may complement the capsular properties or that an array of virulence factors needs to be expressed in a

coordinated way for tissue invasion to be successful [26,31]. The notion that virulence becomes from interaction between genes instead of single genes alone brings the search of invasiveness determinants to a new complexity level.

Quantification of virulence or its related measure, invasiveness, and identification of genetic factors that determine these properties has been broadly approached through have used mutagenesis and knock outs studies followed by in vivo tests in animal models to. Alternatively, these properties can be studied using epidemiological data where pneumococcal strains have been recovered from either disease carrying hosts or asymptomatic hosts. These strains can be grouped together in genetic homogenous groups using different types of molecular typing [31]. Each group can be associated with a higher or lower tendency to cause invasive disease – invasiveness. This property should not be confused with virulence, which quantifies the severity of the disease.

Identifying invasiveness determinants, which are potential virulence factors, consists in finding differences between invasive and non-invasive strains [32,33]. Differences may occur at gene content or gene expression. Since pneumococcal genome is highly variable and dynamic it is expectable that differences able to discriminate between both classes of strains may be found at the gene content level. Differences may be found performing PCR studies or, in a larger scale, with microarray based Comparative Genome Hybridization or with full genome sequencing [34].

Unlike the STM and other in vivo testing experiences, this approach is observational and, therefore, can only detect correlations or associations but not cause-effect relations. It has, however, the advantage of identifying several possible invasiveness determinants simultaneously observing them in multiple genetic backgrounds. Also, it is possible to study interaction between multiple invasiveness determinants and try to infer possible interactions between them. It is also possible to identify new determinants that become from the functional interaction between them and not from the action of the individual genes. This holistic approach and attention to genes interaction fits in a systems biology mold [35]. Systems biology is growing area of life sciences that contrasts with single molecule studies, promoting the study of wider systems [36]. A key feature

of systems biology is the study of properties that emerge from the interaction of elements, which would not be clear studying the elements in separate [37]. The famous sentence sums up the idea of systems biology: "the whole is bigger than the sum of its parts".

Computational methods

With information over genetic content of several strains and over the tendency of these strains to cause invasive diseases, supervised learning algorithms could be used to infer classification rules [38-40]. These algorithms are designed to solve classification problems, such as genotype-phenotype associations, and are evaluated by accuracy of its predictions. Algorithms such as decision trees, neural networks and support vector machines are widely used and with extensive success [41]. However, to the goal of biomolecular comprehension of the mechanisms of invasion, it is more important the interpretation of the rules of classification that the classification accuracy itself. Neural networks in particular have been nicknamed as "black boxes" for it is particularly difficult to interpret their classification rules [42]. Recognition that invasive strains class might not be homogenous but rather achieve invasiveness through different mechanisms further thwarts the problem, as these mechanisms would probably be bundled in a complex function. In this work we attempt to create an alternative to these algorithms that efficiently unveils the mechanisms that lead to invasiveness.

Search of gene modules as invasiveness determinants instead of individual genes represents a further increase of the computational difficulty [43]. The number of possible modules increases quickly with the size of the module. With 1000 individual genes, there are 1000 possible simple determinants, 499500 possible determinants with two genes and the number explodes in combinatory way as we keep searching for modules with more genes. To avoid an exhaustive search in such a huge number of possibilities, it is possible to use a heuristic approach, resorting to previously known interactions between genes [32,33,44]. Interactions may have any biological function affiliation (metabolic, physic, regulation) but are always easily codified in a universal format: networks [45]. Networks are capable of expression complex interaction patterns and can be computationally

handled and researched. To this specific problem, we propose that gene modules associated with invasiveness has some functional interaction between its genes.

Objectives

In this study we propose to identify invasiveness determinants. We define determinant as a group of one or more genes that, together, correlate with invasiveness.

We also propose to design a new algorithm that focus on unveiling rules of classification. Classification rules must be able to discriminate invasive class from non-invasive but are not required to predict all invasive strains. In the end, classification rules should clearly translate into meaningful biological functions.

Methodology

Methodology overview

The aim of this work is to devise an algorithm to find gene modules associated with pneumococcal invasiveness extracted from biological networks. The input information is the pattern of gene presence/absence in a set of strains that contains strains associated with both invasive disease and colonization behaviours. After a brief filtering of non-informative genes, a gene network is created from known biological information. In the present work, this network was based on the CGH presence data and on the invasiveness association of individual genes. Next, gene modules are selected from the network and are evaluated according to their combined association with subsets of invasive strains. In the end, false discovery rate is assessed comparing these modules with modules yielded from strains with random gene composition.

Gene module definition

The core of the methodology proposed in this thesis is the definition of gene module. The gene module is constituted by a set of genes that fulfil the following requirements:

1 - are close neighbours according to a network of gene-gene interactions or to a matrix of pairwise distances between genes;

2 – when a strain contains more than t module genes in its genome, it has a high probability of being associated with invasive disease (or with colonization).

Given the information about gene presence/absence in strains known to be associated with invasive disease or colonization, it is possible to establish the optimal threshold t that maximizes the probability that strains containing the module are in fact associated with invasive disease (or with colonization). The proposed methodology evaluates if the value of t and the achieved predictive probability are statistically significant and robust to noise in the input data.

Noteworthy, the gene modules are not penalized by the number of strains associated with invasive disease that are not detected by the modules presence. The present methodology does not aim to uncover modules that justify the behaviour of all invasive strains. Being a complex phenotype, there may be several different cellular functions that, independently or synergistically, contribute to strain invasiveness.

Methodology description

Gene presence and strain classification data

This work is based on a data set obtained in a microarray based Comparative Genomic Hybridization (aCGH) experiment of 72 strains of pneumococcus. The microarray platform represented genes present in the sequenced genomes of strains Tigr4, G54 and R6. R6 is an avirulent strain descendent of the serotype 2 fully sequenced in 2001 [46]. It derives from D39, the strain used by Avery and co-workers to prove that DNA is the genetic material. Used in laboratories ever since, the strain adapted and lost virulence traits including the capsule. R6 genes are targeted in 2839 spots for each strain. Tigr4 is a strain from the serotype 4 sequenced in 2002 and is highly invasive and virulent in mouse models [15]. In the microarray, 3015 spots target Tigr4 genes. G54 genome was drafted in 2001 [47] and later fully sequenced in JCVI. In the microarray 2763 of its genes were primary targets.

The test group consisted of 72 pneumococcal strains. The association of these strains with invasiveness has been estimated in a previous study [48]. There, two collections of pneumococcal isolates were obtained in Portugal between 2001 and 2003. One of the collections contained carriage isolates and the other disease isolates. The study calculated the association of each group of genetically similar strains with invasive disease or colonization based on the number of isolates in either collection (Supplemental data). As a result, each group of genetically similar strains was classified as invasive, neutral or colonizer. Each of the 72 strains analysed by aCGH is a representative of a group of genetically

similar strains studied, and therefore, inherits its classification as invasive, neutral or colonizer.

After microarray data analysis, each gene was represented as vector v with a length of 72. The i^{th} element of v, v_i, will have the value 1 if the gene was detected in the i^{th} strain, and 0 if it was absent. This vector is here referred to as gene presence vector and the collection of the genes presence profiles is compiled in a matrix g. The matrix element g_{ij} indicates if the i^{th} gene is present (g_{ij}=1) or absent (g_{ij}=0) in the j^{th} strain. The matrix is here referred to as gene presence matrix.

Gene and Strain selection

Due the uncertainty nature of the neutral class, strains with this classification were not used in the search for invasiveness modules. Neutral classification can be a consequence of two different scenarios, one of statistical nature and one with a biological subtext. Strains that appear with similar frequency in invasive disease and colonization collections need a higher statistical power to be significantly associated with either invasive of colonizer behaviours. These strain can share phenotypic traits with invasive or colonizer strains but the neutral class is assigned to them due to statistical power insufficiency. Contrasting with the previous scenario, neutral classification can, in fact, be appropriated to some strains. Neutral behaviour can be a consequence of the accessory genome composition, with a mix of genes promoting invasion with genes promoting colonization, or, the absence of genes promoting either behaviour. In other words, invasiveness may be a continuous property ranging from coloniser to invasive strains and, halfway the scale, neutral strains. We chose to use only invasive and non-invasive strains to train the gene module finding algorithm. The resulting modules can be searched for in the neutral strains, possibly discriminating the reason why the strain was previously classified as neutral.

Not all the genes in the matrix were used in the algorithm. Some of the probes used in the microarray were not specific to only one gene. To avoid a complex

and dubious analysis of the results, all the probes that could hybridize with more than one gene per genome were not used. This accounted for 252 genes. Genes present in all strains (1693 genes) or absent in all strains (4 genes) were also removed. These genes are of no value in the analysis, as they don't allow discrimination between strains and are but a computational weight. Additionally, genes with repeated gene presence vectors (376 genes) were clumped together. The resulting matrix has 47 strains and 1295 genes. The remaining genes have different presence profiles with both low and high presence frequency, which contrasts with the initial distribution (Figure 1 and 2).

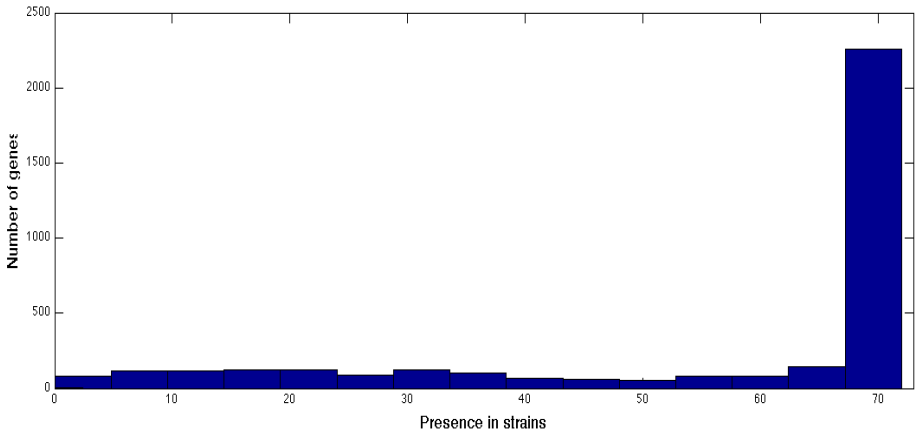

Figure 1 - Gene distribution before gene and strain selection.

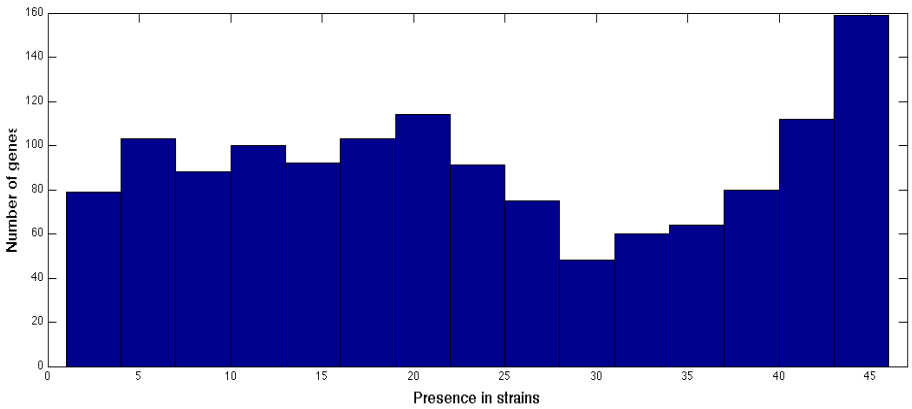

Figure 2 - Gene Distribution after gene and strain selection.

Construction of a Gene-gene distance matrix

The algorithm is devised to use networks like metabolic, transcription regulation or protein-protein interaction networks. In the absence of curated networks of these types for *Streptococcus pneumoniae*, this work is based on a distance matrix computed from the presence profile matrix and the association of each individual gene with the invasive class. The distance is evaluated through two components: co-occurrence and co-invasiveness. Co-occurrence is assessed using the Jaccard distance. This distance evaluates the frequency of co-occurrence between genes in all the strains where at least one of the genes is present. Jaccard avoids low distance scores between two low frequency genes that don't co-occur, as would happen with the Euclidean or the Hamming distance. Table 1 shows two gene presence vectors and the selection of strains to calculate the Jaccard distance. In equation 1, where V_i represents the set of strains where gene *i* is present, the Jaccard distance J_{ij} is calculated as the ratio between the number of the genes present in only one of the strains and the sum of all genes present in at least one of the strains. Using the examples in table 1, the Jaccard distance is $J_{ij}=3/4=0.75$.

Table 1 – Gene Presence vectors with highlights on the genes used to calculate the Jaccard Distance

	Strain 1	Strain 2	Strain 3	Strain 4	Strains 5	Strain 6
Gene *i*	1	0	1	1	0	0
Gene *j*	0	0	1	0	1	0

$$J_{ij} = \frac{\left|V_i \cup V_j\right| - \left|V_i \cap V_j\right|}{\left|V_i \cup V_j\right|} \qquad \text{Equation 1}$$

Co-invasiveness is also a distance and it evaluates the difference between two genes association to the invasive class, measured by Fisher's exact test. Fisher's exact test is a test of statistical significance applied to contingency tables (Table 2). In short, right tailed Fisher's exact test calculates the probability that an association at least or more extreme than the one observed exists in a

contingency table assuming the null hypothesis (H_0) that row and column variables are independent and maintaining the marginal totals (equation 2). The p-value is used as the strength of each gene association to invasiveness. The difference between each gene p-value is the co-invasiveness distance, I_{ij} between them (equation 3).

