Universidade Federal do Rio Grande do Sul
Centro de Biotecnologia
Programa de Pós-Graduação em Biologia Celular e Molecular

Análise evolutiva de processos de redução genômica em bactérias e correlação com aspectos funcionais de *Mycoplasma hyopneumoniae*

Marcos Oliveira de Carvalho

Dissertação submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular da UFRGS como requisito parcial para a obtenção do grau de Mestre

Orientador: Dr. Henrique Bunselmeyer Ferreira

Porto Alegre, Maio de 2008

Este trabalho foi realizado no Laboratório de Genômica Estrutural e Funcional do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (CBIOT-UFRGS), com financiamento da CAPES e do CNPq

À minha mãe, Maria de Fátima, pelo amor incondicional.

Agradecimentos

Agradeço, em primeiro lugar, à minha mãe e à minha irmã, pelo apoio e amor inesgotáveis e, sobretudo, pela compreensão durante a realização desse trabalho. À minha namorada pelo seu amor e dedicação, além de suportar bravamente meses a fio de conversas sobre redução genômica e evolução do parasitismo e, mesmo assim querer continuar namorando comigo. Sem o apoio delas este trabalho não teria sido completado. Agradeço a todos meus tios e tias, que ajudaram das mais diferentes formas desde minha primeira estada em Porto Alegre até agora. Ao meu orientador, Prof. Henrique Ferreira, pela liberdade que me proporcionou para escolher o tema do trabalho e durante a sua condução, o que produziu uma inestimável experiência, da qual pretendo me valer ao longo da minha carreira acadêmica. Seu apoio também foi imprescindível para a realização deste trabalho. À UFRGS, ao PPGBCM e aos seus professores e funcionários, pela oportunidade de participar de um curso de excelência. A CAPES, pelo suporte financeiro na forma de uma bolsa de estudos e ao CNPq, que financiou a infra-estrutura. Aos professores do Grupo de Genômica Funcional e Estrutural do CBIOT, Prof. Arnaldo Zaha, Prof. Augusto Schrank, Profª. Marilene Vainstein, Profª. Irene Schrank e Prof. Sérgio Ceroni. Aos meus colegas de laboratório, pelas discussões frutíferas e pelos ótimos momentos de descontração. Ao fim, agradeço aos três professores que ajudaram a formar minha carreira acadêmica desde a primeira semana em que estive em uma universidade: Profª.Lia Rejane Machado, Profª.Lenira Sepel e Prof. Elgion Loreto.

Índice

Lista de Abreviaturas

C - cytosine (citosina)

G – guanine (guanina)

GC – guanine/ cytosine (citosina/ guanina)

GESPI – *genome strand preference index* (índice genômico de preferência de fita)

CDS – Coding Sequence

PROT* - Porcentagem de identidade de sequência de aminoácidos par-a-par

DNA* - Porcentagem de identidade de sequência de nucleotídeos par-a-par

DNA - Deoxyribonucleic acid (ácido desoxiribonucléico)

Ka – Índice de substituição par-a-par não-sinônima

Ka/Ks – Relação entre índice de substituição par-a-par não-sinônima e índice de substituição par-a-par sinônima

Ks - Índice de substituição par-a-par sinônima

Kb – Quilobase ou 10^3 pares de bases

Mb – Megabase ou 10^6 pares de bases

NCBI - National Center for Biotechnological Information

NJ – Neighbor – Joining

Protein IDs – Identificadores únicos das sequências peptídicas de genomas completos definidos pelo National Center for Biotechnological Information

R^2 – Coeficiente de determinação

RNAP – RNA Polimerase

SCUO - Synonymous Codon Usage Order (Índice usado para determinar numericamente a tendência preferencial de uso de códons)

p – Valor p. Probabilidade de que a amostra possa ser encontrada em uma população sob teste, assumindo-se que a hipótese nula seja verdadeira.

tRNA – RNA transportador

Resumo

Genomas reduzidos de bactérias apresentam-se como complexos sistemas evolutivos, evidenciando diferentes propriedades genômicas de acordo com o nicho ocupado por cada organismo. Em bactérias simbiontes, genomas reduzidos apresentam uma grande estabilidade em relação a rearranjos e taxas mutacionais em comparação com genomas reduzidos de bactérias parasitas. Dentre as espécies parasitas com genoma reduzido, *Mycoplasma hyopneumoniae* distingue-se por apresentar, em diferentes cepas, genomas com características específicas dentro da classe Mollicutes, incluindo maior taxa mutacional, e freqüência de rearranjos.

Através do desenvolvimento de um método quantitativo para avaliar a preferência de disposição preferencial de genes nas fitas senso e anti-senso de genomas de procariotos, foi possível avaliar comparativamente a estrutura genômica de espécies bacterianas em 8 filos de procariotos. Os resultados dessa análise indicam que dentre os genomas analisados as espécies de Mollicutes do grupo hominis são as que possuem a maior desorganização estrutural em relação à colocação preferencial de genes nas fita senso e anti-senso, enquanto que as espécies do grupo pneumoniae apresentam valores intermediários, o que se correlaciona com a maior sintenia observada entre genomas nesse grupo.

Uma análise em grande escala de mutações sinônimas e não-sinônimas em 7 grupos de genomas reduzidos indica a ocorrência de evolução positiva em uma série de genes possivelmente relacionados a patogenicidade. Com base em dados comparativos de rearranjos, presença de recombinação, transferência horizontal e elevadas taxas de mutações sinônimas sobre não-sinônimas, um novo modelo de evolução de genomas reduzidos de parasitas foi proposto. No modelo, sugere-se que, ao contrário dos genomas de bactérias simbiontes, a redução genômica em bactérias parasitas caracteriza-se por um

processo altamente turbulento, no qual a constante geração de variabilidade é um ponto crítico para o sucesso da estratégia de colonização do hospedeiro.

Abstract

Reduced genomes present an interesting evolutionary system, showing different genomic properties accordingly their lifestyle as parasites or symbionts. In symbiont bacteria, reduced genomes present a great stability regarding rearrangements and mutational rates comparing with reduced genomes from parasite bacteria. Among these species, the *Mycoplasma hyopneumoniae* genome distinguish itself by presenting specific characteristics inside the Mollicutes class, including higher mutational and rearrangements rates.

Through the development of a quantitative method of gene strand bias, was possible a comparative analysis of 8 bacterial phyla. The results of such analysis indicated that genomes of Mollicutes bacteria from the hominis group possess the most disordered genomes regarding gene position on the leading or lagging chromosome strand, while genomes from bacteria of the pneumoniae group have intermediary values, in agreement with the higher sinteny between the genomes in this group.

A large-scale analysis of synonymous and non synonymous mutations in 7 groups of reduced genome bacteria determined positive evolution in several genes related to pathogenicity. Based in comparative data from rearrangements, recombination, horizontal transfer and high synonymous over non-synonymous mutational rates, a new model of evolution for parasite species with reduced genomes is proposed. In this model is suggested that, in opposition with symbionts bacteria, genome reduction of parasite prokaryotes is characterized by a high turbulent process, where the constant variability generation is a critical point for the success of the host colonization.

1. Introdução geral

1.1 Mollicutes e estudos genômicos

Coletivamente, os Mollicutes apresentam-se como interessantes modelos para o estudo de processos evolutivos de redução genômica e sua correlação com o parasitismo. Isso se deve à diversidade de hospedeiros que as espécies dessa classe colonizam, à presença de espécies não-parasitas e à disponibilidade da seqüência de genomas completos.

Taxonomicamente, os Mollicutes englobam bactérias que não possuem parede celular, sendo o termo "micoplasmas" aplicado de forma usual para determinar os microorganismos com tal propriedade. No entanto, com o decorrer dos estudos taxonômicos desta classe, foram propostos novos termos para separar os organismos com ausência de parede celular e que possuíam diferentes nichos ecológicos, surgindo as denominações Phytoplasma, para espécies parasitas exclusivas de vegetais e insetos e atribuição de ordem indefinida. Dessa forma a classe Mollicutes é subdividida em 4 ordens, sendo estas: *Mycoplasmatales*, a mais extensa e que engloba os gêneros *Mycoplasma* e *Ureaplasma*; *Entomoplasmatales*, possuindo os gêneros *Entomoplasma* e Mesoplasma; *Spiroplasmataceae*, possuindo apenas o gênero Spiroplasma; Acholeplasmatales, incluindo apenas o gênero *Acholeplasma* e *Anaeroplasmatales* englobando os gêneros *Anaeroplasma* e *Asteroleplasma*.

Atualmente existem mais de 180 espécies de Mollicutes conhecidas, incluindo tanto espécies parasitas, como *Mycoplasma hyopneumoniae*, *Mycoplasma pulmonis* e *Mycoplasma genitalium*, quanto de vida livre, como *Mesoplasma florum*. Mollicutes parasitas colonizam uma ampla gama de hospedeiros, incluindo plantas (Phytoplasmas), peixes (*Mycoplasma mobile*), aves (*Mycoplasma synoviae*) répteis (*Mycoplasma alligatorum*) e mamíferos (*Mycoplasma genitalium*). Dentre os Mollicutes 17 espécies

tiveram genomas seqüenciados, algumas delas para mais de uma linhagem ou isolado. Três destes genomas (de *M. hyopneumoniae* J, *M hyopneumoniae* 7448 e *M. synoviae*) foram seqüenciados por consórcios brasileiros de pesquisa genômica, com a participação de nosso grupo de pesquisa (Vasconcelos *et al*, 2005).