Table 2 - Contingency table

	Invasive strains	Coloniser strains	
Gene *i* Present	a=18	b=3	a+b=21
Gene *i* Absent	c=13	d=13	c+d=26
	a+c=31	b+d=16	47

$$p_i = P(a > a_{obs} | H_0) \quad \text{Equation 2}$$

$$I_{ij} = |p_i - p_j| \quad \text{Equation 3}$$

2x2 Contingency tables have only one degree of freedom because they maintain the marginal totals constant. In this example, there are only 3 possible tables were the positive association would be more extreme (higher counts in the left upper and right bottom cells). Right tailed Fisher's exact test calculates the p-value based on the probability of these three tables under the hypothesis of independence between gene *i* presence and strain class. In this case the p-value is 0.011.

As co-occurrence and co-invasiveness distances may have different amplitudes, they are converted to their respective rank (in the ordered list of all pair-wise gene distances). The distance between each pair of genes is the maximum of the two ranked distances (equation 4). The choice of the maximum guarantees that genes pairs with a small final distance have simultaneously small co-occurrence and co-invasiveness distances.

$$D_{ij} = \max(rank(J_{ij}), rank(I_{ij}))$$ Equation 4

Selection of gene modules

An invasive gene module is defined as a group of genes *m* (that are related according to a given network) with a determined presence threshold *t*, such that strains that have in their genome more than *t* genes belonging to the module tend to be invasive strains. When a strain has more that *t* genes of the module, the strain is said to have the module present. A similar definition can be stated for colonizer gene modules, but for simplicity, in the remaining methodology section only invasive gene modules are referred.

Each gene in the distance matrix is used as a seed for module selection. Each seed is a module itself that grows in steps by adding the next closest neighbour of the seed. Each module differs from the previous module in exactly one gene. If two or more genes have the same distance to the seed, the gene added is the one with the smaller average distance to the genes already included in the module. Module creation is stopped with a size of 50 to reduce computational weight and avoid the evaluation of large modules with reduced statistical confidence. As shown in Figure 3, module size is inversely correlated with its presence on strains. Some modules with 50 genes are still fully present in several strains. By fully present it is meant that all the genes of the module are present in a particular strain. If a large module is present in very few strains, it is difficult to test if those strains tend to be invasive or colonizers due to a loss of statistical power.

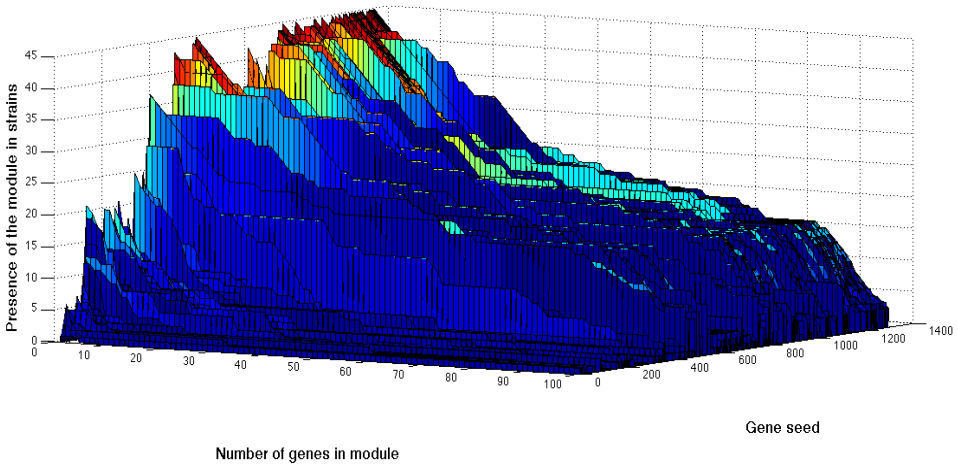

Figure 3 - Influence of module length on module presence in strains

Gene modules are evaluated through three successive filtering steps. The first is a runs test, which tests for the random distribution of strain classes (invasive or colonizer). Strains are organized as an ordered list based on the number of module genes present. Each strain is labelled as invasive or colonizer. The runs test statistic *R* is the number of times the strain label changes when the list read from start to end. Runs test has the null hypothesis that labels are randomly distributed, where *R* has intermediate values. In the left tailed version applied in this work, the alternative hypothesis is that labels are clustered together, which produces low *R* values. The probability of observing *R* runs is obtained according to the combinatorial expressions in equations 5 (when *R* is even) and 6 (when *R* is odd), were *m* is the number of invasive labels and n is the number of colonizer labels. The parenthesis notation in equations 5 and 6, with an upper number *u* and a lower number *l*, refer to the number of possible combinations of the *u* elements in subgroups with *l* elements. Statistical significance is obtained through equation 7, and is required to be equal or lower than 0.05 for each module to be maintained.

$$P(R=2x)=\frac{2\binom{m-1}{x-1}\binom{n-1}{x-1}}{\binom{n+m}{n}} \quad \text{Equation 6}$$

$$P(R=2x+1)=\frac{\binom{m-1}{x-1}\binom{n-1}{x}+\binom{m-1}{x}\binom{n-1}{x-1}}{\binom{n+m}{n}} \quad \text{Equation 7}$$

$$P(R\le R_{obs})=\sum_{r=2}^{r=R_{obs}}P(R=r) \quad \text{Equation 8}$$

The runs test is applied to speed the module search, since it has a low execution time, and evaluates the modules prior to threshold definition. In practice, it lowers the number of modules that have to be evaluated in the remaining filtering steps that are more time consuming. The next steps are still necessary because all the desired modules should present a significant runs test, but some modules that survive the runs test can be discarded later. This happens when invasive strains tend to have a similar number of module genes present, which causes the clustering of the invasive labels in the runs test list, but that clustering happens in the middle of the list or in the extreme that corresponds to the presence of a small number of module genes in the strain's genome.

The second filtering step estimates the Invasive Predictive Value (IPV) of each module. IPV is a measure that assesses the probability with which the module correctly identifies invasive strains. In other words, among the strains that have the module present (that is, have *t* or more module genes present), IPV gives the fraction of those strains that are invasive. IPV=1 means that the module is only present in invasive strains. Module thresholds *t* are fixed on the lowest value that allows modules to have an IPV=1. In other words, thresholds are the largest number of module genes present in a non-invasive strain. If no threshold can be found that gives the module an IPV=1 the module discarded. The requirement of IPV=1 was set after the observation that our gene presence matrix and the associated gene distance matrix could easily generate modules with this maximal IPV. For other applications this requirement can be relaxed, but still a high IPV should be chosen to assure the module predictive power. The selection of modules through the IPV measure does not require that an invasive module

should be present in the majority of invasive strains, only that it is present mainly in invasive strains. This allows the detection of modules that may contribute to the invasiveness of some strains, but not others, which is coherent with the biological perspective that invasiveness is a complex phenotype that can be influenced by different cellular pathways.

In the third step the significance of threshold *t* is evaluated. For this purpose a one tailed test was developed that calculates the probability of the threshold to be as the one observed or lower. This test is based in the classical theory of extreme value statistics [49]. First, a list with the number of module genes present in every strain is obtained, from which the frequency of each number of presences is calculated. This frequency f(*n*) represents the probability of a strain having *n* genes from the module *m* (equation 9). The corresponding cumulative frequency F(*n*) is the probability of a strain having at the most *n* genes (equation 10).

$$f(n) = P(\sum_i g_{ij} = n : i \in m) \qquad \text{Equation 9}$$

$$F(n) = \sum_{m=1}^{n} f(m) = P(\sum_i g_{ij} \leq n : i \in m) \qquad \text{Equation 10}$$

$$F_S(n) = F(n)^S \qquad \text{Equation 11}$$

The probability $F_S(n)$ of *S* independent strains having each at the most *n* genes is the cumulative frequency F(*n*) powered to *S* (equation 11). A threshold of *t* means that among S colonizer strains, one strain has *t* module genes and the remaining strains have less than *t* module genes. The probability $F_S(t)$ includes this case but also sets of S colonizer strains where all have less than t module genes. The latter cases are all included in $F_S(t\text{-}1)$. Therefore, the probability of observing a threshold *t* is determined by the difference of those two probabilities. The p-value for threshold significance is obtained by the sum of the probabilities of thresholds lower or equal to the one observed (equation 12). Statistical significance of 0.05 is required for each module to be maintained.

$$p(t_{obs}) = \sum_{t=1}^{t_{obs}} F_s(t) - F_S(t-1) \qquad \text{Equation 12}$$

Module trimming

The initial analysis of the modules resulting from the described selection process showed that in some cases, the last genes added to the growing modules where highly frequent genes. These genes did not contribute significantly to the predictive power of the module, although they did not prevent correct predictions, since they were also present in the invasive strains that contained the remaining genes of the module. This observation suggested the need for a module trimming step that would remove non-informative genes. Knowing that each module is present in a subset of the invasive strains, a Fisher's right tailed exact test was used to assess the individual association of module genes with their presence in that subset of invasive strains. Genes that were not significantly associated with the subset of invasive strains were excluded from the module. After this trimming, the threshold is recalculated and re-evaluated as described above. From this step forward we select only the largest module among the ones generated from the same seed.

False discovery rate

To correct for multiple testing, 1000 random gene presence matrices were used as inputs. Random matrices maintain the number of genes and strains as the original matrix and also maintain each gene frequency. Among other module properties, the module length distribution obtained with the random matrices presented values that tended to be lower than the module lengths obtained from the original gene presence matrix (Figure 4). This means that in the absence of a real association between the gene presence and the invasive classification of the strains it is rare to find large gene modules that are only present in invasive strains. Using this property, a module size cut-off s was applied, such that only modules with k or more genes were selected. For every possible cut-off value the False Discovery Rate (FDR) was estimated as the ratio between the mean

number of gene modules selected with random gene presence matrices and the number of gene modules selected with the original gene presence matrix. For k=24 FDR was lower than 0.05, meaning that only 5% of the modules selected from the original data were expected to be falsely associated with invasive strains.

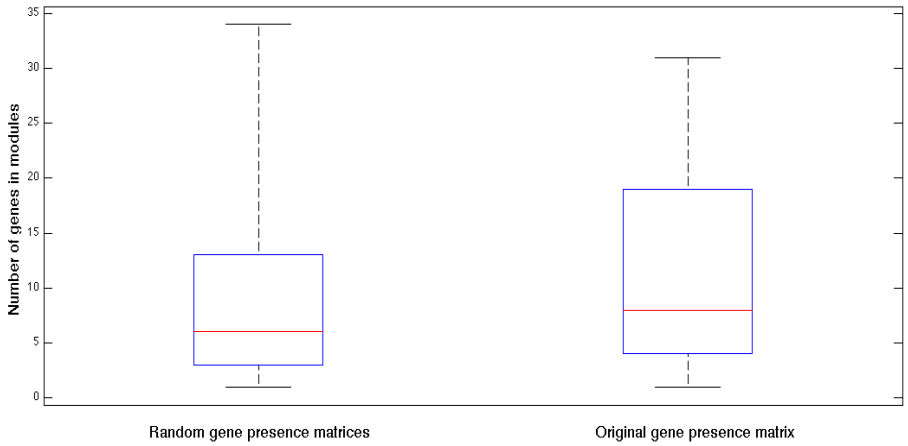

Figure 4 - Boxplot of module lengths with random and original gene presence matrices

Module's robustness

The microarray CGH data analysis has an inherent error. Validation experiments using the same microarray platform indicate an 8% error in presence/absence classification [50]. To measure the impact a misclassification would have on the module's IPV, the gene presence matrix was randomly changed in 8% of its values. Each module's IPV is then calculated using the changed matrix. This is performed 5000 times and the robustness estimated as the percentage of the repetitions that resulted in an IPV above 0.95.

Implementation

The proposed method for module search and selection was implemented in the Matlab computational environment (version 7.10.0 (R2010a), Mathworks Inc.). Functions and scripts used in this thesis are included in the supplemental files.

Results and Discussion

The search for coherent gene modules has wielded 30 modules, from which 26 are unique. Each module components and evaluated properties are presented in a supplemental file.

Table 3 – Evolution of the number of modules through the selection steps

	Modules	Seeds with modules
Initial	64750	1295
Significant runs test	16989	1048
IPV=1	6665	364
Significant threshold	3534	197
Trimmed	3018	190
FDR corrected	-	30
Unique	-	26

The runs test proved to be an efficient filter to reduce the computation time in subsequent steps (Table 3). Among the modules selected in the runs test, over one third had a threshold value that allowed IPV=1. The fact that these modules were selected in the runs test and the adaptability of threshold definition enables such elevated fraction. Similarly, more than half of the modules with IPV=1 presented a statistically significant threshold value. The multiple testing correction showed that modules with more than 24 genes are unlikely (FDR<0.05) to result from random gene presence matrices. Still, random datasets were able to yield modules more efficiently than expected. One reason for this observation is the fact that the gene distance matrix is created using the same data that is later used to evaluate the modules. This way, it is easier to successfully create gene modules with IPV=1, significant in the runs test and in the threshold test. If, during the analysis of random datasets, the original gene distance matrix is used, instead of redefining it for each random gene presence matrix, the number of modules that survived the three filtering steps would

strongly decrease. Using established biological networks, that encode information that is obtained independently of the gene presence matrix, would also avoid this situation. Additionally, it would have an impact on the composition of selected modules, facilitating the interpretation of the association between a biological function or pathway and invasiveness.

The modules are highly overlapped in their content (Figure 5). Overlap is calculated as the percentage of a module's genes (vertical axis) present in another module (horizontal axis). With 26 modules and a minimal size of 24 genes (due to the multiple testing correction), only 111 genes are used. These genes have varied frequencies of occurrence in the 47 strains (Figure 6). Overlap between modules was expected to result from the network approach used. If two gene seeds are each other closest neighbours, than they are likely to generate similar modules.

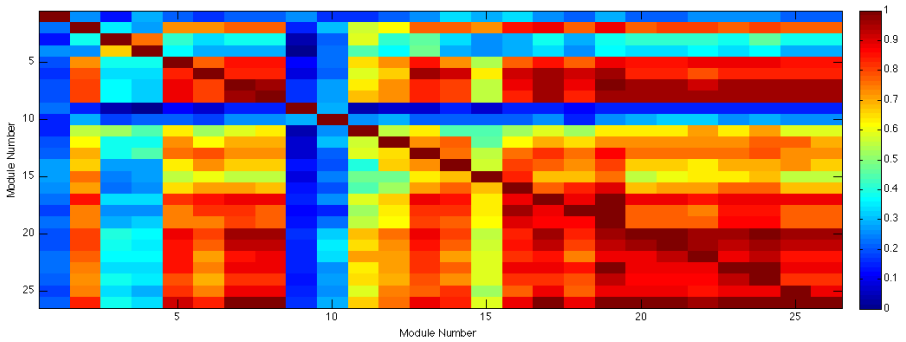

Figure 5 - Module Overlap

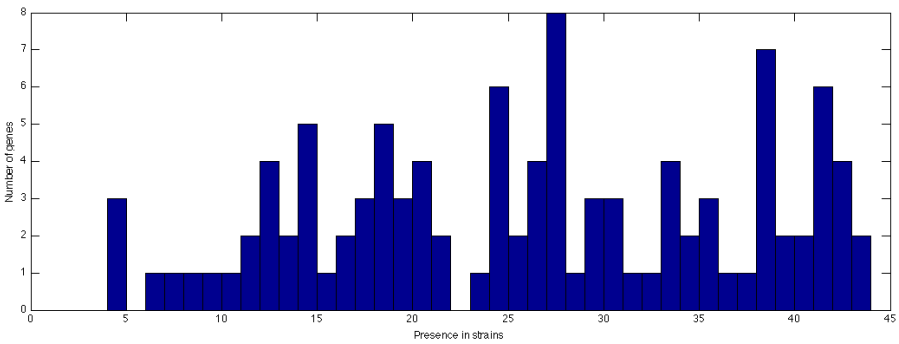

Figure 6 - Presence of the selected genes in strains

In Figure 7, the modules are linked according to their content similarity through a dendrogram, and it is clear that some sub-modules are present in multiple modules. Connections between leaves in the dendrogram are made based on Jaccard distance between modules and with a complete or "furthest neighbour" linkage, meaning the distance between two clusters is the largest distance between two of their elements. Using only the closely related modules on the left side of the dendrogram (modules 5-8, 13, 14, 16-26) and comparing the genes they include, some of the genes are almost constant while others are sporadic (Figure 8), strengthening the idea that sub-modules are the cause of the high overlap. If these sub-modules were selected as independent modules they were later discarded for not being the largest module from a seed or by the multiple testing correction.

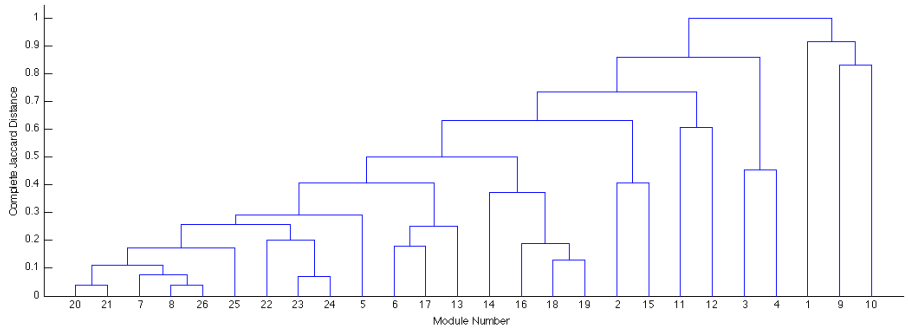

Figure 7 - Dendrogram of module based on gene composition.

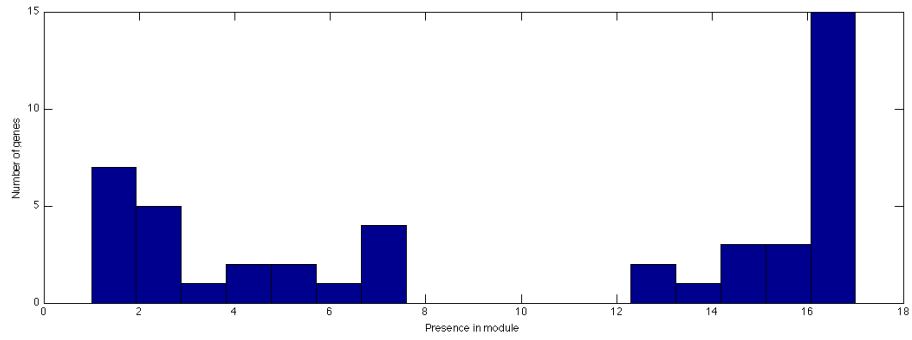

Figure 8 - Bimodal distribution of gene presence

Selected modules were not required to be present in all invasive strains but only to be absent in colonizer strains. Nevertheless, the modules are present in at

least 10 (out of 31) invasive strains and up to 23 strains. Such positive association was unsuccessfully pursued with colonization. Modules with a Colonization Predictive Value (CPV) equal to one were not found. This may result from inappropriate criteria of module selection when applied to colonization or some other technical reason. On the other hand, it may stem from a biological property of *Streptococcus pneumoniae*, where all strains are naturally colonizers and therefore possess the same mechanisms to assure a successful colonization.

The modules selected are robust against minor changes in the gene presence matrix (all modules present an IPV>0.95 in more than 99% of the analysis with gene presence matrices randomly changed in 8% of their elements). The microarray experiment from which the data was originated had an inherent error and this high robustness provides confidence in the authenticity of the findings, showing they are unlikely to be a consequence of microarray error. Existence of a threshold to define presence of modules instead of requiring presence of all modules' components is likely to be a source of robustness against minor disturbances in the presence profiles of the genes.

The number of gene module present in a strains correlates with the strains' invasiveness odd's ratio. Invasiveness' odd's ratio expresses the ratio between the frequency of invasive disease and the frequency of asymptomatic colonization caused by a related group of strains. This analysis included neutral strains, which were not used to search and select modules (Figure 9 and Supplemental files).

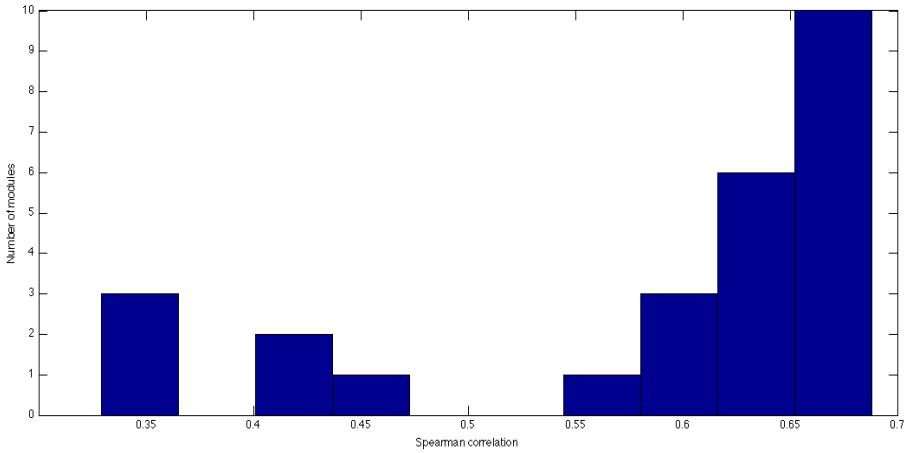

Figure 9 – Correlation between individual module gene presence and invasiveness odd's ratio

Spearman correlation analysis was used to test if there is a monotonic relationship between the two variables (number of module genes present and invasiveness odd's ratio). Correlation values (*r*) range from -1 (inverse correlation) to 1 (positive correlation). When the variables are not related, the *r* is expected to be 0. For all the modules the correlation value was positive and statistically greater than 0. The majority (20) of the modules present strong correlations (0.55-0.70). These correlation values are considered strong due both to the origin of the invasiveness odd's ratios and to the complexity of the invasive phenotype. The odd's ratios were estimated in observational epidemiological studies, with high associated variability, and the invasiveness can be influenced by so many variables that it would be difficult to achieve very high correlations. Using the global number of genes that belong to any module present in each strain also shows a significant and strong correlation with the invasiveness odd's ratio (rho=0.6160, $p<10^{-8}$), even when considering the neutral strains (Figure 10). Correlation with strains' odd's ratio is an important confirmation of the modules validity. Neutral strains were not used as input data, nor were the invasive and colonizer strains odd's ratio and therefore the algorithm doesn't influence the correlation of modules gene presence with the odd's ratio.

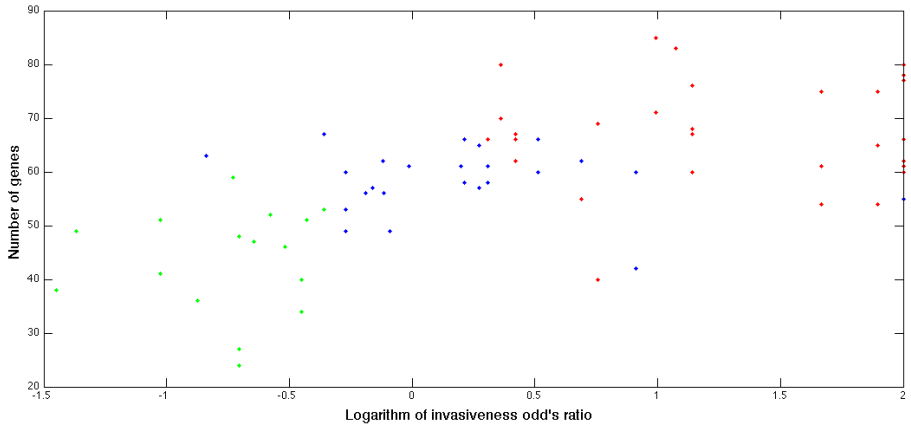

Figure 10 – Correlation between global modules genes presence and invasiveness odd's ratio

Gene functional analysis

One objective of the present work is to uncover cellular functions that contribute to the invasiveness of pneumococcal strains. This can be achieved through the functional analysis of the 111 genes that are members of the 26 invasive modules. A fraction of those genes have no known function. Their presence in invasive modules suggests that their function should be directly or indirectly related with the invasiveness of the strains containing these genes. For the remaining genes it was possibly to collect functional information from JCVI annotation (http://cmr.jcvi.org) and literature search.

Statistical enrichment of specific gene functions was assessed using the annotation from JCVI, both for each module individually and for the set of 111 genes present in all the selected modules. Although some functions are frequent among the selected genes, these functions were also very frequent in the total pool of pneumococcal genes, disabling a statistically significant enrichment. Absence of functional enrichment should not be interpreted as lack of relevance of that function for strain invasiveness. Transport associated genes, which is the most frequent function among the selected genes, are highly frequent in pneumococcal genomes. However, additional or mutated genes may provide advantage in a different medium or in the use of an additional subtract. Module-wise functional enrichment was not found either. Functional enrichment in a module would suggest it was mainly associated with a concrete pathway or

defined biological function. Here, lack of functional enrichment suggests that modules include several different functions that may need to act synergistically to impact strain invasiveness.

The 111 genes present in all the modules were manually assigned to 17 functional classes that are presented and discussed in the following sections.

Anaerobic Response

Table 4 - Genes associated with anaerobic response

Spots	Primary targets	Gene symbol	Description
153	SP0202	NrdD	Anaerobic Ribonucleoside-Tryphosphate reductase
156	SP0205	NrdG	Anaerobic Ribonucleoside-Tryphosphate reductase activating protein
2506	SPN08064	-	Anaerobic Nucleotide Reductase

Ribonucleotide reductases (RNR) provide all living organisms with the deoxyribonucleoside triphosphates required for DNA synthesis [51]. Class III RNRs, such as nrdD, are homodimeric and contain a stable oxygen-sensitive glycyl radical formed by S-adenosylmethionine (SAM) and a second protein, nrdG, that it is not required for catalysis once the radical has been generated [52]. This class is functional only under strict anaerobic conditions, contrasting with Class I – strict aerobic conditions – and with Class II – both aerobic and anaerobic. Importance of nrdD to virulence has been reported in several organisms such as *E. Coli* [53] , *S. Aureus* [54], *P. Aeruginosa* [53] and *S. Sanguinis* [55], but studies in *S. Pneumoniae* have not been conducted so far.