Os Mollicutes têm como sua principal característica a ausência de uma parede celular, genomas com tamanho geralmente inferior a $1x10^6$ nucleotídeos (1 megabase), a necessidade de colesterol para seu crescimento em cultivo e um genoma rico em nucleotídeos AT, distinguindo-se de outros genomas reduzidos pela presença em praticamente todas as espécies da classe, de elementos de transposição (Loreto *et al*, 2007), inclusive grandes elementos como elementos integrativos conjugativos (Pinto *et al*, 2007) e no caso do genoma de *M. hyopneumoniae*, de uma seqüência de inserção ainda pouco caracterizada, denominada tmH1. Esta manutenção de seqüências de inserção é justificada como um componente não-neutro no processo evolutivo dos Mollicutes, sendo responsável pela geração de variabilidade, podendo, portanto, representar um componente adaptativo do genoma (Loreto *et al*, 2007).

Interessantemente, uma extensa região do genoma de *Mycoplasma synoviae* foi identificada como tendo sua origem a partir de transferência horizontal (Vasconcelos *et al*, 2007), o que estabelece que mesmo genomas reduzidos podem eventualmente adquirir novas regiões genômicas e mantê-las, caso elas ofereçam alguma vantagem adaptativa. No caso de *M. synoviae*, a região identificada como transferida horizontalmente é responsável pela codificação de uma classe de proteínas de membrana com provável papel no processo de patogenicidade, comum à *Mycoplasma gallisepticum*, outro mollicute com o mesmo hospedeiro que *M. synoviae*.

1.2 Aspectos evolutivos de genomas bacterianos reduzidos

Dois estudos pioneiros (Kingsbury, 1969; Bak *et al*, 1969) mostraram uma relativa diversidade com relação ao tamanho de vários genomas bacterianos. O trabalho realizado por Bak *et al* (1969), utilizando microscopia eletrônica para determinação do tamanho do genoma em micoplasmas, já apontava a possibilidade de estas bactérias possuírem os menores genomas de procariotos. Os mesmos autores também sugeriram que o fato de terem o genoma distintamente menor do que o das outras bactérias já estudadas na época seria um indicativo de que os micoplasmas integrariam uma classe separada de organismos.

O conceito de que os micoplasmas, como bactérias possuidoras de genomas muito pequenos, participariam de uma linhagem inicial a partir da qual ocorreram sucessivos eventos de expansão genômica predominou até o início da década de 1980, quando um estudo filogenético de espécies de Mollicutes estabeleceu as relações de descendência entre os micoplasmas e espécies da ordem Clostridiales (Woese *et al*, 1980). Tal trabalho sugeriu que os Mollicutes haviam sofrido um processo de redução genômica, ao invés de serem espécies representativas de organismos ancestrais já originalmente portadores de genomas de pequeno tamanho. Esse mesmo estudo já indicava que micoplasmas possuíam uma alta taxa de mutação, o que era evidenciado pela elevada variabilidade observadas entre genes ortólogos de RNA ribossômico 16S. Esta característica, por sua vez, tem importante significado para a biologia dos Mollicutes.

A diversidade de hospedeiros das diferentes espécies de Mollicutes (de plantas a mamíferos) reflete estratégias de extremo sucesso durante o processo de evolução do parasitismo nas espécies dessa classe, principalmente se for considerado o fato de que,

devido à falta de parede celular, as populações de espécies de Mollicutes ficam sujeitas a constantes gargalos populacionais, uma vez que necessitam permanentemente de contato com o hospedeiro para sua dispersão. Estes gargalos populacionais, por sua vez, diminuem a taxa efetiva de seleção sobre uma determinada população, levando a uma maior fixação de mutações deletérias.

O acúmulo de mutações acabaria por levar uma espécie à extinção, em um fenômeno, conhecido como "*Muller´s rachet*", que já foi proposto para as proteobactérias endossimbiontes de insetos, outro grupo de bactérias que apresentam genomas reduzidos (Moran, 1996). As proteobactérias simbiontes de insetos também enfrentariam o mesmo processo de gargalo populacional, devido à natureza intracelular de sua relação com o hospedeiro e à herança materna essencialmente vertical.

Há, contudo, diferenças fundamentais entre os processos evolutivos de genomas reduzidos de bactérias endossimbiontes e parasitas. A primeira diferença diz respeito à estrutura genômica. Bactérias endossimbiontes vêm mantendo seus genomas essencialmente intactos por mais de 50 milhões de anos (Tamas *et al*, 2002), enquanto Mollicutes parasitas apresentam altas taxas de rearranjos e recombinação. Por exemplo, grandes rearranjos foram reportados em genomas de micoplasmas estreitamente relacionados (Vasconcelos *et al*, 2005) e estudos de variação populacional em linhagens de espécies de micoplasmas obtidas em diversas regiões geográficas demonstraram que a variação genômica dessas espécies acontece em nível mutacional e recombinacional (Calus *et al*, 2007 e Mayor, 2008).

Pode-se afirmar que tal nível de variação está relacionado com o nicho adotado pelos micoplasmas como parasitas. Enquanto espécies endossimbiontes desfrutam de um ambiente altamente estável, sendo provisionadas por seu hospedeiro, espécies parasitas

necessitam de contínua adaptação a novos hospedeiros ou a pressões seletivas impostas pelos mecanismos de defesa de seu hospedeiro. Esta disputa adaptativa entre parasitos e parasitados leva a um processo de co-evolução constante, onde as espécies hospedeiras são pressionadas a evoluir sob influência dos parasitos, e os parasitos são pressionados a evoluir pelos mecanismos de defesa de seus hospedeiros (Van Valen, 1973).

Caso haja um desequilíbrio na relação co-evolutiva entre hospedeiro e parasito, pode ocorrer a extinção de uma das espécies. Como exemplo, podemos supor uma relação parasito-hospedeiro na qual o hospedeiro desenvolveu uma rápida capacidade de identificação do parasito no momento da infecção. Esta capacidade adquirida poderia torná-lo resistente à infecção pelo parasito em poucas gerações e forçaria o parasito a trocar de espécie hospedeira ou adaptar-se à nova condição a partir da aquisição, por co-evolução, de um mecanismo capaz de burlar a nova capacidade do hospedeiro. Caso nenhuma destas situações ocorra, a espécie parasita tenderia a ser extinta.

A relação co-evolutiva também poderia ocorrer no sentido inverso. Nesta situação, haveria primeiro o surgimento de uma vantagem adaptativa para o parasito, à qual o hospedeiro deveria responder, co-evolutivamente, desenvolvendo uma adaptação que a suplantasse.

Mudanças bruscas como as supostas acima não são comuns na natureza. Em situações naturais, o processo de relação parasito-hospedeiro desenvolve-se ao longo de milhões de anos. Entretanto, os genomas reduzidos de Mollicutes parasitas constituem um interessante exemplo de como essas adaptações podem ser rápidas em alguns casos, se considerada a escala evolutiva geral. Através do mecanismo de variação de fase, os genomas de Mollicutes parasitas podem rapidamente reestruturar componentes de sua membrana celular que potencialmente seriam identificados pelo sistema imune do

hospedeiro (Iverson-Cabral *et al*, 2006). Ao produzir constantemente indivíduos variantes em relação aos genes de proteínas de membrana em sua população, tais bactérias submetem-se ao processo de seleção darwiniana clássica pelo sistema imune do hospedeiro, sendo que as variantes não identificadas pelo sistema imune rapidamente, aumentam sua freqüência na população e iniciam a colonização do hospedeiro, instaurando o processo infeccioso crônico peculiar às espécies de Mollicutes parasitas.

Portanto, a alta taxa mutacional e a elevada freqüência de rearranjos e de recombinação são aspectos cruciais para o sucesso dos Mollicutes parasitas. Seus genomas reduzidos e, portanto, com repertórios restritos de genes parálogos, que poderiam proporcionar mecanismos de escape, precisam rapidamente gerar variabilidade para contrabalançar pressões seletivas proporcionadas pelos hospedeiros.

É preciso ressaltar, no entanto, que tal processo de geração de variabilidade acontece em nível populacional. Isso significa que muitas das variantes geradas são reconhecidas pelo hospedeiro e eliminadas da população, sendo então a freqüência das variantes não reconhecidas aumentada de acordo com o desvio do equilíbrio de Hardy-Weinberg (Nagylaki, 1976).

1.3 Justificativas e objetivos

O fenômeno da redução genômica em bactérias, apesar de identificada já em meados dos anos 80, só pôde ser estudada com profundidade a partir da disponibilidade de genomas completos de bactérias com genomas reduzidos e de espécies relacionadas com genomas não-reduzidos. A partir de tais dados genômicos, foi possível iniciar o desenvolvimento de correlações evolutivas e traçar modelos para o entendimento dos processos que levaram à redução de genomas em procariotos. No entanto, estes estudos concentraram-se em genomas de endossimbiontes do filo Proteobacteria, sendo as espécies do filo Firmicutes geralmente apenas utilizadas como modelos para a identificação de um genoma mínimo.

Dessa forma, existe uma grande carência de estudos com relação aos processos evolutivos moleculares que moldaram genomas de parasitas, especialmente, de espécies de parasitos do filo Firmicutes.

Além disso, a maioria das bactérias com genomas reduzidos do filo Firmicutes são importantes patógenos de plantas, animais domésticos e seres humanos. Isso justifica adicionalmente a necessidade de um entendimento em profundidade de seus processos evolutivos e das características biológicas daí resultantes, o que pode facilitar o desenvolvimento de estratégias de controle dos patógenos para a preservação de seus hospedeiros naturais.

Dado o exposto este trabalho objetivou:

- Determinar as principais diferenças evolutivas entre genomas reduzidos de espécies

de Mollicutes parasitas e simbiontes

- Identificar genes em parasitos com genomas reduzidos que apresentem relação com patogenicidade, utilizando métodos filogenéticos.