Choline binding proteins

Table 5 - Genes associated with Choline binding proteins

Spots	Primary targets	Gene symbol	Description
2861	SPN14033	PcpC	Choline binding protein F
3613	spr1995	PspC	Choline binding protein A

Streptococcus pneumoniae possesses a family of proteins that bind the phosphocholine present in the membrane and the cell wall [56,57]. The choline-binding proteins (CBP) of pneumococci and other gram positive organisms contain structurally similar choline-binding domains, which are composed of multiple tandem aminoacid repeats. CBPs have been reported to contribute to adherence of pneumococci through effects on surface charge [57,58].

PspC (Pneumococcal surface protein), also known as CbpA, is a highly polymorphic protein and its impact on pneumococcal virulence varies between strains [59,60]. Of the many biological functions that have been attributed to PspC, its role in adherence is one of the most studied [61]. While PspC does not exclusively govern pneumococcal adherence, its absence significantly reduces adherence and invasion of human cells. PspC also binds the secretory component of human secretory immunoglobulin A and human factor H [61], as well as complement component C3 [58,62].

PcpC is a paralogue of CbpF [63]. A study on cell adherence showed that mutants deficient in only CbpF had no observable phenotype [64]. On the other hand, it appears to be important to counteract cellular lysis, as a negative regulator of LytC activity in *S. Pneumoniae*. Deletion of PcpC was found to promote a significant increase in competence-induced lysis [63].

Transcription regulation

Table 6 - Genes associated with transcription regulation

Spots	Primary targets	Gene symbol	Description
293	SP0395	-	Transcriptional regulator, putative
103	SP0141	-	Transcriptional regulator
286	SP0386	-	Sensor histidine kinase, putative
880	SP1187	LacT	Transcription antiterminator LacT
3054	spr0104	-	Hypothetical protein
3307	spr0907	PhtD	Pneumococcal histidine triad protein D precursor

Transcriptional regulation is a highly complex process with many known and unknown interactions. Several promoters, enhancers and transcription factors among other molecules interact to either repress or active transcription of genes. Lac operon contains genes involved in transport and metabolism of lactose. LacE and LacF code for a lactose-specific phosphoenol-pyruvate-dependent phosphotransferase system (PTS) and LacG for phospho- β -galactosidase, an enzyme that hydrolyzes lactose to glucose and galactose-6-P [65]. LacT, also one of the proteins encoded in the operon, is an antiterminator. It binds to premature transcriptional termination sequences in the intercistronic region of the operon and prevents transcription termination [66]. Mutants with complete deletion of lacT present a Lac^- phenotype [67].

Pht is a protein family restricted to the *Streptococcus* genus that share a histidine triad motif, HxxHxH, repeated five to six times in their aminoacid sequences [59,68]. PhtD in particular is highly frequent in strains and displays litle variability in nucleotide sequence [68]. PhtD is involved in pneumococcal lung-specific virulence [28], without further indication about their biological function.

Insertion Sequences/Transposons

Table 7 - Genes associated with insertion sequences/transposons

Spots	Primary targets	Gene symbol	Description
533	SP0700	-	Transposase, IS30 family, degenerate
1061	SP1441	-	IS66 family element, Orf3, degenerate
1470	SP2018	-	Transposase, IS3 family, degenerate
3225	spr0612	IS1239	Degenerate transposase ; truncayion
3330	spr0958	-	Tn5252, relaxase, truncation
3433	spr1298	Transposase E	IS66 family, Degenerate transposase

A unique recombination-mediated genetic plasticity is a distinctive feature of *Streptococcus pneumoniae* and a key to its success as a pathogen [69]. Transposable elements (TEs) are mobilized to change their location in DNA by the action of a transposase (TPase), an enzyme which catalyses transposition. The presence and activity of TEs may lead to structural changes in both the size and composition of a genome. TEs can generate various mutations, such as insertions, deletions, duplications, inversions and translocations of even large DNA fragments [70].

Insertion sequences (ISs) comprise a large group of bacterial transposable DNA elements. These elements vary in size from 0.7 to 3.5 kb. IS elements generally encode a transposase and duplicate a sequence of several base pairs at the target site on transposition. IS elements are classified in families. IS66 is a family of IS with similar composition, coding for 3 ORFs with unknown function. [71]

Transposons are larger TEs (>20 kB). Tn5252 (~47.5 kb) is a conjugative transposon containing a chloramphenicol resistance gene (cat) and an integrase gene (int5252) driving its site-specific insertion into the host cell genome. Other genetic and functional properties of Tn5252 are still largely unknown. Multiresistant *S. pneumoniae* isolates have disclosed new larger genetic elements both carrying catpC194 in their Tn5252-like transposon and sharing an identical genome integration site [72].

tRNA

Table 8 - Genes associated with tRNA

Spots	Primary targets	Gene symbol	Description
93	SP0129	Gcp	Probable tRNA threonylcarbamoyladenosine biosynthesis protein Gcp; glycoprotease family protein
1712	SPN01230	-	Glucose inhibited division protein A, fragment
3186	spr0492	ValS	Valyl-tRNA synthetase
3441	spr1329	GlyQ	Glycyl-tRNA synthetase alpha subunit

The accuracy of protein synthesis depends on the specific recognition of aminoacids and tRNAs by aminoacyl-tRNA synthetases. Several aminoacyl-tRNA synthetases, however, have difficulty discriminating between cognate and structurally similar aminoacids, resulting in misactivation of noncognate amino acids and misacylation of tRNA [73]. Valyl-tRNA synthetase (ValS) should discriminate the cognate valine from the noncognate, isosteric threonine, which has a quite similar shape and size as valine and a hydroxyl group in its side chain instead of a methyl group [74]. To maintain high translational fidelity, these aminoacyltRNA synthetases catalyse proofreading (editing) reactions in which the misproducts are hydrolyzed.

Gcp homologs are ubiquitous in all three kingdoms of life however, these Gcp homologs do not exhibit glycoprotease activity as predicted [75]. It has been revealed that these Gcp homologs are required for cell viability of many bacterial species examined to date, including *Streptococcus pneumoniae* [76]. Recently it has been reported that Gcp plays an important role in the modulation of the branched-chain aminoacids biosynthesis pathway [77], being involved in the biosynthesis of N6-threonylcarbamoyladenosine (t6A) and tRNA modification [78].

Cell envelope

Table 9 - Genes associated with cell envelope

Spots	Primary targets	Gene symbol	Description
2771	SPN12014	-	YFMI protein, efflux transporter.
3310	spr0916	MesH	MesH protein
254	SP0349	Cps4D	Capsular polysaccharide biosynthesis protein Cps4D
2591	SPN08230	-	53.6 KDa protein
2593	SPN08232	-	25.5 KDa protein; putative chain length regulator; putatitve chain length terminator
3116	spr0304	Pbpx	Penicillin-binding protein 2X
3591	spr1903	GalU	UTP-glucose-1-phosphate uridylyltransferase

Pneumococcal capsule genes are expressed in one single operon, cap [79]. One of the most striking features of the pneumococcal cap locus is its huge genetic divergence, since only a few genes are conserved among different clusters. Cps4D is part of the cap locus and is refered in the literature as being involved in virulence but with no information regarding function [80,81].

A functional gal regulon, including galU, regulates the supply of UDP-galactose and UDP-glucose, two major precursors for the biosynthesis of capsular polysaccharide (CPS) in bacteria [82,83]. GalU is an UDP-glucose pyrophosphorylase responsible for synthesis of UDP-glucose from glucose 1-phosphate and UTP [82,84]. UDP-glucose plays a well-established biochemical role as a glycosyl donor in the enzymatic biosynthesis of carbohydrates [82,85]. galU mutants were dominated by abnormal capsule synthesis, and several of the galU mutants also had defects in other surface proteins [84].

Streptococci are among the most penicillin-sensitive organisms but are known to develop high level resistance [9]. The target proteins of β-lactams such as penicillin are enzymes that catalyse late steps in murein biosynthesis, a constituent of the cell wall [9,86]. In *Streptococcus pneumoniae*, altered low-affinity forms of

penicillin-binding proteins (PBPs) 1a, 2x, and 2b are found invariably in penicillin-resistant clinical isolates [69,87]. Several mutations in one PBP may be required to cause a substantial decrease in penicillin affinity, and more than one PBP may have to be changed into low affinity variants in order to achieve high resistance levels. In *Streptococcus pneumoniae*, alterations in the PBP2x and PBP2b confer low resistance and are the prerequisite for high levels of resistance [9,23].

Spermidine metabolism

Table 10 - Genes associated with spermidine metabolism

Spots	Primary targets	Gene symbol	Description
683	SP0918	SpeE	Spermidine synthase
685	SP0920	NspC	Carboxynorspermidine decarboxylase

Polyamines, such as spermidine and norspermidine, are small ubiquitous aliphatic hydrocarbon molecule and polycationic at physiological pH [88]. Polyamines exist primarily as complexes with RNA molecules in prokaryotes and are essential for efficient DNA replication, transcription and translation [89,90].

In the *Streptococcus* genus, choline-binding proteins are important virulence factors. Pneumococci can substitute choline with polyamines, under laboratory conditions, which then get incorporated in the bacterial cell wall [91]. Additionally, environmental polyamine acquisition may play an important role in pneumococcal response to temperature, oxidative and pH stress. In *Streptococcus pneumoniae*, polyamines have been linked to *in vivo* fitness, pathogenesis and virulence factor expression [88,91]. SP0918 and SP0920 synthetize spermidine and norspermidine, respectivily, and are therefore linked to virulence.

Osmotic Regulation

Table 11 - Genes associated with osmotic regulation

Spots	Primary targets	Gene symbol	Description
747	SP1010	MscL	Large conductance mechanosensitive channel
3516	spr1604	AqpZ	Aquaporin Z - water channel

Aqueous environments are prone to drastic changes in extracellular osmolality [92-94]. In order to respond to osmotic pressure shifts, *Streptococcus pneumoniae* has membrane proteins that change its permeability to water or ions. Besides the osmotic changes in aqueous mediums, pneumococcus also suffers from osmotic shifts when invades a new tissue and, therefore, response to osmotic pressure shifts are essential for invasive traits.

Aquaporins belong to a large family of proteins that increase the rate of diffusion of water and glycerol across cell membranes [95]. Aquaporin Z in particular is a water channel protein implicated in important processes such as osmoregulation and volume expansion in rapidly growing cells [96]. The aqpZ gene was found to be osmotically regulated, as its expression was significantly increased in steady-state hypoosmotic conditions, but greatly reduced in hyperosmolar media [95]. The need for an aquaporin to adapt to hypoosmotic stresses has been regarded as puzzling, because the presence of water channels may result in an excessive build-up of turgor pressure, which may lead to ruptures of the cell wall and consequent cell death [96]. AqpZ-like proteins seem to be necessary for the virulence expressed by some pathogenic bacteria [96].

MscL is a mechanosensitive channel with low ion selectivity and high conductance that opens in response to mechanically imposed membrane stress or osmotic shifts [97]. In response to changes in the lipid bilayer tension, the MscL channel proteins gate and open transiently to form large aqueous pores, through which both solutes and solvents can pass rapidly [94]. Consequently, the concentration of water-attracting solutes inside the cells is lowered, turgor is reduced, and cell lysis is prevented. These channels are major routes for the

release of cytoplasmic solutes to achieve a rapid reduction of the turgor pressure during the transition from media of high to low osmolality [98].

Co-enzymes

Table 12 - Genes associated with co-enzymes

Spots	Primary targets	Gene symbol	Description
506	SP0665	-	Chorismate binding enzyme
2051	SPN04086	-	Hydroxyethylthiazole kinase (EC 2.7.1.50)
3565	spr1783	KdtB	Phosphopantetheine adenylyltransferase

Co-enzymes, consist of small organic molecules that many enzymes require as cofactors to be catalytically active. Their impact on cellular physiology can be profound as many different processes may be affected.

Chorismate is an important precursor of several biomolecules such as aromatic aminoacids, vitamins K, ubiquinone and folic acid [99]. It is of special importance as it is the branch point in the biosynthesis pathway of the referred molecules.

Hydroxyethylthiazole kinase participates in thiamine metabolism, which is a cofactor of several key enzymes, particularly involved in carbohydrate metabolism [100].

KdtB is the phosphopantetheine adenylyltransferase enzyme, responsible for the biosynthesis of Coenzyme A [101].

Proteolysis

Table 13 - Genes associated with Proteolysis

Spots	Primary targets	Gene symbol	Description
3055	spr0105	-	Transporter, truncation
3056	spr0106	-	Transporter, truncation
1790	SPN02074	PepF	Oligoendopeptidase F homolog (EC 3.4.24.-)
3222	spr0607	-	Dipeptidase

Streptococcus pneumoniae is a microorganism with multiple aminoacid requirements and bacteria are often able to utilize peptides as a source of aminoacids as well or better than utilizing the aminoacids themselves [102]. This way, enzymes with proteolytic activity play a major role in fulfilling the cell aminoacid requirements. Spr0607 activity is highly specific to dipeptides and has a putative role in protein synthesis and degradation [102]. SPN02074 is an homolog of the oligoendopeptidase F (pepF) which has collagenase activity. PepF is a cytoplasmic endopeptidase that hydrolyses oligopeptides but cannot degrade proteins[103]. An homologue enzyme in Group B Streptococci was found not to be a collagenase and it has been suggested to be involved in degradation of peptides transported to the cell's interior. This would require a oligopeptide transport system but nothing similar has yet been reported in pneumococcus [104]. PepF expression has been found to be increased during the competence state as a late gene [21].

Competence

Table 14 - Genes associated with competence

Spots	Primary targets	Gene symbol	Description
89	SP0123	Ccs1	Competence-induced protein Ccs1
708	SP0954	CelA	Competence protein CelA
3553	spr1724	SsbB	Single-strand DNA-binding protein
3581	spr1861	CglD	Competence protein

In *Streptococcus pneumoniae*, the competence state, sometimes called X-state [19], is triggered by a peptide called CSP (competence stimulating peptide). Upon interacting with CSP, a two component signalling system (comD and comE) initiates the transcription of 24 gene set (early genes), including comX, which encodes for an alternative sigma factor, σX. ComX is responsible for the transcription of a second class of genes, termed late genes, which code for effectors for DNA uptake and recombination and other functions that are only beginning to be understood [105-107]. The competence regulon of *Streptococcus pneumoniae* has been shown to cross regulate virulence [108]

The four genes present in invasive modules are all late genes [21,109]. SsB is a single-strand DNA (ssDNA) binding protein. It plays a direct role in the stabilization of internalized ssDNA and its cellular concentration is adjusted so as to handle very large quantities of ssDNA [110].

CglD gene codes for a 135 aminoacid protein with a N-terminal hydrophobic segment. This N-terminal part resembles a hydrophobic region that is conserved in pilins from various gram-negative bacterial species and was implicated in protein-protein interaction of pilin subunits involved in the assembly of pili. Similar pilin-like proteins in *B. subtilis*, ComG3 to ComG7, are thought to form a multimeric complex for the binding and uptake of transforming DNA [111].

CelA gene is about 40% identical to a *B. subtilis* 205 aminoacid transmembrane protein ComEA, which is required for DNA uptake during genetic transformation [111].

Ccs1 is also a late gene of the competence state but has no known function. [21]

DNA recombination/repair

Table 15 - Genes associated with DNA recombination/repair

Spots	Primary targets	Gene symbol	Description
89	SP0123	Ccs1	Competence-induced protein Ccs1
708	SP0954	CelA	Competence protein CelA
3553	spr1724	SsbB	Single-strand DNA-binding protein
3581	spr1861	CglD	Competence protein
86	SP0119		MutT-nudix family protein
130	SP0173	HexB	DNA mismatch repair
850	SP1147	-	Integrase-recombinase, phage integrase family, truncation
1215	SP1669	-	MutT-nudix family protein
1301	SP1783	-	MutT-nudix family protein
1354	SP1849	DpnD	DpnD protein
3058	spr0108	-	Hypothetical protein
3093	spr0238	RuvB	Holliday junction DNA helicase RuvB

Strains of *Streptococcus pneumoniae* can harbour one of two restriction systems, DpnI or DpnII. The DpnI cassette consists of an operon containing two genes, dpnC and dpnD. While dpnC has been shown to encode the DpnI restriction endonuclease, dpnD encodes an 18 kDa protein of unknown function [112].

During genetic recombination a heteroduplex joint is formed between two homologous DNA molecules. The heteroduplex joint plays an important role in recombination since it accommodates sequence heterogeneities (mismatches, insertions or deletions) that lead to genetic variation. RuvB is responsible for widening the disparity margin allowed between the two DNA molecules and this way promoting genetic variation. In *E. Coli*, RuvB action was shown to be facilitated by the presence of ssb, a ssDNA binding protein [113].

HexB is a DNA replication editor correcting potentially mutagenic mismatches [114]. While it is crucial to maintain the genome integrity, it does not prevent intra or inter species recombination. However, it has been reported that HexB mutants display higher transformation efficiency [115].

The MutT proteins or "Nudix" hydrolases are a family of versatile, widely distributed, "housecleaning" enzymes. Computer searches revealed that this protein family is present in organisms ranging from viruses to humans. All the enzymes of this family characterized so far hydrolyze a nucleoside diphosphate linked to some other moiety, X (nu-di-X).[116] Studies show that MutT mutants have mutation frequencies from 100 to 10000 fold higher [117]. Interestingly, these mutT knock outs causes, exclusively, a single, unidirectional AT to CG transversion [118].

Transport

Table 16 - Genes associated with transport

Spots	Primary targets	Gene symbol	Description
3055	spr0105	-	Transporter, truncation
3056	spr0106	-	Transporter, truncation
747	SP1010	MscL	Large conductance mechanosensitive channel protein MscL
3516	spr1604	AqpZ	Aquaporin Z - water channel protein
2771	SPN12014	-	YFMI protein, efflux transporter.
3310	spr0916	MesH	MesH protein
293	SP0395	-	Transcriptional regulator, putative
185	SP0250	-	PTS system, IIC component
228	SP0310	-	PTS system, IIC component

294	SP0396	MtlF	PTS system, mannitol-specific IIA component
711	SP0957	-	ABC transporter, ATP-binding protein
1138	SP1553	-	ABC transporter, ATP-binding protein
1179	SP1618	-	PTS system, IIB component
1588	SP2198	-	ABC transporter, permease protein
3089	spr0221	ABC-MSP	ABC transporter membrane-spanning permease - iron transport
3108	spr0278	PTS-EII	Phosphotransferase system sugar-specific EII component
3143	spr0356	MtlA	Mannitol PTS EII
3393	spr1183	ABC-NBD-truncation	ABC transporter ATP-bindng protein - possibly multidrug efflux, truncation
3425	spr1290	ABC-N/P	ABC transporter ATP-binding/membrane-spanning protein - unknown substrate
3428	spr1293	ABC-NBD	ABC transporter ATP-binding protein - unknown substrate

Streptococcus pneumoniae has demanding nutritional needs and the availability of key nutrients in the host modulates the expression of bacterial phenotypes that may affect disease outcome [88]. A large proportion of the pneumococcal genome is devoted to basic metabolic functions such as transport of nutrients, and this is of particular importance, as it spends most of its life cycle on nutritionally restricted mucosal surfaces [15,119].

PTS transporters (phosphoenolpyruvate:carbohydrate phosphotransferase system) are the most abundant group of carbohydrate transporters in the pneumococcus [120] and each pneumococcal genome contains between 15 and 20

PTS transporters [121]. PTS systems are involved in the transport of a large number of carbohydrates and its phosphorylation during transport [122]. Their role in regulation of metabolic networks has been described several times [123-125]. The basic composition of the PTS is similar across all species studied so far [122,123]. It is comprised of two "general" cytoplasmic components, EI and HPr, which are common to all PTS carbohydrates. Carbohydrate specificity resides in EII, and hence, bacteria usually contain many different EIIs [123].

The second most abundant group of transport systems are ABC (ATP-binding cassetes) transporters [121,126]. Generally, ABC transporters consist of two permeases, two ATP binding cassetes called NBD (nucleotide binding domain) and a domain with substrate specificity [127]. These transport complexes transport not only sugars, but also metal ions, like iron [128] or manganese, polyamines [126] and export peptides, toxins and multiple drugs [126,129].

Carbohydrate metabolism

Table 17 - Genes associated with carbohydrate metabolism

Spots	Primary targets	Gene symbol	Description
3310	spr0916	MesH	MesH protein
293	SP0395	-	Transcriptional regulator, putative
185	SP0250	-	PTS system, IIC component
228	SP0310	-	PTS system, IIC component
294	SP0396	MtlF	PTS system, mannitol-specific IIA component
1179	SP1618	-	PTS system, IIB component
3108	spr0278	PTS-EII	Phosphotransferase system sugar-specific EII component
3143	spr0356	MtlA	Mannitol PTS EII
223	SP0303	BglA	6-phospho-beta-glucosidase
295	SP0397	MtlD	Mannitol-1-phosphate 5-

			dehydrogenase
3247	spr0704	-	Hypothetical protein

Streptococcus pneumoniae is able to utilize at least thirty-two substrates as carbon sources, including sugars like mannose, galactose and hyaluronic acid, but also many food-borne monosaccharides, disaccharides and polysaccharides [121,124,125]. This is a hallmark of a bacteria adapted to a wide variety of ecological niches, each one providing a wide variety of carbon sources. The sequencing of Tigr4's genome showed that a large part of it is dedicated to the transport and metabolism of sugars [15]. When more than one substrate is available, the cell has mechanisms to regulate its protein expression in order to use only one of them at a time. CcpA [119,125,130] and RegM [131] are two of the most studied proteins to play this regulatory role and its possible that SP0395 also participates in this process.

MtlD is a mannitol-1-phosphate 5-dehydrogenase. After mannitol is translocated by the respective PTS and phosphorilated to mannitol-1-phosphate, MtlD catalyses its conversion to fructose-6-phosphate [132].

BglA, also called celA, is a 6-phospho-beta-glucosidase. It belongs to the cellobiose metabolism operon which also contains a PTS transporter, a putative DNA-binding protein and two proteins of unknown function [133,134].

Acetyltransferases

Table 18 - Genes associated with acetyltransferases

Spots	Primary targets	Gene symbol	Description
92	SP0128	-	Ribosomal-protein-alanine acetyltransferase, putative
707	SP0953	-	Acetyltransferase, GNAT family
1995	SPN03274	-	Hypothetical protein homologous to SPT02104
3290	spr0850	-	Acetyltransferase, GNAT family, hypothetical protein

The transfer of an acetyl group from one molecule to another is a fundamental biochemical process and the best understood set of acetyltransferases is the GNAT superfamily (Gcn5-related N-acetyltransferase) [135]. Gcn5 is a histone acetyltransferase (HAT) and plays a role in epigenetic regulation. The GNAT superfamily is spread throughout eukariotic and prokariotic species and its members present a wide range of substrate specificity [135,136]. SP0953 and spr0850 substrates are unknown.

Toxins

Table 19 - Genes associated with toxins

Spots	Primary targets	Gene symbol	Description
3055	spr0105	-	Transporter, truncation
2861	SPN14033	-	Choline binding protein F; PCPC.
3613	spr1995	PspC	Choline binding protein A
400	SP0531	BlpI	Bacteriocin BlpI
401	SP0532	BlpJ	Bacteriocin BlpJ
404	SP0535	PncG	PncG immunity protein; hypothetical protein
405	SP0536	BlpL	Immunity protein BlpL
778	SP1050	-	Antitoxin PezA, transcriptional regulator, putative
1984	SPN03247	-	YMFA protein.
2505	SPN08061	-	Cardiolipin synthetase (EC 2.7.8.-)
2851	SPN14017	-	Hypothetical protein homologous to SPT01142

Bacteriocins are small heat stable peptides common in gram-positive bacteria. The production of bacteriocins has been related to the producer's ability to colonize a host more efficiently due the ability of these peptides to eliminate competitor strains, which often include other species [137].

The blp locus of pneumococcus encodes a number of putative bacteriocin like peptides which have been implied in intra and inter species competition [138]. The locus contains open reading frames for a typical two-component regulatory system (blpR and blpH), a small peptide pheromone (blpC), and a dedicated ABC transporter (blpA and blpB). BlpI (or pncA) and blpJ (or pncD) were found to code for a bacteriocin like peptide [137,138], while blpL still has no known function

[137]. Lux et al also identified pncG has a putative immunity protein [137].

PezA is an anti-toxin part of a toxin anti-toxin system (TA) with its counterpart PezT, a toxin. TA systems functions are not consensual and roles as growth control, stabilization of mobile genetic elements, or programmed cell death have been proposed. In pneumococcus, PezA mutants enable survival to PezT toxicity. PezT mutants display attenuated infection progress and were out-competed by wild type strains [139]. Attenuation of infectiveness was not observed upon disturbance of the other TA system identified in pneumococcus [140]. Together, this observations point PezAT to play a virulence bound role rather than a classical TA system role ensuring stable inheritance [139].

Cardiolipin is a lipid which has been reported as indirect inhibitor of phage-associated cell lysis [141]. It has also been shown to be required for the synthesis of hyaluronan [142,143]. Hyaluronan is a saccharide polymer found throughout several species, playing very different roles. In Streptococcus group A and C (beta-haemolytic Streptococci, while pneumococcus is alpha-haemolytic), hyaluronan is important to the capsule synthesis and, in this way to its pathogenicity [143].

Aminoacid metabolism

Table 20 - Genes associated with aminoacid metabolism

Spots	Primary targets	Gene symbol	Description
506	SP0665		Chorismate binding enzyme
3052	spr0102	ArgG	Argininosuccinate synthase
3053	spr0103	ArgH	Argininosuccinate lyase
3291	spr0852	Ald-truncation	Alanine dehydrogenase, truncation
3292	spr0853	Ald-truncation	Alanine dehydrogenase, truncation
3293	spr0854	Ald-truncation	Alanine dehydrogenase, truncation

3374	spr1096	MetY-truncation	O-acetylhomoserine sulfhydrylase, truncation
3391	spr1181	GdhA	Glutamate dehydrogenase

Analysis of the genome sequence of S. pneumoniae D39 showed that it contains only two genes that are putatively involved in arginine biosynthesis, argG and argH, encoding the enzymes for the conversion of citrulline (and aspartate) to arginine. There are, however, other pathways that lead to the synthesis of arginine, e.g., carA and carB synthetize carbamoyl-phosphate from glutamate, which is a precursor of arginine [144].