- Estabelecer um modelo evolutivo para espécies de Mollicutes parasitas com genoma reduzido, com ênfase para a explicação das propriedades de genomas de linhagens parasitas de *Mycoplasma hyopneumoniae*.

2. Resultados

2.1 Artigo “de Carvalho MO, Ferreira HB. 2007. Quantitative determination of gene strand bias in prokaryotic genomes. Genomics. 90:733-40.”

2.1.1 Introdução

A disponibilidade de genomas bacterianos completos a partir do seqüenciamento de *Haemophilus influenza* (Fleischmann *et al*, 1995) abriu a possibilidade de interpretação da seqüência de DNA cromossômico de diferentes formas (Grigoriev *et al*, 1998). Diferentes estratégias foram criadas para conversão das seqüências nucleotídicas em outras formas de dados que pudessem ser interpretadas por métodos analíticos numéricos, como *wavelets* (Song *et al*, 2003) e modelos ocultos de Markov (Waack *et al*, 2006).

Tais métodos foram inicialmente utilizados para a identificação de assinaturas características das seqüências genômicas, que poderiam representar blocos funcionais de um cromossomo, como os determinantes de início e final de genes. Adicionalmente, a conversão das propriedades da seqüência de DNA em genomas completos, tais como conteúdo GC e padrões de seqüência revelaram importantes características funcionais dos cromossomos bacterianos como a pronunciada tendência de acúmulo de G sobre T na fita senso (Necşulea & Lobry, 2007).

Estas análises podem ser realizadas na forma de “janela deslizante”, onde a seqüência genômica é analisada de forma contínua ao longo de intervalos pré-definidos do genoma, possibilitando a identificação de regiões diferenciais da seqüência em relação ao genoma completo (Thurman *et al*, 2008).

Este método é empregado com relativo sucesso para identificação de ilhas genômicas que apresentam conteúdo GC com níveis estatisticamente relevantes em relação à média cromossômica, evidenciando transferência horizontal ou restrições\incrementos a desvios mutacionais (Hsiao *et al*, 2003). A abundância relativa de G sobre C tem uma importância prática na análise de genomas completos, uma vez que o cálculo cumulativo do desvio de G sobre C da fita senso em relação à anti-senso pode ser usado para graficamente determinar a origem de replicação em genomas bacterianos (Tillier & Collins, 2004).

No entanto, os genomas bacterianos de forma geral apresentam em primeiro nível uma tendência em relação a disposição dos genes na fita senso. Esta orientação foi primeiramente interpretada como uma forma de evitar erros de replicação que são geralmente associados ao processo descontínuo de replicação da fita anti-senso (Rocha & Danchin, 2001).

Contudo, trabalhos posteriores indicam que a colocação preferencial de genes na fita senso ou anti-senso está relacionada à essencialidade de cada gene (Rocha, 2004). Esta seria uma estratégia empregada pelos genomas bacterianos para evitar que genes essenciais tenham sua expressão prejudicada por colisões entre a maquinaria de replicação e transcrição durante o período de divisão celular, otimizando dessa forma a disponibilidade de transcritos de importância crucial para a célula.

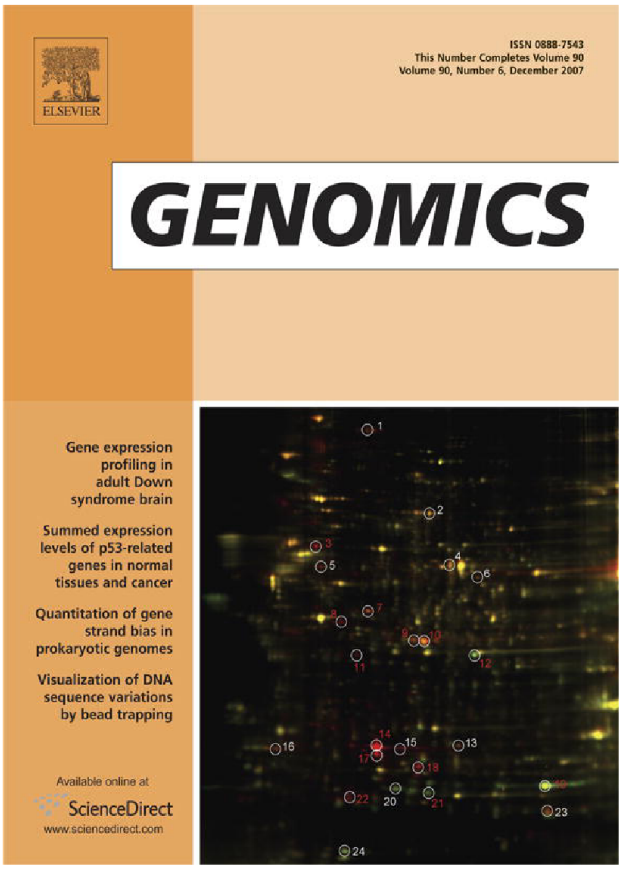
A disposição preferencial de genes na fita senso ou anti-senso caracteriza-se, portanto, como uma importante propriedade estrutural dos genomas bacterianos, com correlação direta entre os processos de transcrição e replicação. Considerando esta importância, um método de quantificação da disposição preferencial de genes estabelece uma referência para a comparação analítica das diferenças na distribuição dos genes ao

longo do cromossomo em genomas de diferentes espécies de bactérias. Esta quantificação pode então ser utilizada para o estabelecimento de relações evolutivas entre genomas

Especificamente para genomas reduzidos de Mollicutes, foi observado durante o trabalho de seqüenciamento e análise do genoma de *M. hyopneumoniae*, que os genomas de espécies parasitas apresentavam um padrão pouco definido de disposição preferencial de genes nas fitas senso e anti-senso, enquanto o genoma do mollicute de vida livre *M. florum* apresentava uma grande estruturação do genoma em relação à disposição dos genes nas fitas senso e anti-senso. Porém, foi verificado que, mesmo entre os micoplasmas parasitas, há uma grande diferença em relação aos níveis de estruturação, sendo os genomas de micoplasmas do grupo Pneumoniae mais estruturados do que os de micoplasmas do grupo Hominis.

Para determinar de forma analítica a preferência de disposição gênica na fita senso ou anti-senso de cromossomos bacterianos, foi então desenvolvido um método numérico de normalização dos dados de colocação gênica. Este método produz como resultado um índice denominado GESPI, que é um indicador absoluto do nível de estruturação cromossômica em relação à disposição gênica nas fitas senso ou anti-senso, podendo ser usado para comparações entre genomas bacterianos de diferentes classificações filogenéticas.

No artigo apresentado a seguir, descrevemos o desenvolvimento deste método quantitativo para determinação da tendência de distribuição de genes na fita senso e anti-senso e a sua aplicação para o estabelecimento de correlações entre as características estruturais de genomas procarióticos.



Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 733–740
GENOMICS
www.elsevier.com/locate/ygeno

Quantitative determination of gene strand bias in prokaryotic genomes

Marcos Oliveira de Carvalho [a,b], Henrique B. Ferreira [a,c,*]

[a] *Laboratório de Genômica Estrutural e Funcional, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, CEP 91591-970, Porto Alegre, RS, Brazil*
[b] *The Bioinformatics Organization*
[c] *Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, CEP 91591-970, Porto Alegre, RS, Brazil*

Received 19 April 2007; accepted 23 July 2007
Available online 24 October 2007

Abstract

Comparative genometrics of microorganisms is a relatively new area, in which genome properties are translated into numerical indexes. Such indexes can be used for a comprehensive and comparative analysis of microbial genomes, contributing to the understanding of their evolution. This work presents a new method for quantitative determination of gene strand bias in prokaryotic chromosomes, in which data transformation of gene position skew leads to a numerical index that can be applied to quantitative comparisons of genome organization. It was applied in the comparative analysis of 49 completely sequenced Firmicutes genomes, allowing the distinction of groups defined according to their patterns of gene strand preference. The resulting groups revealed that, regarding gene strand bias, reduced genomes are, in general, the more disordered among Firmicutes, while genomes of extremophile organisms comprehend those with the highest degree of genome organization in this phylum.

Keywords: Comparative genomics; Gene order; Chromosome organization

Gene distribution and orientation within genomes have important functional implications [1]. One example is the organization of operons, in which contiguous and co-oriented genes are cotranscribed in polycistronic RNAs [2]. The replication process also influences gene organization in the chromosome, as exemplified by the common location of dnaA genes close to the replication origin as a strategy to regulate the cell division process [3]. Another important pattern of gene distribution within genomes is their preferential location in the leading strand of the chromosome, which can be interpreted as evidence of gene essentiality [4].

Some methods to estimate gene strand bias have been described [5,6]. However, these methods do not allow quantitative estimates, and, therefore, more specialized methods are needed to survey the massive amount of genome sequence data currently available. With that in mind, we developed an algorithm for quantitative determination of gene strand bias in prokaryotic chromosomes based on the preference of gene distribution and orientation along the chromosome. This method produces a numeric value that represents the degree of gene strand bias of a genome and can be used for the comparison of different genomes.

Results and discussion

The developed algorithm is applicable to prokaryotic genomes with single replication origins (OriC) and for which the OriC position was either experimentally determined or computationally predicted. Prior to inclusion in the analyzed samples, the OriC position was verified for each genome, based on previously described methods [7]. For calculation purposes, OriC was defined as the first base of the file containing the genome sequence, provided this position corresponded to the functionally or theoretically predicted OriC position. Genome files that were not in agreement with this criterion were adjusted accordingly.

* Corresponding author. Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, CEP 91591-970, Porto Alegre, RS, Brazil. Fax: +55 51 3308 7309.
E-mail address: henrique@cbiot.ufrgs.br (H.B. Ferreira).
URL: http://bioinformatics.org/ (M.O. de Carvalho).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.07.010

The algorithm is essentially based on the cumulative gene strand bias index (C_n), which was calculated as described under Materials and methods. After C_n is calculated for each gene in a genome, starting from OriC, it can be plotted against n or the gene start coordinate, and the resulting plots represent the tendency of gene orientation along the entire chromosome. According to this formula, a portion of a genome with an accumulation of genes in the leading strand will be seen as an increment in C_n values, as genes in the leading strand have an additive effect, while a preponderancy of genes in the lagging strand causes a decrement in C_n. This situation is exemplified well by the *Mesoplasma florum* genome [8], in which the gene strand bias toward the leading strand in the first half of the genome (starting from OriC) is clearly inverted at the terminus of replication (TerC), as can be seen by the change in the graph direction (from an ascending to a descending curve) (Fig. 1).

Although the C_n plot gives a good representation of gene distribution on the two strands of a genome, it is not possible to take C_n absolute values for comparisons between genomes, as the differences in gene number from one genome to another lead to incomparable C_n data, even if the chromosomes have similar gene distributions. Therefore, it was necessary to find a way to normalize the values of gene strand preference among different genomes. The observation that a completely ordered genome, with its first half containing all genes in the leading strand and its second half containing all genes in the lagging strand (Fig. 2A), produces gene strand bias graphs with Pearson's correlation coefficients (r) [9] of 1 and −1 between C_n and n, if the two halves are considered separately (Figs. 2B and 2C, respectively), allowed the use of that statistic descriptor to normalize C_n values among different genomes. Proportionally higher Pearson's coefficients between the set of values of C_n and n would correspond to genome portions with stronger bias toward gene location in the leading strand, while proportionally lower ones are expected to correspond to genome portions with the opposite behavior. The observed curve and coefficient behaviors are the expected ones for most prokaryotic genomes, which typically present genes co-oriented with the direction of replication in each genome half [10]. Deviances from this model occur [11], and for such atypical genomes, comparative analyses based on plots of C_n-derived indexes would not be possible, although such indexes would be still valid to describe the gene strand bias for each individual genome.

A simple form to describe the level of gene strand bias for any prokaryotic genome in a way that would allow quantitative comparisons between genomes would be to calculate the square of the Pearson's correlation coefficient (r^2) between C_n and n for the two halves of the genome and then determine the average between the obtained values. The square calculation was necessary because Pearson's coefficients cannot be used for direct averaging, since their values are not linear functions of the magnitude of the relation between the variables. The r^2 value, however, does not represent all the information on a genome's gene strand preference. For some genomes, genes can be distributed in stretches with alternating gene strand preference along the chromosome. In situations like that, a stretch with gene strand preference opposite to that of its flanking regions corresponds to a genome structure (called here an inversion island) that is likely to have important functional implications, as, for instance, in the cases of pathogenicity islands [12] and clusters of genes coding for restriction-modification systems [13]. A simple way to reflect the presence of inversion islands within a chromosome in a single index is to calculate r^2 in a sliding window fashion, which will produce higher values for more structured regions (i.e., those with more genes positioned in the same strand). By varying the window

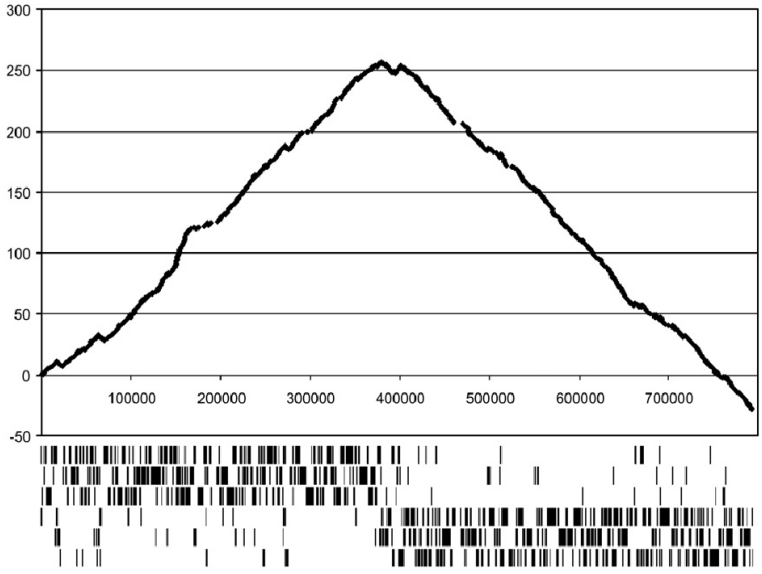


Fig. 1. Cumulative gene strand bias graph of the *M. florum* genome. Dark lines at the bottom represent gene positions along the genome. Numbers on the x axis represent genome base pairs and numbers on the y axis indicate absolute values of the cumulative gene strand bias (C_n).

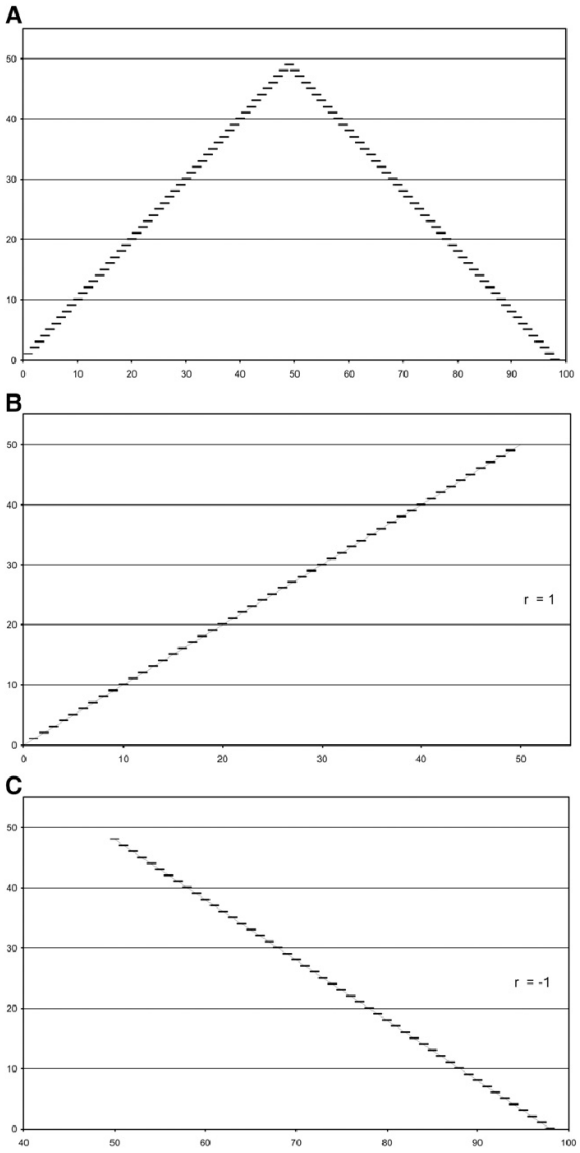


Fig. 2. (A) C_n value plot for a completely ordered conceptual genome. To illustrate the concept of C_n graph partition, the (B) first and (C) second halves of this genome were considered separately, producing gene strand bias graphs with Pearson's correlation coefficients (r) of 1 and −1, respectively.

size, it is possible to obtain different resolutions (levels of sensitivity) for the identification of positions in the genome that deviate from the neighbor current C_n increment or decrement tendency. This generates a gene strand preference index (*gespi*) that provides a quantitative estimate of gene strand bias for any given prokaryotic genome (including the linear ones), provides