Glutamate dehydrogenase (GDH) activity catalyses the reversible oxidative deamination of glutamate to 2-oxoglutarate and ammonia, mainly using NADP as the co-factor [145]. This allows pneumococcus to degrade glutamate and to convert aminoacids to aromatic compounds. GDH activity has been correlated with virulence in *Streptococcus suis* and *Clostridium botulinum* [146].

Methionine synthesis occurs by methylation of homocysteine. In turn, homocysteine can be synthetized through a sulfhydrylation pathway, in which homocysteine is synthesized from O-acylhomoserine and sulfide by O-acylhomoserine sulfhydrylase (metY) [147,148].

Functional interactions with invasiveness

Although no functional enrichment was found using JCVI annotation, some selected genes have clear functional interactions. The ribozyme nrdD is activated by nrdG and both were selected in the same module [149]. argG and argH catalyse sequential reactions that constitute an alternative arginine syntheses pathway [144]. The enzyme mannitol-1-phosphate 5-dehydrogenase (mtlD) uses as substrate mannitol-1-phosphate, which is the result of the transport of mannitol by PTS system (mtlA and mtlF) [132]. RuvB activity as an enhancer of recombination is facilitated by the presence of ssDNA binding protein, ssb [113]. Assembly of an ABC transporter requires several components such as permeases and ATP binding cassettes (or NBDs) and several genes of both of them were selected [126]. Aquaporin Z (aqpZ) action in hipoosmotic conditions puzzles the science community because its action would seemingly build up excessive turgor

pressure in the cell [95]. Large mechanosensitive channel (mscL) provides an effective response to excessive turgor pressure and may be a biological counterpart of aqpZ. Polyamines, as spermidine and norspermidine, have been reported to be possible substitutes of choline and are therefore important in cell wall structure and possible in interaction with choline binding proteins [98].

Most genes selected have been previously linked with invasiveness or have a plausible connection with invasiveness. Cell envelope proteins and choline binding proteins play a major role in protection against host immunity [64,79]. Both are important in inhibiting human complement action on the cell, either by removal of its components or by binding factor H. Several transposable elements (TEs) have been identified in or near the capsular locus and have been reported to affect the transcriptional regulation [71].

Invasion of new tissue requires a swift adaptation to a new environment, both in terms of physical properties and nutrient availability. Osmotic response genes present in invasive modules are directed towards strong shifts in osmolality rather than fine tuning and therefore are of particular interest in adaptation to new environments [95,98]. Anaerobic response genes nrdD and its activator nrdG are hardly functional in the nasopharynx, since they are strict anaerobic [149]. Inside the human organism however, oxygen concentration is low since it is almost always bound to biological molecules such as haemoglobin and nrdD may be crucial to compensate the function of the aerobic ribonucleoside-tryphosphate reductases. Adaptation to different carbon sources has also been proven to be crucial for invasiveness [150]. The high number of sugar transporters is related with the capability of invasive strains to strive in different medium composition. Similarly, some selected genes code for enzymes involved in the metabolism of different sugars. Proteolysis genes are likely related with the nutritional demand for aminoacids [102].

Protein synthesis is a constant process in any bacteria and demands a permanent pool of aminoacids and tRNA. Aminoacid synthesis genes are part of some invasive modules and provide alternative pathways of synthesis, using alternative substrates or the same substrates in a more economic way [144,151].

Genes related to the synthesis and editing of tRNA bound with the respective aminoacid were also selected [78,152].

Finally, the huge heterogeneity of pneumococcus genome originates from its recombination capabilities [13]. Gene transfer allows the bacteria to adapt and overcome different adversities. Some of the genes present in invasive modules promote genome heterogeneity by enhancing recombination with extracellular DNA [113]. Among the genes selected, DNA internalization, stabilization and enhancement of recombination with heterogenic DNA are promoted [153]. The natural competence of pneumococcus is complemented with the capacity to induce lysis in neighbour cells, which releases DNA fragments. Several bacterocins were associated with invasiveness, as well as genes that inhibited cell lysis [139]. These genes give the cell a natural advantage in competition with other pathogens.

Globally, the algorithm has proven to fulfil the objective of module identification. It has room for improvement, especially in the handling of false discovery results. To avoid false discovery results it is possible that a large number of true associations have been dismissed. Analysis of the modules provides optimism about the authenticity of the selected modules. The modules robustness to minor changes in the gene presence matrix discards the influence of microarray errors on results. It is, however, impossible to discard it completely as not only there is doubt concerning the presence or absence of a gene but also concerning its true nature. Identification of gene presence uses a probe with 70 nucleotides and no information exists about the conservation of the rest of the gene. On the other hand, gene absence may be caused by few mutations on the sequence recognized by the probe while the majority of the gene is preserved.

Furthermore, interaction between genes is hardly understood due to the many loopholes in known pathways. To understand the true biological function of a gene it is crucial to fill the gaps in the overall cell knowledge and, this way, understand the pathways in which it is involved, the physical interactions it makes and how it is regulated.

Conclusion

In this thesis a new algorithm to predict phenotype from genotype is presented and is used to find gene modules associated with invasive strains. Unlike other approaches, the main goal is not to achieve optimal predictive power of invasiveness in strains but rather to unveil the gene modules that may be responsible for the invasive traits. The method uses a network-based heuristic to minimize the solution search space and to restrain the positive results to gene modules with biological relevance. As a first approach an artificial gene distance matrix was created based on gene co-occurrence and co-invasiveness. In the end, 26 gene modules were found to be present (above a specific threshold) only in invasive strains and to be robust against estimated input data error. Presence of gene modules in strains was found to correlate positively with the strain's invasiveness odd's ratio. This correlation is valid to neutral strains as well, even though they were not used in module selection.

The modules include 111 unique genes. Some of the genes were previously described as crucial to invasiveness in *Streptococcus Pneumoniae*, such as PsPC or PthD, while other genes' impact on invasiveness has not been determined experimentally. Many of the genes associated with invasiveness have unknown or only putative biological functions that need to be identified and characterized. The ones that have identified functions may be strong candidates to be invasiveness determinants when co-present with other genes in the same module. In this scenario are genes linked with cell envelope, transport, sugar metabolism, osmotic response, aminoacid synthesis, spermidine synthesis and proteolysis.

The algorithm proved to be able to establish association of gene modules with the invasive class. Moreover, it was able to return rules that dictate associations that are easily translated to a biological level: increase of module genes presence increases the invasiveness odd's ratio.

Future perspectives

The present work has identified several genes with unknown function, with putative functions or with identified functions in other organism. To understand their part in invasiveness it is necessary to understand their biological role first and, therefore, laboratorial work needs to be executed. Moreover, while the algorithm has associated them with invasiveness, it is not possible to clarify what influence, if any, they have on invasiveness. This work is based in the analysis of observational studies and therefore it is not possible to ascertain more than correlation relationships and the genes identified may cause invasive traits or be no more than a genetic lineage marker of invasive strains.

The algorithm rationale was devised to use biological networks. In this work an artificial gene distance matrix was created and used as a first approach. Use of a biological network is the logical next step, allowing the algorithm to perform with the input data it was devised for. An additional step will need to be added to the algorithm where a gene distance matrix is generated from the network. This is thought to eliminate some of the problems of using a gene distance matrix generated from the gene presence data. In this scenario the same data source is used to generate the modules and to evaluate them, leading more easily to false positive results. Using a different source to generate the module will likely tackle this issue.

The algorithm itself is a work in progress and for further development new datasets are required. It is a tool to solve classification problems using a network-based heuristic and is, theoretically, adaptable to classification problems of any background. It was, nonetheless, designed for a specific biological problem and therefore incorporates specific assumptions. In the near future, biological classification problems, preferentially phenotype-genotype association problems would be the perfect input data for testing and developing the algorithm.

References

1. Bogaert, D., De Groot, R. & Hermans, P. W. M. Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *Lancet Infect Dis* **4**, 144–154 (2004).
2. Mehr, S. & Wood, N. Streptococcus pneumoniae – a review of carriage, infection, serotype replacement and vaccination. *Paediatric Respiratory Reviews* 1–7 (2012).doi:10.1016/j.prrv.2011.12.001
3. Austrian, R. The pneumococcus at the millennium: not down, not out. *J. Infect. Dis.* **179**, S338–S341 (1999).
4. Lopez, R. Pneumococcus: the sugar-coated bacteria. *Int. Microbiol.* **9**, 179–190 (2006).
5. Austrian, R. The Gram stain and the etiology of lobar pneumonia, an historical note. *Bacteriological Reviews* **24**, 261 (1960).
6. van der Poll, T. & Opal, S. M. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. *The Lancet* **374**, 1543–1556 (2009).
7. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.* **6**, 288–301 (2008).
8. Tillett, W. S., Cambier, M. J. & McCormack, J. E. The Treatment of Lobar Pneumonia and Pneumococcal Empyema with Penicillin. *Bull N Y Acad Med* **20**, 142–178 (1944).
9. Hakenbeck, R., Grebe, T., Zähner, D. & Stock, J. B. beta-lactam resistance in Streptococcus pneumoniae: penicillin-binding proteins and non-penicillin-binding proteins. *Molecular Microbiology* **33**, 673–678 (1999).
10. Cardozo, D. M., Nascimento-Carvalho, C. M. C., Souza, F. R. & Silva, N. M. S. Nasopharyngeal colonization and penicillin resistance among pneumococcal strains: a worldwide 2004 update. *Braz J Infect Dis* **10**, 293–304 (2006).
11. Austrian, R. The Jeremiah Metzger Lecture: Of gold and pneumococci: a history of pneumococcal vaccines in South Africa. *Trans. Am. Clin. Climatol. Assoc.* **89**, 141–161 (1978).
12. Weinberger, D. M., Harboe, Z. B., Flasche, S., Scott, J. A. & Lipsitch, M. Prediction of Serotypes Causing Invasive Pneumococcal Disease in Unvaccinated and Vaccinated Populations. *Epidemiology* **22**, 199–207 (2011).
13. Donati, C. *et al.* Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biology* **11**, R107 (2010).
14. Hava, D. L., LeMieux, J. & Camilli, A. From nose to lung: the regulation behind Streptococcus pneumoniae virulence factors. *Molecular Microbiology* **50**, 1103–1110 (2003).
15. Tettelin, H. *et al.* Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. *Science* **293**, 498–506 (2001).
16. Mitchell, A. M. & Mitchell, T. J. Streptococcus pneumoniae: virulence

factors and variation. *Clinical Microbiology and Infection* **16**, 411–418 (2010).
17. Lapierre, P. & Gogarten, J. P. Estimating the size of the bacterial pan-genome. *Trends in Genetics* **25**, 107–110 (2009).
18. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11**, 472–477 (2008).
19. Claverys, J.-P., Martin, B. & Håvarstein, L. S. Competence-induced fratricide in streptococci. *Molecular Microbiology* **64**, 1423–1433 (2007).
20. Bartilson, M. *et al.* Differential fluorescence induction reveals Streptococcus pneumoniae loci regulated by competence stimulatory peptide. *Molecular Microbiology* **39**, 126–135 (2001).
21. Peterson, S. N. *et al.* Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays. *Molecular Microbiology* **51**, 1051–1070 (2003).
22. Rimini, R. *et al.* Global analysis of transcription kinetics during competence development in Streptococcus pneumoniae using high density DNA arrays. *Molecular Microbiology* **36**, 1279–1292 (2000).
23. Hakenbeck, R. *et al.* Mosaic Genes and Mosaic Chromosomes: Intra- and Interspecies Genomic Variation of Streptococcus pneumoniae. *Infection and Immunity* **69**, 2477–2486 (2001).
24. Weiser, J. N. The pneumococcus: why a commensal misbehaves. *J Mol Med* **88**, 97–102 (2009).
25. Weinberger, D. M. *et al.* Association of Serotype with Risk of Death Due to Pneumococcal Pneumonia: A Meta-Analysis. *CLIN INFECT DIS* **51**, 692–699 (2010).
26. Obert, C. *et al.* Identification of a Candidate Streptococcus pneumoniae Core Genome and Regions of Diversity Correlated with Invasive Pneumococcal Disease. *Infection and Immunity* **74**, 4766–4777 (2006).
27. Melin, M., Trzciński, K., Meri, S., Käyhty, H. & Väkeväinen, M. The capsular serotype of Streptococcus pneumoniae is more important than the genetic background for resistance to complement. *Infection and Immunity* **78**, 5262–5270 (2010).
28. Hava, D. L. & Camilli, A. Large-scale identification of serotype 4 Streptococcus pneumoniae virulence factors. *Molecular Microbiology* **45**, 1389–1406 (2002).
29. Lau, G. W. *et al.* A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. *Molecular Microbiology* **40**, 555–571 (2001).
30. Polissi, A. *et al.* Large-Scale Identification of Virulence Genes fromStreptococcus pneumoniae. *Infection and Immunity* **66**, 5620–5629 (1998).
31. Obert, C. A., Gao, G., Sublett, J., Tuomanen, E. I. & Orihuela, C. J. Assessment of molecular typing methods to determine invasiveness and to differentiate clones of Streptococcus pneumoniae. *Infect. Genet. Evol.* **7**, 708–716 (2007).
32. Schmidt, M. C. *et al.* NIBBS-Search for Fast and Accurate Prediction of Phenotype-Biased Metabolic Systems. *PLoS Comput Biol* **8**, e1002490