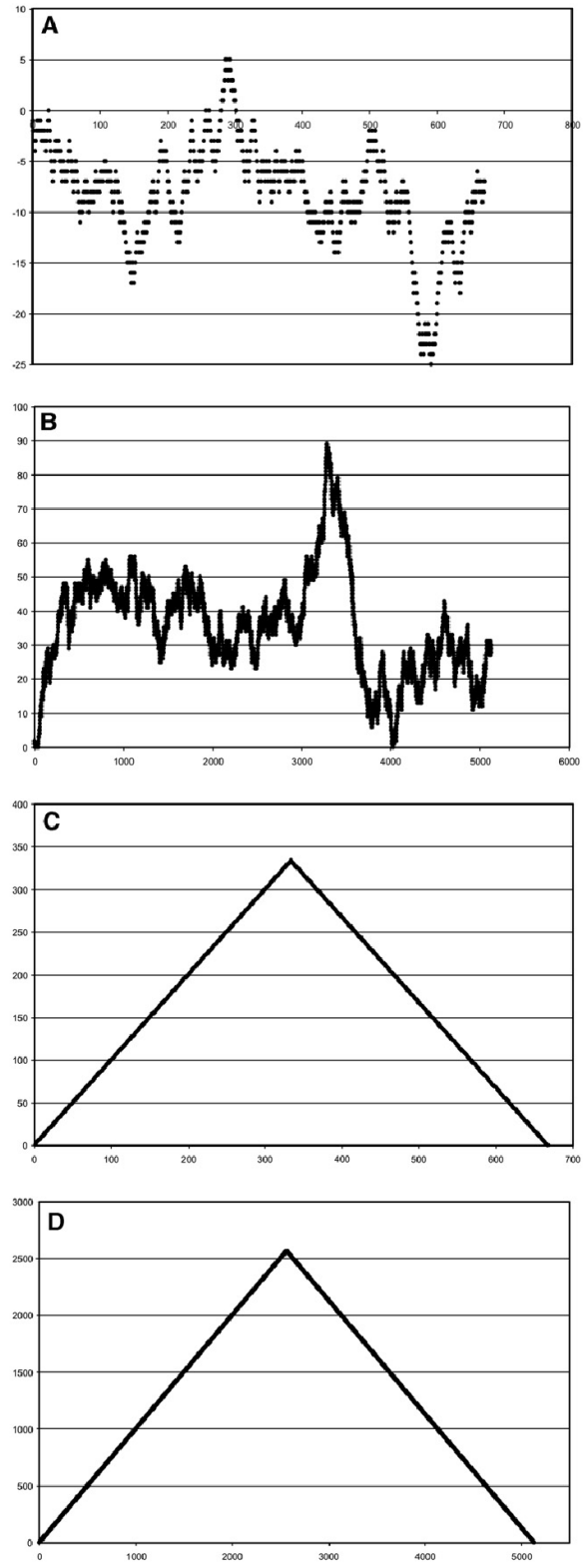
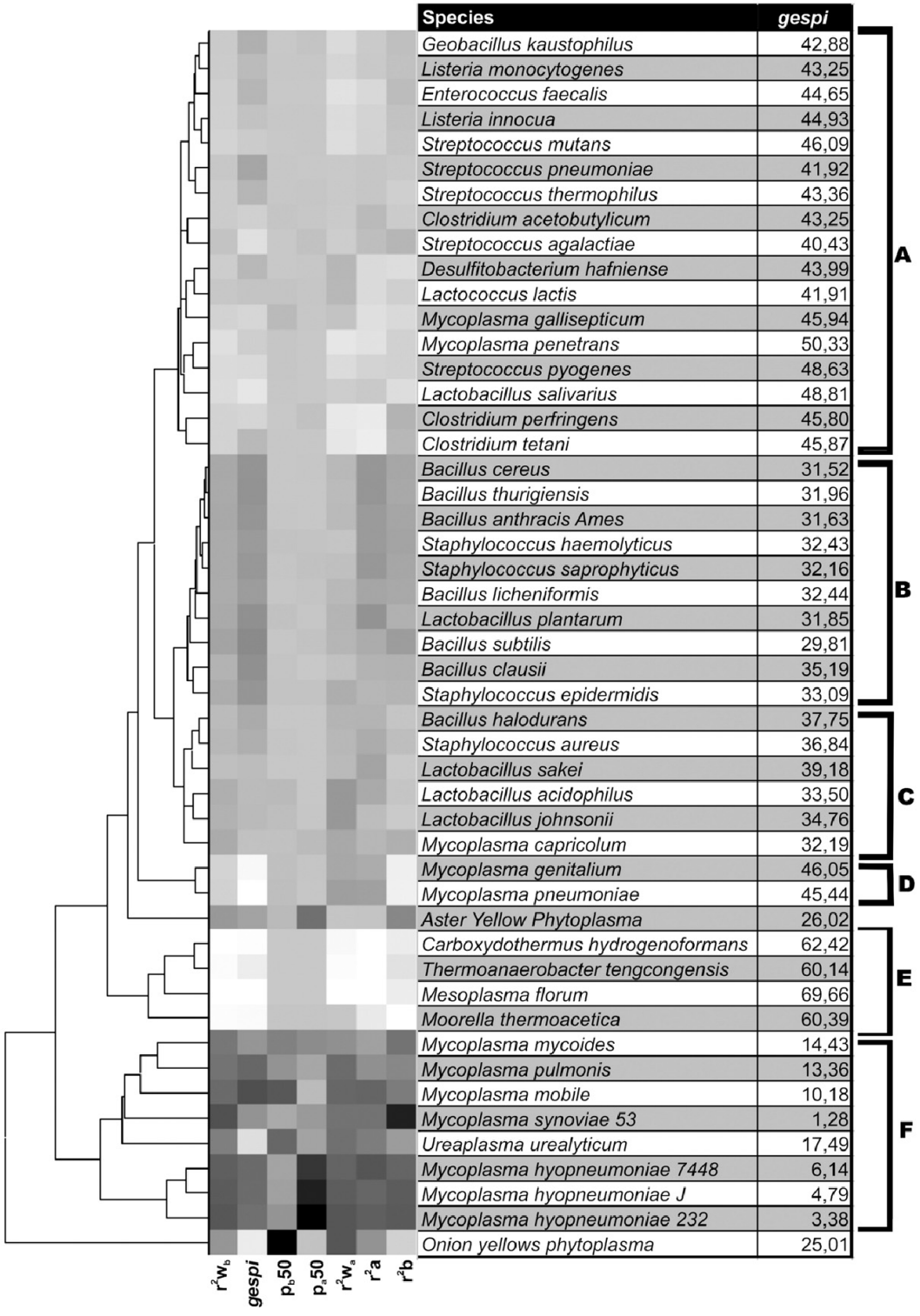


Fig. 3. Simulation of accumulated strand bias graphs for different types of gene strand distribution. (A) and (B) correspond to highly disordered genomes, with gene positions derived from pseudo-random number generation. (C) and (D) correspond to highly ordered genomes, with gene positions determined according to an equitable gene partition between the two chromosome replication halves. (A) and (C) were designed to represent small genomes, comprising 671 genes each, while (B) and (D) represent large genomes, comprising 5134 genes each. Each data point on the x axis corresponds to the position of one gene in the simulated genome, while each y-axis value is the corresponding accumulated gene strand bias.

information on the degree of overall co-orientation between genes and the replication process, and can be compared between different genomes.

To test whether *gespi* values are representative of gene strand bias for both small and large genomes, we simulated data for four hypothetical genomes, two with 671 genes (one completely



ordered regarding gene strand bias and the other with random gene strand preference) and two with 5134 genes (again with one completely ordered regarding gene strand bias and the other with random gene strand preference). As can be seen in Fig. 3, the simulated data confirm that *gespi* is representative of gene strand bias for both small and large genomes, with *gespi* values falling between the predicted boundaries both for completely ordered genomes (*gespi* values of 99.51 and 99.93 for the small and the large genome, respectively) and for those with randomly ordered genes (*gespi* values of 0.0002 and 0.0001 for the small and the large genome, respectively).

Our method was initially validated for the genomic comparative analysis, calculating *gespi* and the corresponding C_n-plotting graphs (Supplementary data, part 1) for 49 Firmicutes genomes. A strong correlation was observed between *gespi* and the C_n graphs, indicating that they are complementary indicators of gene strand bias. Like TS plots [5], C_n plots are visual indicators of large differences in gene strand preference. On the other hand, the *gespi* method, derived from C_n data, is able to detect both small and large differences in gene strand bias between genomes. For instance, C_n plots (or TS plots) of *Bacillus subtilis* and *M. florum* are quite similar, suggesting comparable levels of gene strand preference. However, when the *gespi* method is applied to these genomes, we can demonstrate that small differences in gene strand bias, scattered along the genomes and not detected in the plots, result in an overall quantitative difference of more than 40% (*gespi* values of 29.81, for *B. subtilis,* and 69.66, for *M. florum*).

It was proposed that gene strand bias is more evident in Firmicutes than in other bacteria groups [14]. However, our *gespi*-based analysis showed that, in the Mollicutes class, the members of the Pulmonis group show a nearly random gene distribution. For instance, strains J, 232, and 7448 of *Mycoplasma hyopneumoniae* [15,16] and *Mycoplasma synoviae* [15] present highly disordered genomes, with *gespi* values of 4.79, 3.38, 6.14, and 1.28, respectively These are in contrast with the highly ordered genomes of the non-Pulmonis mollicute *M. florum* (*gespi*=69.66) and other Firmicutes (*gespi* values between 29.81 and 62.42). This contrast may be correlated with an evolutionary change occurred early in the Pulmonis group. It has been suggested that the use of alternative DNA polymerase subunits (DnaE or PolC) in the replication of each strand in Firmicutes is responsible, at least in part, by the generation of more pronounced gene strand bias [14]. However, both dnaE and polC genes are present in all Pulmonis group genomes, which excludes a loss of one of these genes during the process of gene reduction as a simple explanation for the observed low degree of gene strand bias. Alternatively, we hypothesize that Pulmonis group species have a higher rate of genomic recombination and rearrangement than other Mollicutes. According to that, we found that the level of synteny, which is reduced by recombination and rearrangement events, is higher in non-Pulmonis genomes, such as those of the Pneumoniae group, comprising *Mycoplasma genitalium* [17] and *M. pneumoniae* [18], which have corresponding higher *gespi* indexes (46.05 and 45.44, respectively).

The *gespi* component values can also be informative, since they can be specifically analyzed to identify tendencies in gene distribution in the whole chromosome. Therefore, they are useful for comparative analysis, such as hierarchical clustering, from which patterns of genome structuring can be inferred. To demonstrate that, 49 Firmicutes genomes were grouped according to their patterns of gene strand distribution based on *gespi* and its components (Fig. 4). From the resulting clustering, it was possible to delineate different groups of genomes according to their values of *gespi,* defined as groups A to F in Fig. 4. For instance, the most disordered genomes (group F), comprehending Firmicutes subjected to genome reduction, were consistently grouped, with few exceptions. Additionally, it was possible to define a group of highly ordered genomes (group E), essentially formed by extremophile organisms, which is suggestive of a possible evolutionary relationship between gene strand bias and adaptation to extreme environments and remains to be investigated.

Discrepancies in gene strand preferences evidenced by *gespi* analysis can be further analyzed using complementary bioinformatics approaches. For instance, approximately the same *gespi* index was calculated for the two Phytoplasma species analyzed, aster yellow phytoplasma [19] (AYWP) and onion yellow phytoplasma [20] (OYP), but they did not group in the hierarchical cluster analysis, evidencing differences in their genome structures regarding gene strand bias. While the AYWP first replichore is highly organized, with genes mostly positioned in the leading strand, the OYP chromosome presents an opposite architecture, with the second replichore holding most genes in the leading strand. The alignment of the AYWP and OYP genomes (Fig. 5) confirms that, indeed, there are different architectures in these two phytoplasma chromosomes, which are due to a large inversion of the structured region, although the OriC and TerC regions of both chromosomes were kept syntenic.

It can be argued that the simple ratio between the number of genes in the leading strand and the total number of genes would be a good parameter to evaluate gene strand bias. However, our approach allows a more comprehensive view of gene location, taking into account both the overall gene distribution in the two strands and the occurrence of local structural patterns. Although

Fig. 4. Hierarchical clustering results using both the calculated *gespi* and its components as variables. Darker tones indicate lower values. The origin of replication was assumed to start at the first nucleotide position of the genome, as is usual for annotated genomes deposited in the NCBI database. Defined groups are identified by letters on the right. Groups A and B are constituted mainly of Bacilli bacteria, with the exception of Clostridia species in group A. Group A species have intermediate *gespi* values (45.80 to 50.33), while group B have low–intermediate *gespi* values (29.81 to 35.19). Groups E and F correspond to the extremes of *gespi* distribution in the Firmicutes phylum, with group E comprehending the most ordered genomes (*gespi* values from 60.14 to 69.66) and group F, the most disordered ones (*gespi* values from 1.28 to 17.49). Group E is essentially constituted of extremophile organisms (with the sole exception of *M. florum*), while group F is constituted mainly of Mollicutes species from the Pulmonis group.

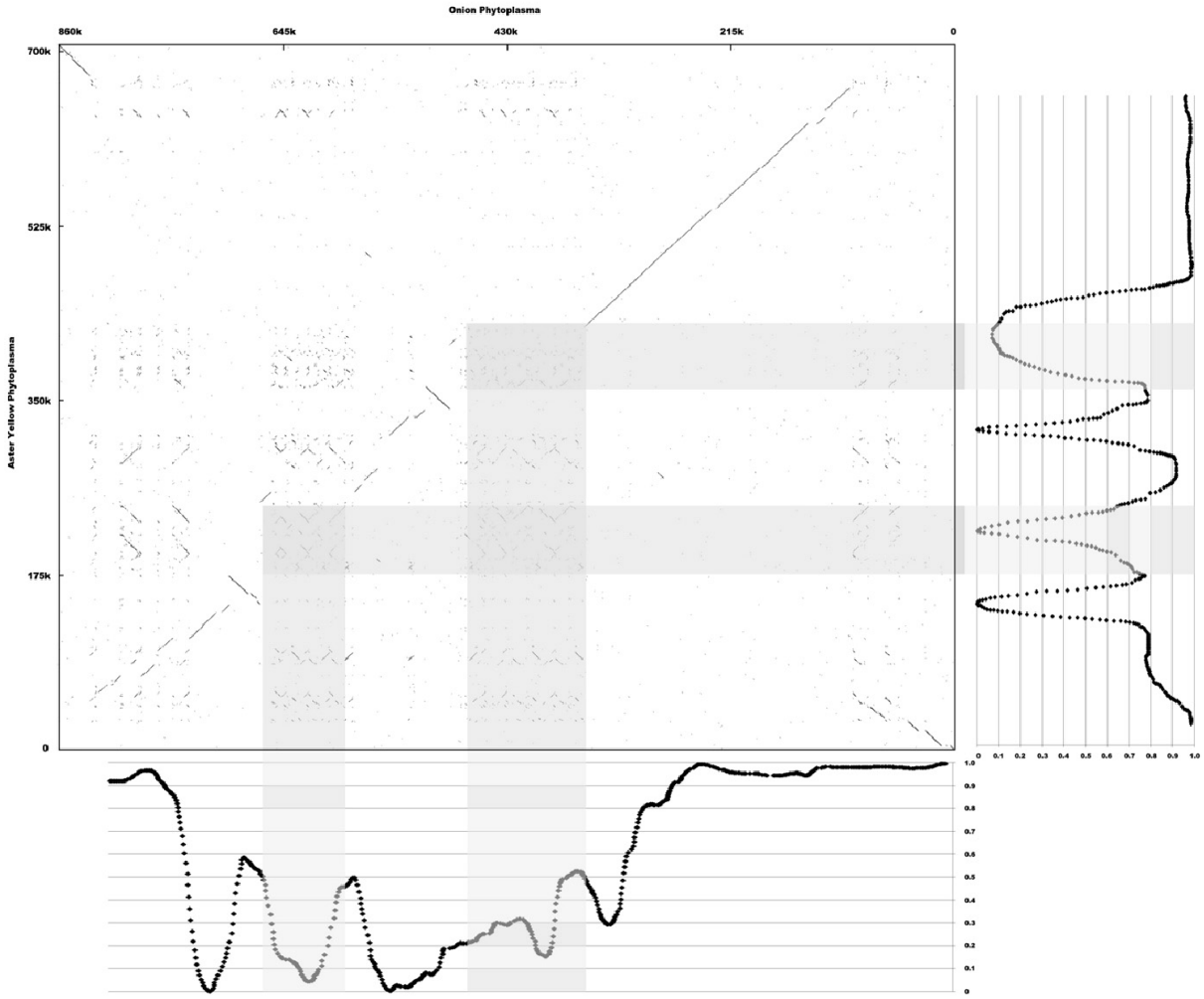


Fig. 5. Dot-plot alignment of the aster yellow phytoplasma (*y* axis) and onion yellow phytoplasma (*x* axis) genomes, and the corresponding graphs of r^2_w values for the whole chromosomes. Shaded areas correspond to rearranged genome regions.

one of the *gespi* components is the percentage of genes that are in the leading strand, this value is considered only after the value of 0.5, corresponding to the expected gene fraction in a strand in a case of random distribution (50%), is subtracted from it. A simple example of the advantage of using *gespi* instead of the simple ratio between genes in the leading strand and the total number of genes would be any case in which we have a clear preponderance of genes in one strand of a replichore, while the genes in the other replichore are randomly distributed between the two strands. Situations like that happen, for instance, in the OYP and AYWP genomes, in which gene distributions in the lagging and leading strands are not equitable in the two replichores. Using the *gespi* approach, these irregular gene distributions between replichores can be both detected and measured, resulting in corresponding lower index values (25.01, for OYP, and 26.02, for AYWP) in comparison to those obtained for genomes with nearly equal gene distributions (*gespi* values around 45.00).

Finally, to demonstrate that *gespi* is applicable to broader phylogenetic surveys, we performed a *gespi* analysis for 35 additional genomes, including representative samples from 12 Eubacteria phyla outside Firmicutes (Supplementary data, part 2). This allowed us to establish interesting evolutionary correlations based on gene strand bias. For instance, we found that extremophile genomes appear among the most organized genomes (higher gene strand bias) in the Deltaproteobacteria (e.g., *Geobacter sulfurreducens,* with *gespi*=22.08) and Firmicute (e.g., *Carboxydothermus hydrogenoformans,* with *gespi*=62.42) groups. This is in agreement with a recent study [21] that showed that co-orientation of transcription and replication is due to selective pressure for processive, efficient, and accurate replication, considering that an organized pattern of gene orientation in relation to OriC would help extremophile organisms to overcome difficulties in transcription and replication under harsh environmental conditions. However, extremophiles from the Thermotogae and Deinococcus–Thermus group are far less organized regarding gene strand bias (with *gespi* values of 5.56 and 1.82 for *Thermotoga maritima* and *Thermus thermophilus,* respectively), suggesting that within this group alternative biological mechanisms for DNA and RNA synthesis coordination may be in operation, deserving specific investigation.

Previous studies [5] also addressed the problem of quantitative measurement of gene strand preference, but in a simplified form and with less resolution, in which the cumulative index of gene strand preference (TS) corresponded to the percentage of genes in a given strand (leading or lagging) in a window of several genes. Although our method also uses a sliding window strategy, C_n, its main data component is obtained using each gene position as a data point, allowing for a more discriminatory analysis of gene strand bias.

Overall, our results suggest that *gespi,* along with its components, is a good numeric indicator of gene strand preference and its application in comparative genometrics is an interesting approach for a preliminary search of patterns of chromosome structuring. Since the method is not very demanding in terms of hardware, it is well suited to performing extensive genome surveys in any laboratory with access to basic computational resources. Future *gespi* analyses are expected to help us understand the evolutionary mechanisms that drive the unequal gene distribution between the two chromosome strands, as well as the overall relationships between patterns of genome organization and lifestyles for different organisms.

Materials and methods

To calculate the cumulative gene strand bias index (C_n), let n be the gene order number, counted from OriC, and s_i be the index designating its location in the leading or lagging strand, with 1 corresponding to a gene in the leading strand and −1 to a gene in the lagging strand. C_n is then expressed as $C_n = C_{n-1} + s_i$, with $C_1 = s_{i1}$.

The calculation of *gespi* for a given genome is done according the following assumptions: N being the total number of genes of the genome, let N_h be the order number of the nearest gene to the chromosome's $S/2$ coordinate, where S corresponds to the last nucleotide in the genome file and, therefore, is the genome size in nucleotides. Let r^2_a be the square of the Pearson's correlation coefficient between (n_1/N_h) and (C_1/C_{Nh}), corresponding to the genes in the first half of the genome, and r^2_b the square of the Pearson's correlation coefficient between (n_{Nh+1}/N) and (C_{Nh+1}/C_N), corresponding to the genes in the second half of the genome. With r^2_w being the square of the Pearson's correlation coefficient for each window (w) along the genome, the average of the r^2_w values can be calculated for each half of the genome, with the first half being $r^2 w_a$ and the second half being $r^2 w_b$. Although r^2 datasets from different genome segments are not really additive to represent an r^2 value for the whole genome, it is valid to calculate the $r^2 w$ averages for each genome segment, which are necessary for the identification of regions with different gene strand biases within a genome.