(2012).
33. Slonim, N., Elemento, O. & Tavazoie, S. Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks. *Mol Syst Biol* **2**, 1–14 (2006).
34. Önskog, J., Freyhult, E., Landfors, M., Rydén, P. & Hvidsten, T. R. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. *BMC Bioinformatics* **12**, 390 (2011).
35. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annual review of genomics and human genetics* **2**, 343–372 (2001).
36. Chuang, H.-Y., Hofree, M. & Ideker, T. A Decade of Systems Biology. *Annu. Rev. Cell Dev. Biol.* **26**, 721–744 (2010).
37. Sieberts, S. K. & Schadt, E. E. Moving toward a system genetics view of disease. *Mamm Genome* **18**, 389–401 (2007).
38. Bhaskar, H., Hoyle, D. C. & Singh, S. Machine learning in bioinformatics: A brief survey and recommendations for practitioners. *Computers in Biology and Medicine* **36**, 1104–1125 (2006).
39. Larranaga, P. Machine learning in bioinformatics. *Briefings in Bioinformatics* **7**, 86–112 (2006).
40. Szymczak, S. *et al.* Machine learning in genome-wide association studies. *Genet. Epidemiol.* **33**, S51–S57 (2009).
41. Jensen, L. J. & Bateman, A. The rise and fall of supervised machine learning techniques. *Bioinformatics* **27**, 3331–3332 (2011).
42. Kamruzzaman, S. M. & Sarkar, A. M. J. A New Data Mining Scheme Using Artificial Neural Networks. *Sensors* **11**, 4622–4647 (2011).
43. Moore, J. H., Asselbergs, F. W. & Williams, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **26**, 445–455 (2010).
44. Winter, C. *et al.* Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. *PLoS Comput Biol* **8**, e1002511 (2012).
45. Hurley, D. *et al.* Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic Acids Research* **40**, 2377–2398 (2012).
46. Hoskins, J. *et al.* Genome of the Bacterium Streptococcus pneumoniae Strain R6. *J. Bacteriol.* **183**, 5709–5717 (2001).
47. Dopazo, J. *et al.* Annotated Draft Genomic Sequence from a *Streptococcus pneumoniae* Type 19F Clinical Isolate. *Microbial Drug Resistance* **7**, 99–125 (2001).
48. Sa-Leao, R. *et al.* 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. *Journal of Clinical Microbiology* **49**, 1369–1375 (2011).
49. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. (Springer: 2001).
50. Cardoso, L. *Classificação de genes em hibridação genómica*

comparativa de estirpes de Streptococcus pneumoniae. (Dissertação de Mestrado em Bioestatística na Faculdade de Ciências da Universidade de Lisboa: 2009).

51. Eliasson, R. *et al.* The anaerobic ribonucleoside triphosphate reductase from Escherichia coli requires S-adenosylmethionine as a cofactor. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 3314–3318 (1990).
52. Torrents, E. The Anaerobic Ribonucleotide Reductase from Lactococcus lactis. INTERACTIONS BETWEEN THE TWO PROTEINS NrdD AND NrdG. *Journal of Biological Chemistry* **276**, 33488–33494 (2001).
53. Sjoberg, B. M. & Torrents, E. Shift in Ribonucleotide Reductase Gene Expression in Pseudomonas aeruginosa during Infection. *Infection and Immunity* **79**, 2663–2669 (2011).
54. Masalha, M., Borovok, I., Schreiber, R., Aharonowitz, Y. & Cohen, G. Analysis of Transcription of the Staphylococcus aureus Aerobic Class Ib and Anaerobic Class III Ribonucleotide Reductase Genes in Response to Oxygen. *J. Bacteriol.* **183**, 7260–7272 (2001).
55. Paik, S. *et al.* Identification of Virulence Determinants for Endocarditis in Streptococcus sanguinis by Signature-Tagged Mutagenesis. *Infection and Immunity* **73**, 6064–6074 (2005).
56. Brooks-Walter, A., Briles, D. E. & Hollingshead, S. K. The pspC gene of Streptococcus pneumoniae encodes a polymorphic protein, PspC, which elicits cross-reactive antibodies to PspA and provides immunity to pneumococcal bacteremia. *Infection and Immunity* **67**, 6533–6542 (1999).
57. Rosenow, C. *et al.* Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of Streptococcus pneumoniae. *Molecular Microbiology* **25**, 819–829 (1997).
58. Ogunniyi, A. D. *et al.* Contributions of Pneumolysin, Pneumococcal Surface Protein A (PspA), and PspC to Pathogenicity of Streptococcus pneumoniae D39 in a Mouse Model. *Infection and Immunity* **75**, 1843–1851 (2007).
59. Melin, M. *et al.* Interaction of Pneumococcal Histidine Triad Proteins with Human Complement. *Infection and Immunity* **78**, 2089–2098 (2010).
60. Yuste, J. *et al.* The Effects of PspC on Complement-Mediated Immunity to Streptococcus pneumoniae Vary with Strain Background and Capsular Serotype. *Infection and Immunity* **78**, 283–292 (2009).
61. Quin, L. R. *et al.* Factor H Binding to PspC of Streptococcus pneumoniae Increases Adherence to Human Cell Lines In Vitro and Enhances Invasion of Mouse Lungs In Vivo. *Infection and Immunity* **75**, 4082–4087 (2007).
62. Kerr, A. R. *et al.* The Contribution of PspC to Pneumococcal Virulence Varies between Strains and Is Accomplished by Both Complement Evasion and Complement-Independent Mechanisms. *Infection and Immunity* **74**, 5319–5324 (2006).
63. Eldholm, V. *et al.* The Pneumococcal Cell Envelope Stress-Sensing System LiaFSR Is Activated by Murein Hydrolases and Lipid II-

Interacting Antibiotics. *J. Bacteriol.* **192**, 1761–1773 (2010).
64. Gosink, K. K., Mann, E. R., Guglielmo, C., Tuomanen, E. I. & Masure, H. R. Role of Novel Choline Binding Proteins in Virulence of Streptococcus pneumoniae. *Infection and Immunity* **68**, 5690–5695 (2000).
65. Alpert, C. A. & Siebers, U. The lac operon of Lactobacillus casei contains lacT, a gene coding for a protein of the Bg1G family of transcriptional antiterminators. *J. Bacteriol.* **179**, 1555–1562 (1997).
66. Rutberg, B. Antitermination of transcription of catabolic operons. *Molecular Microbiology* **23**, 413–421 (1997).
67. Gosalbes, M. J., Esteban, C. D. & Pérez-Martínez, G. In vivo effect of mutations in the antiterminator LacT in Lactobacillus casei. *Microbiology (Reading, Engl.)* **148**, 695–702 (2002).
68. Rioux, S. *et al.* Transcriptional regulation, occurrence and putative role of the Pht family of Streptococcus pneumoniae. *Microbiology* **157**, 336–348 (2011).
69. Claverys, J. P., Prudhomme, M., Mortier-Barrière, I. & Martin, B. Adaptation to the environment: Streptococcus pneumoniae, a paradigm for recombination-mediated genetic plasticity? *Molecular Microbiology* **35**, 251–259 (2000).
70. Dziewit, L. *et al.* Insights into the Transposable Mobilome of Paracoccus spp. (Alphaproteobacteria). *PLoS ONE* **7**, e32277 (2012).
71. Han, C. G., Shiga, Y., Tobe, T., Sasakawa, C. & Ohtsubo, E. Structural and Functional Characterization of IS679 and IS66-Family Elements. *J. Bacteriol.* **183**, 4296–4304 (2001).
72. Mingoia, M., Tili, E., Manso, E., Varaldo, P. E. & Montanari, M. P. Heterogeneity of Tn5253-Like Composite Elements in Clinical Streptococcus pneumoniae Isolates. *Antimicrobial Agents and Chemotherapy* **55**, 1453–1459 (2011).
73. Tardif, K. D. Functional group recognition at the aminoacylation and editing sites of E. coli valyl-tRNA synthetase. *RNA* **10**, 493–503 (2004).
74. Fukunaga, R. Structural Basis for Non-cognate Amino Acid Discrimination by the Valyl-tRNA Synthetase Editing Domain. *Journal of Biological Chemistry* **280**, 29937–29945 (2005).
75. Zheng, L., Yu, C., Bayles, K., Lasa, I. & Ji, Y. Conditional Mutation of an Essential Putative Glycoprotease Eliminates Autolysis in Staphylococcus aureus. *J. Bacteriol.* **189**, 2734–2742 (2007).
76. Lei, T. *et al.* The C-Terminal Domain of the Novel Essential Protein Gcp Is Critical for Interaction with Another Essential Protein YeaZ of Staphylococcus aureus. *PLoS ONE* **6**, e20163 (2011).
77. Lei, T. *et al.* The Essentiality of Staphylococcal Gcp Is Independent of Its Repression of Branched-Chain Amino Acids Biosynthesis. *PLoS ONE* **7**, e46836 (2012).
78. Deutsch, C., Yacoubi, El, B., de Crécy-Lagard, V. & Dirk Iwata-Reuyl Biosynthesis of threonylcarbamoyl adenosine (t6A), a universal tRNA nucleoside. *J. Biol. Chem* 13666–13673 (2012).
79. Moscoso, M. & García, E. Transcriptional regulation of the capsular polysaccharide biosynthesis locus of streptococcus pneumoniae: a

bioinformatic analysis. *DNA Res.* **16**, 177–186 (2009).
80. Pandya, U., Sinha, M., Luxon, B. A., Watson, D. A. & Niesel, D. W. Global transcription profiling and virulence potential of Streptococcus pneumoniae after serial passage. *Gene* **443**, 22–31 (2009).
81. Jothi, R., Manikandakumar, K., Ganesan, K. & Parthasarathy, S. On the analysis of the virulence nature of TIGR4 and R6 strains of Streptococcus pneumoniae using genome comparison tools. *Journal of chemical sciences-Bangalore-* **119**, 559 (2007).
82. Aanensen, D. M., Mavroidi, A., Bentley, S. D., Reeves, P. R. & Spratt, B. G. Predicted Functions and Linkage Specificities of the Products of the Streptococcus pneumoniae Capsular Biosynthetic Loci. *J. Bacteriol.* **189**, 7856–7876 (2007).
83. Chang, H. Y., Lee, J. H., Deng, W. L., Fu, T. F. & Peng, H. L. Virulence and outer membrane properties of a galU mutant of Klebsiella pneumoniae CG43. *Microb. Pathog.* **20**, 255–261 (1996).
84. Vilches, S. *et al.* Mesophilic Aeromonas UDP-glucose pyrophosphorylase (GalU) mutants show two types of lipopolysaccharide structures and reduced virulence. *Microbiology* **153**, 2393–2404 (2007).
85. Stimson *et al.* Meningococcal pilin: a glycoprotein substituted with digalactosyl 2,4- diacetamid-2,4,6-trideoxyhexose. *Mol Microbiol* 1201–1214 (1995).
86. Ghuysen, J. M. Molecular structures of penicillin-binding proteins and beta-lactamases. *Trends in Microbiology* **2**, 372–380 (1994).
87. Rieux, V., Carbon, C. & Azoulay-Dupuis, E. Complex relationship between acquisition of β-lactam resistance and loss of virulence in Streptococcus pneumoniae. *J. Infect. Dis.* **184**, 66–72 (2001).
88. Shah, P., Nanduri, B., Swiatlo, E., Ma, Y. & Pendarvis, K. Polyamine biosynthesis and transport mechanisms are crucial for fitness and pathogenesis of Streptococcus pneumoniae. *Microbiology* **157**, 504–515 (2011).
89. Igarashi, K. & Kashiwagi, K. Polyamines: Mysterious Modulators of Cellular Functions. *Biochem. Biophys. Res. Commun.* **271**, 559–564 (2000).
90. Shah, P. & Swiatlo, E. A multifaceted role for polyamines in bacterial pathogens. *Molecular Microbiology* **68**, 4–16 (2008).
91. Shah, P., Romero, D. G. & Swiatlo, E. Role of polyamine transport in Streptococcus pneumoniae response to physiological stress and murine septicemia. *Microb. Pathog.* **45**, 167–172 (2008).
92. Wahome, P. G. & Setlow, P. Growth, osmotic downshock resistance and differentiation of Bacillus subtilis strains lacking mechanosensitive channels. *Arch Microbiol* **189**, 49–58 (2007).
93. Wood, J. M. *et al.* Osmosensing and osmoregulatory compatible solute accumulation by bacteria. *Comparative Biochemistry and Physiology-Part A: Molecular & Integrative Physiology* **130**, 437–460 (2001).
94. Wood, J. M. Osmosensing by bacteria: signals and membrane-based sensors. *Microbiology and Molecular Biology Reviews* **63**, 230–262 (1999).
95. Soupene, E., King, N., Lee, H. & Kustu, S. Aquaporin Z of Escherichia