It was defined that p_a and p_b represent the gene fractions that are co-oriented with replication direction in the first and second genome halves, respectively. Assuming that a random gene distribution would result in 50% of them in each strand, any value above 0.5 for p_a and p_b represents an additive bias toward gene strand preference, suggesting that selection is acting on the gene distribution. This defined $p_a 50 p_a - 0.5$ and $p_b 50 p_b - 0.5$.

The *gespi* was then determined as $gespi = \{[(r^2 a \times r^2 w_a \times p_a 50) + (r^2 b \times r^2 w_b \times p_b 50)] \times 100\}$. This index has values between 0 and 100, with 0 representing a complete absence of gene strand bias and 100 representing a genome completely structured regarding gene strand preference.

Clustering of *gespi* values and their components for the analyzed genomes was performed with the HCE3 software (http://www.cs.umd.edu/hcil/hce/hce3.html), using the UPGMA linkage method and the Euclidean method for distance measurement. The dot-plot alignment was constructed with the lbdot software (http://www.lynnon.com/dotplot/files.html) using recommended configuration values. Genome sequences were obtained from the NCBI FTP site (ftp://ncbi.nlm.nih.gov) on August 10, 2006 (Accession Nos. NC_007716, NC_003997, NC_006274, NC_006582, NC_002570, NC_006322, NC_000964, NC_005957, NC_007503, NC_003366, NC_004557, NC_007907, NC_004668, NC_006510, NC_006814, NC_005362, NC_004567, NC_007576, NC_007929, NC_002662, NC_003212, NC_003210, NC_006055, NC_007644, NC_007633, NC_004829, NC_000908, NC_006360, NC_007332, NC_007295, NC_006908, NC_005364, NC_004432, NC_000912, NC_002771, NC_007294, NC_005303, NC_002951, NC_004461, NC_007168, NC_007350, NC_004116, NC_004350, NC_003098, NC_004606, NC_006449, NC_003869, NC_002162, NC_008752.1, NC_003228.3, NC_002929.2, NC_004463.1, NC_004545.1, NC_002620.2, NC_007899.1, NC_005085.1, NC_009342.1, NC_002971.3, NC_008025.1, NC_007354.1, NC_002939.4, NC_008571.1, NC_000915.1, NC_006369.1, NC_008576.1, NC_002678.2, NC_002977.6, NC_002677.1, NC_002755.2, NC_008767.1, NC_008820.1, NC_008027.1, NC_007969.1, NC_007778.1, NC_007940.1, NC_003197.1, NC_007606.1, NC_007513.1, NC_000853.1, NC_006461.1, NC_002967.9, NC_002978.6, NC_004088.1).

Acknowledgments

We thank Arnaldo Zaha (Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul), Sérgio Ceroni da Silva (Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul), and Cristiano Valim Bizarro (Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul) for critically reading the manuscript. M.O.C. is a recipient of a CAPES M.Sc. fellowship. This work was supported by grants from MCT/CNPq.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.07.010.

References

[1] J.G Lawrence, Gene organization: selection, selfishness, and serendipity, Annu. Rev. Microbiol. 57 (2003) 419–440.
[2] P.P. Dennis, M. Ehrenberg, H. Bremer, Control of rRNA synthesis in Escherichia coli: a systems biology approach, Microbiol. Mol. Biol. Rev. 4 (2004) 639–668.
[3] J. Kato, Regulatory network of the initiation of chromosomal replication in Escherichia coli, Crit. Rev. Biochem. Mol. Biol. 6 (2005) 331–342.
[4] E.P. Rocha, A. Danchin, Gene essentiality determines chromosome organization in bacteria, Nucleic Acids Res. 22 (2003) 6570–6577.
[5] P. Lopez, H. Philippe, Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation, C. R. Acad. Sci. Ser. III Sci. Vie 324 (2001) 201–208.
[6] W. Li, Statistical properties of open reading frames in complete genome sequences, Comp. Chem. 23 (1999) 230–283.
[7] C. Nikolaou, Y. Almirantis, A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species, Nucleic Acids Res. 33 (2005) 6816–6822.
[8] *Mesoplasma florum* L1, complete genome. Direct submission to NCBI. Genome RefSeq: NC_006055.
[9] N.C. Smeeton, P. Sprent, Applied Nonparametric Statistical Methods, 3rd ed. CRC Press, Boca Raton, FL, 2000.
[10] S.D. Bentley, J. Parkhill, Comparative genomic structure of prokaryotes, Annu. Rev. Genet. 38 (2004) 771–792.
[11] S. Casjens, The diverse and dynamic structure of bacterial genomes, Annu. Rev. Genet. 32 (1998) 339–377.
[12] S.H. Yoon, C.G. Hur, H.Y. Kang, Y.H Kim, T.K Oh, J.F. Kim, A computational approach for identifying pathogenicity islands in prokaryotic genomes, BMC Bioinformatics 6 (2005) 184–195.

[13] F. Collyn, L. Guy, M. Marceau, M. Simonet, C.A. Roten, Describing ancient horizontal gene transfers at the nucleotide and gene levels by comparative pathogenicity island genometrics, Bioinformatics 9 (2006) 1072–1079.
[14] E.P. Rocha, Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol. 9 (2002) 393–395.
[15] A.T. Vasconcelos, H.B. Ferreira, C.V. Bizarro, et al., Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae, J. Bacteriol. 16 (2005) 5568–5577.
[16] F.C. Minion, E.J. Lefkowitz, M.L. Madsen, B.J. Cleary, S.M. Swartzell, G.G. Mahairas, The genome sequence of Mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis, J. Bacteriol. 21 (2004) 7123–7133.
[17] C.M. Fraser, J.D. Gocayne, O. White, et al., The minimal gene complement of Mycoplasma genitalium, Science 5235 (1995) 397–403.
[18] R. Himmelreich, H. Hilbert, H. Plagens, E. Pirkl, B.C. Li, R. Herrmann, Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae, Nucleic Acids Res. 22 (1996) 4420–4449.
[19] X. Bai, J. Zhang, A. Ewing, et al., Living with genome instability: the adaptation of phytoplasmas to diverse environments of their insect and plant hosts, J. Bacteriol. 10 (2006) 3682–3696.
[20] K. Oshima, S. Kakizawa, H. Nishigawa, Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma, Nat. Genet. 1 (2004) 27–29.
[21] J.D. Wang, M.B. Berkmen, A.D. Grossman, Genome-wide coorientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis, Proc. Natl. Acad. Sci. USA 104 (2007) 5608–5613.

2.2 Manuscrito "de Carvalho MO, Ferreira HB. 2008. Different evolutionary scenarios for parasites and symbionts reduced genomes" (a ser submetido à Molecular Biology and Evolution)

2.2.1 Introdução

Os resultados apresentados no artigo que constituiu a seção 2.1 demonstraram que os genomas de espécies de Mollicutes apresentam uma grande variabilidade em relação à tendência de posicionamento de genes na fita senso ou anti-senso. Para determinar a existência de padrões relacionados às variáveis constituintes do GESPI, foi realizada uma análise de agrupamento hierárquico.

Esta análise demonstrou um interessante padrão na distribuição dos valores de cada variável de acordo com a distribuição filogenética dos micoplasmas, sendo o grupo hominis, onde se encontra *Mycoplasma hyopneumoniae*, o que apresenta a menor média de organização estrutural do genoma em relação à disposição dos genes nas fitas senso e anti-senso.

A organização dos genes ao longo do cromossomo em bactérias é bastante relevante, uma vez que genes considerados essenciais são mais observados na fita senso, a fim de evitar colisões entre a maquinaria de replicação e de transcrição e, assim, diminuir o número de transcritos abortados (Rocha & Danchin, 2003).

Neste aspecto, os genomas de diferentes linhagens de *M. hyopneumoniae* constituiriam um paradoxo, uma vez que um mecanismo que supostamente aumentaria a adaptabilidade da espécie, gerando diversidade através de rearranjos e recombinação, criaria, ao mesmo tempo, dificuldades para o mecanismo de replicação e para a expressão

de genes essenciais. No caso de uma bactéria parasita, a falha em coordenar a replicação e a expressão de genes importantes para o processo de infecção poderia reduzir a probabilidade de sucesso na colonização do hospedeiro, o que poderia resultar em extinção.

Mecanismos evolutivos compensatórios já foram descritos para genomas reduzidos, como o de *Buchnera aphidicola*, onde o gene groEL, apresenta forte evolução negativa em relação ao restante do genoma. Esta taxa evolutiva diferencial do gene groEL está dissociada do fenômeno de desvio mutacional ocasionado pelos gargalos populacionais aos quais a bactéria endossimbionte *B. aphidicola* está sujeita (Herbeck *et al*, 2003).

Este desvio mutacional induz a fixação de mutações ligeiramente deletérias em populações submetidas a gargalos populacionais, que, por sua vez, promovem uma desestabilização estrutural das proteínas codificadas pela espécie em questão (van Ham *et al*, 2003). A manutenção de uma alta seleção purificadora no gene groEL, cujo produto tem por função prover assistência ao correto dobramento de proteínas recém traduzidas (Li *et al*, 2008), evidencia um mecanismo compensatório do genoma de *B. aphidicola* frente à fixação de mutações desestabilizadoras induzidas pelas características populacionais da sua espécie.