coli: Reassessment of Its Regulation and Physiological Role. *J. Bacteriol.* **184**, 4304–4307 (2002).
96. Calamita, G. The Escherichia coli aquaporin-Z water channel. *Molecular Microbiology* **37**, 254–262 (2000).
97. Hoffmann, T., Boiangiu, C., Moses, S. & Bremer, E. Responses of Bacillus subtilis to Hypotonic Challenges: Physiological Contributions of Mechanosensitive Channels to Cellular Survival. *Applied and Environmental Microbiology* **74**, 2454–2460 (2008).
98. Levina, N. *et al.* Protection of Escherichia coli cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. *EMBO J.* **18**, 1730–1737 (1999).
99. Fernandes, C. L., Breda, A., Santiago Santos, D., Augusto Basso, L. & Osmar Norberto de Souza A structural model for chorismate synthase from Mycobacterium tuberculosis in complex with coenzyme and substrate. *Computers in Biology and Medicine* **37**, 149–158 (2007).
100. Müller, I. B. *et al.* The Vitamin B1 Metabolism of Staphylococcus aureus Is Controlled at Enzymatic and Transcriptional Levels. *PLoS ONE* **4**, e7656 (2009).
101. Freiberg, C. *et al.* Identification of novel essential Escherichia coli genes conserved among pathogenic bacteria. *J. Mol. Microbiol. Biotechnol.* **3**, 483–489 (2001).
102. Johnson, M. K. Physiological roles of pneumococcal peptidases. *J. Bacteriol.* **119**, 844–847 (1974).
103. Kanamaru, K., Stephenson, S. & Perego, M. Overexpression of the PepF oligopeptidase inhibits sporulation initiation in Bacillus subtilis. *J. Bacteriol.* **184**, 43–50 (2002).
104. Lin, B. *et al.* Characterization of PepB, a group B streptococcal oligopeptidase. *Infection and Immunity* **64**, 3401–3406 (1996).
105. Johnsborg, O. & HÃ varstein, L. S. Regulation of natural genetic transformation and acquisition of transforming DNA in Streptococcus pneumoniae. *FEMS Microbiology Reviews* **33**, 627–642 (2009).
106. Yother, J., McDaniel, L. S. & Briles, D. E. Transformation of encapsulated Streptococcus pneumoniae. *J. Bacteriol.* **168**, 1463–1465 (1986).
107. Piotrowski, A., Luo, P. & Morrison, D. A. Competence for Genetic Transformation in Streptococcus pneumoniae: Termination of Activity of the Alternative Sigma Factor ComX Is Independent of Proteolysis of ComX and ComW. *J. Bacteriol.* **191**, 3359–3366 (2009).
108. Zhu, L. & Lau, G. W. Inhibition of Competence Development, Horizontal Gene Transfer and Virulence in Streptococcus pneumoniae by a Modified Competence Stimulating Peptide. *PLoS Pathog* **7**, e1002241 (2011).
109. Luo, P. & Morrison, D. A. Transient Association of an Alternative Sigma Factor, ComX, with RNA Polymerase during the Period of Competence for Genetic Transformation in Streptococcus pneumoniae. *J. Bacteriol.* **185**, 349–358 (2003).

110. Attaiech, L. *et al.* Role of the Single-Stranded DNA–Binding Protein SsbB in Pneumococcal Transformation: Maintenance of a Reservoir for Genetic Plasticity. *PLoS Genet* **7**, e1002156 (2011).
111. Pestova, E. V. & Morrison, D. A. Isolation and Characterization of Three Streptococcus pneumoniae Transformation-Specific Loci by Use of alacZ Reporter Insertion Vector. *J. Bacteriol.* **180**, 2701–2710 (1998).
112. la Campa, de, A. G., Springhorn, S. S., Kale, P. & Lacks, S. A. Proteins encoded by the DpnI restriction gene cassette. Hyperproduction and characterization of the DpnI endonuclease. *Journal of Biological Chemistry* **263**, 14696–14702 (1988).
113. Parsons, C. A., Stasiak, A. & West, S. C. The E. coli RuvAB proteins branch migrate Holliday junctions through heterologous DNA sequences in a reaction facilitated by SSB. *EMBO J.* **14**, 5736 (1995).
114. Humbert, O., Prudhomme, M., Hakenbeck, R., Dowson, C. G. & Claverys, J. P. Homeologous recombination and mismatch repair during transformation in Streptococcus pneumoniae: saturation of the Hex mismatch repair system. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9052–9056 (1995).
115. Burghout, P. *et al.* Search for Genes Essential for Pneumococcal Transformation: the RadA DNA Repair Protein Plays a Role in Genomic Recombination of Donor DNA. *J. Bacteriol.* **189**, 6540–6550 (2007).
116. Bessman, M. J., Frick, D. N. & O'Handley, S. F. The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed,"housecleaning" enzymes. *Journal of Biological Chemistry* **271**, 25059–25062 (1996).
117. Treffers, H. P., Spinelli, V. & Belser, N. O. A Factor (or Mutator Gene) Influencing Mutation Rates in Escherichia Coli. *Proc. Natl. Acad. Sci. U.S.A.* **40**, 1064–1071 (1954).
118. Yanofsky, C., Cox, E. C. & Horn, V. The unusual mutagenic specificity of an E. Coli mutator gene. *Proc. Natl. Acad. Sci. U.S.A.* **55**, 274 (1966).
119. Iyer, R., Baliga, N. S. & Camilli, A. Catabolite control protein A (CcpA) contributes to virulence and regulation of sugar metabolism in Streptococcus pneumoniae. *J. Bacteriol.* **187**, 8340–8349 (2005).
120. Ren, Q., Chen, K. & Paulsen, I. T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research* **35**, D274–D279 (2007).
121. Bidossi, A. *et al.* A Functional Genomics Approach to Establish the Complement of Carbohydrate Transporters in Streptococcus pneumoniae. *PLoS ONE* **7**, e33320 (2012).
122. Postma, P. W., Lengeler, J. W. & Jacobson, G. R. Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiol. Rev.* **57**, 543–594 (1993).
123. Deutscher, J., Francke, C. & Postma, P. W. How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria. *Microbiology and Molecular Biology Reviews* **70**, 939–1031 (2006).

124. Marion, C. *et al.* Streptococcus pneumoniae Can Utilize Multiple Sources of Hyaluronic Acid for Growth. *Infection and Immunity* **80**, 1390–1398 (2012).
125. Kaufman, G. E. & Yother, J. CcpA-Dependent and -Independent Control of Beta-Galactosidase Expression in Streptococcus pneumoniae Occurs via Regulation of an Upstream Phosphotransferase System-Encoding Operon. *J. Bacteriol.* **189**, 5183–5192 (2007).
126. Davidson, A. L., Dassa, E., Orelle, C. & Chen, J. Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiology and Molecular Biology Reviews* **72**, 317–364 (2008).
127. Marion, C., Aten, A. E., Woodiga, S. A. & King, S. J. Identification of an ATPase, MsmK, Which Energizes Multiple Carbohydrate ABC Transporters in Streptococcus pneumoniae. *Infection and Immunity* **79**, 4193–4200 (2011).
128. Brown, J. S., Gilliland, S. M. & Holden, D. W. A Streptococcus pneumoniae pathogenicity island encoding an ABC transporter involved in iron uptake and virulence. *Molecular Microbiology* **40**, 572–585 (2001).
129. Putman, M., van Veen, H. W. & Konings, W. N. Molecular Properties of Bacterial Multidrug Transporters. *Microbiology and Molecular Biology Reviews* **64**, 672–693 (2000).
130. Carvalho, S. M., Kloosterman, T. G., Kuipers, O. P. & Neves, A. R. CcpA Ensures Optimal Metabolic Fitness of Streptococcus pneumoniae. *PLoS ONE* **6**, e26707 (2011).
131. Giammarinaro, P. & Paton, J. C. Role of RegM, a Homologue of the Catabolite Repressor Protein CcpA, in the Virulence of Streptococcus pneumoniae. *Infection and Immunity* **70**, 5454–5461 (2002).
132. Honeyman, A. L. & Curtiss, R. Isolation, characterization, and nucleotide sequence of the Streptococcus mutans mannitol-phosphate dehydrogenase gene and the mannitol-specific factor III gene of the phosphoenolpyruvate phosphotransferase system. *Infection and Immunity* **60**, 3369–3375 (1992).
133. McKessar, S. J. & Hakenbeck, R. The Two-Component Regulatory System TCS08 Is Involved in Cellobiose Metabolism of Streptococcus pneumoniae R6. *J. Bacteriol.* **189**, 1342–1350 (2007).
134. Shafeeq, S., Kloosterman, T. G. & Kuipers, O. P. CelR-mediated activation of the cellobiose-utilization gene cluster in Streptococcus pneumoniae. *Microbiology* **157**, 2854–2861 (2011).
135. Sterner, D. E. & Berger, S. L. Acetylation of Histones and Transcription-Related Factors. *Microbiology and Molecular Biology Reviews* **64**, 435–459 (2000).
136. Dyda, F., Klein, D. C. & Hickman, A. B. GCN5-related N-acetyltransferases: a structural overview. *Annu Rev Biophys Biomol Struct* **29**, 81–103 (2000).
137. Lux, T., Nuhn, M., Hakenbeck, R. & Reichmann, P. Diversity of Bacteriocins and Activity Spectrum in Streptococcus pneumoniae. *J. Bacteriol.* **189**, 7741–7751 (2007).
138. Dawid, S., Roche, A. M. & Weiser, J. N. The blp Bacteriocins of

Streptococcus pneumoniae Mediate Intraspecies Competition both In Vitro and In Vivo. *Infection and Immunity* **75**, 443–451 (2006).
139. Mutschler, H., Reinstein, J. & Meinhart, A. Assembly dynamics and stability of the pneumococcal epsilon zeta antitoxin toxin (PezAT) system from Streptococcus pneumoniae. *Journal of Biological Chemistry* **285**, 21797–21806 (2010).
140. Mutschler, H. & Meinhart, A. ε/ζ systems: their role in resistance, virulence, and their potential for antibiotic development. *J Mol Med* **89**, 1183–1194 (2011).
141. Garcia, P., López, R., Ronda, C., García, E. & Tomasz, A. Mechanism of phage-induced lysis in pneumococci. *J. Gen. Microbiol.* **129**, 479–487 (1983).
142. Tlapak-Simmons, V. L., Baggenstoss, B. A., Clyne, T. & Weigel, P. H. Purification and lipid dependence of the recombinant hyaluronan synthases from Streptococcus pyogenes and Streptococcus equisimilis. *Journal of Biological Chemistry* **274**, 4239–4245 (1999).
143. Tlapak-Simmons, V. L., Kempner, E. S., Baggenstoss, B. A. & Weigel, P. H. The active streptococcal hyaluronan synthases (HASs) contain a single HAS monomer and multiple cardiolipin molecules. *Journal of Biological Chemistry* **273**, 26100–26109 (1998).
144. Kloosterman, T. G. & Kuipers, O. P. Regulation of Arginine Acquisition and Virulence Gene Expression in the Human Pathogen Streptococcus pneumoniae by Transcription Regulators ArgR1 and AhrC. *Journal of Biological Chemistry* **286**, 44594–44605 (2011).
145. Tanous, C., Chambellon, E. & Yvon, M. Sequence analysis of the mobilizable lactococcal plasmid pGdh442 encoding glutamate dehydrogenase activity. *Microbiology* **153**, 1664–1675 (2007).
146. Kutz, R. & Okwumabua, O. Differentiation of Highly Virulent Strains of Streptococcus suis Serotype 2 According to Glutamate Dehydrogenase Electrophoretic and Sequence Type. *Journal of Clinical Microbiology* **46**, 3201–3207 (2008).
147. Sperandio, B. *et al.* Control of Methionine Synthesis and Uptake by MetR and Homocysteine in Streptococcus mutans. *J. Bacteriol.* **189**, 7032–7044 (2007).
148. Kovaleva, G. Y. & Gelfand, M. S. Transcriptional regulation of the methionine and cysteine transport and metabolism in streptococci. *FEMS Microbiology Letters* **276**, 207–215 (2007).
149. Garriga, X. *et al.* nrdD and nrdG genes are essential for strict anaerobic growth of Escherichia coli. *Biochem. Biophys. Res. Commun.* **229**, 189–192 (1996).
150. Mcallister, L. J., Ogunniyi, A. D., Stroeher, U. H. & Paton, J. C. Contribution of a Genomic Accessory Region Encoding a Putative Cellobiose Phosphotransferase System to Virulence of Streptococcus pneumoniae. *PLoS ONE* **7**, e32385 (2012).
151. Hartel, T. *et al.* Impact of Glutamine Transporters on Pneumococcal Fitness under Infection-Related Conditions. *Infection and Immunity* **79**, 44–58 (2010).
152. Cho, K. H. & Caparon, M. G. tRNA Modification by GidA/MnmE Is Necessary for Streptococcus pyogenes Virulence: a New Strategy To

Make Live Attenuated Strains. *Infection and Immunity* **76**, 3176–3186 (2008).

153. Pestova, E. V. & Morrison, D. A. Isolation and Characterization of Three Streptococcus pneumoniae Transformation-Specific Loci by Use of alacZ Reporter Insertion Vector. *J. Bacteriol.* **180**, 2701–2710 (1998).