O trabalho apresentado a seguir procura estabelecer relações entre os padrões de evolução positiva/negativa encontrados em genomas reduzidos de espécies bacterianas parasitas e simbiontes com características funcionais da espécie. Para tal, utiliza de forma inédita uma análise em larga escala de valores de substituição sinônima (Ks), não-sinônima (Ka) e sua proporção (Ka/Ks) analisados através da técnica de agrupamento hierárquico. Tal metodologia permite distinguir grupos de genes com características evolutivas

semelhantes e que, portanto, devem estar sujeitos as mesmas forças evolutivas.

Different evolutionary scenarios for parasites and symbionts reduced genomes

Marcos Oliveira de Carvalho[1,3], Henrique B. Ferreira[1,2,*]

[1] Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[2]Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[3]The Bioinformatics Organization, Bioinformatics.Org

Running title: Parasite genome reduction

Key words: Comparative genomics, Chromosome organization, Genome reduction, Genome evolution

Correspondence:

Centro de Biotecnologia - Depto. Biologia Molecular e Biotecnologia.

Av. Bento Gonçalves, 9500, Prédio 43421

CEP 91591-970, Fax: (51) 33167309

Porto Alegre, RS.

e-mail: henrique@cbiot.ufrgs.br

* Corresponding Author.

Abstract

Genome reduction, the shrinkage of chromosomal size of a given species over time from an ancestor with a larger genome, is a complex phenomenon, which occurs mainly in endosymbiont or parasite bacteria from different classes. Few studies have compared so far the evolution of reduced genomes from parasite bacteria to that of reduced genomes from symbiont and free-living bacteria. Such comparative studies are necessary to provide information on processes involved in the shaping of reduced bacterial genomes and to identify adaptative patterns that emerge to allow the establishment of symbiotic or parasite life styles. To identify adaptative patterns in reduced genomes, a large-scale phylogenomic comparison was done in genomes of both symbionts and parasite bacteria and in bacterial genomes believed to be currently under reduction. The analyses were based on synonymous and non-synonymous mutation (Ka/Ks) ratios and genomic architecture, including repeats, duplications, and rearrangements. The analysis of Ka/Ks ratios revealed that genomes from parasite bacteria have higher Ka/Ks ratios in comparison to those observed for symbionts, and that they retain genes under high purifying evolution. From the analyzed species, genes found to be under positive evolution are most from membrane proteins in Mycoplasmas and Chlamydia, with also a number of genes related to pathogenicity such as incA on Chlamydiae genomes and adk on Mycoplasmas. A model of differential evolution for parasite and symbiont bacteria with reduced genomes is proposed.

Introduction

Reduced prokaryotic genomes are subject of much interest as they pose an intriguing evolutionary puzzle, for being adaptable to either free-living, parasite or symbiont life styles. Interestingly, prokaryotic genomes share common features irrespective to bacterial life styles, such as a low GC content (Khachane *et al*, 2007; Goto *et al* 2000), loss of DNA repair pathways (Klasson and Andersson, 2006; Carvalho *et al*, 2005) and obfuscated replication start origin (Mackiewicz *et al*, 2004). On the other hand, reduced genomes from bacterial species with different life styles have prominently differences in repeat content (Rocha, 2003), horizontal transfer (Silva *et al*, 2003), genome rearrangements (Silva *et al*, 2003) and overall gene strand bias (de Carvalho and Ferreira, 2007). The observed similarities among reduced genomes from free-living, parasite or symbiont species may be inherent to the process of genome reduction they all have undergone, but the differences are likely to be results from adaptation processes to a particular lifestyle, especially in cases of symbiont or parasite bacteria.

Genomes from symbiont bacteria present an enduring evolutionary stability, with virtually no rearrangements over a long time period (Tamas *et al*, 2002) and congruence between their evolutionary trees and the evolutionary trees from their hosts (Conord *et al*, 2008). Such stability, however, contrasts with an increased accumulation of mildly deleterious mutations (Moran, 1996), and continuing genome degradation (Pérez-Brocal, 2006). Genomes from parasite bacteria, on the other hand, which are constantly subject to strong selective pressures from the host environment, evolved complex mechanisms of diversity generation,

involving the use of repeats (Rocha and Blanchard, 2002), rearrangements (Vasconcelos *et al*, 2005) and recombination (Mayor *et al*, 2008). Considering the presently different genome features and possible evolutionary changes related to the lifestyle of each species, it is reasonable to assume distinct evolutionary scenarios for genomes from symbionts and parasites. To address this hypothesis, the evolutionary changes on synonymous and non-synonymous sites in small genomes from symbiont and parasite species were examined, with emphasis in the analysis of parasite Mollicutes. A model for the evolution of reduced genomes from parasite and symbiont bacteria was proposed based on the findings of Ka/Ks analysis and genome organization.

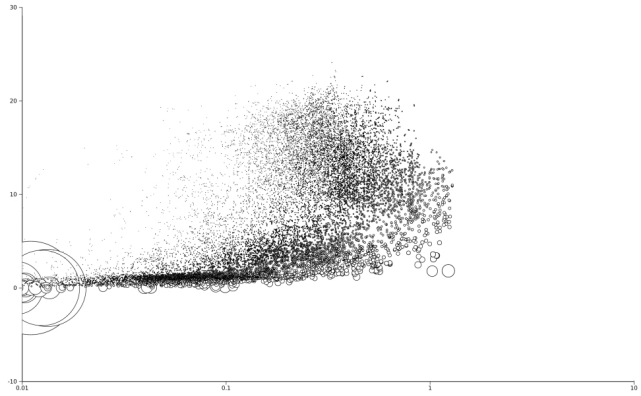
Materials and Methods

Ortholog gene detection was done with the OrthoMCL (Li *et al*, 2003) software, using the complete proteome sequence for each analyzed genome. Supplemental Material contains the list of genomes used in this work, as well their taxonomic classification. Genomes were considered under reduction if their pseudogene count were above six percent of their total gene number and they do not have a genome smaller than 1.5 Mb. Ortholog sequences were codon aligned with the help of a Bioperl (Stajich *et al*, 2002) script, being subsequently submitted to analysis by the codeml software from de PAML package (Yang, 2007). Protein IDs from the determined orthologs were mapped on their respective DNA sequences with custom scripts. TRF (Benson, 1999) was used to determine tandem repeats, and RepeatScout (Price *et al*, 2007) was used for the identification of direct and inverse repeats. To determine rearrangement distances, the genomes were first aligned with MUMMER (Delcher *et al*, 2002), and the corresponding breakpoint coordinates were transformed in ordered data, which were analyzed with the GRAPPA software (Moret *et al*, 2001). GESPI gene strand bias analysis was carried out as described elsewhere (de Carvalho and Ferreira, 2007). The phylogenetic tree for species used in this study was constructed based on 16S genes with MEGA (Tamura *et al*, 2007). Twenty nine outgroups, listed on Supplemental Material S2, were used. The tree, available as Supplemental Material S3, was reconstructed using the Neighbor-Joining method and Kimura 2-parameters for distance calculations with 1000 bootstrap replications for accessing its consistency. Hierarchical cluster analysis was performed using the HCE3 software (Seo *et al*, 2006), using Euclidean distance and total linkage for the

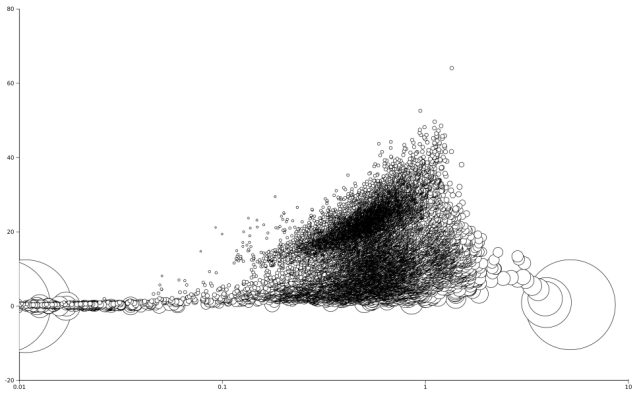
calculation of evolutionary relationship between ortholog gene pairs. Statistical tests were conducted with the R statistical language (Ihaka and Gentleman, 1996).

Results and Discussion

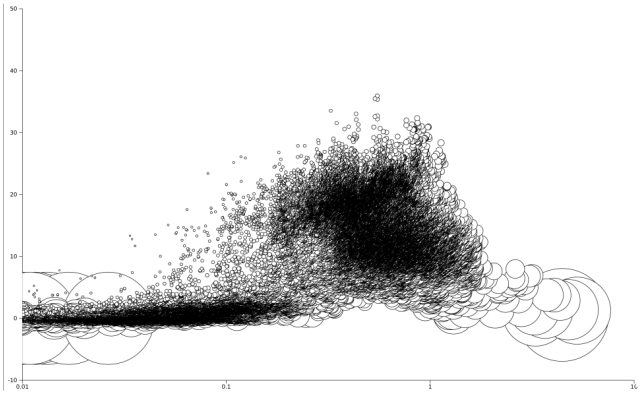
The analysis of Ka/Ks substitutions performed for 66,523 genes from 68 genomes (Supplemental Material S4) yielded 118,170 ortholog gene pair comparisons, which were subsequently submitted to statistical analysis (Supplemental Material S5). Each genome in the chosen dataset was assigned to a group according to its phylogenetic position, lifestyle, and classification as reduced or under reduction. A total of seven groups (assigned in Supplemental Material S4) were established, namely "Firmicutes|reduced genome|parasite/non-parasite"; "Chlamidiae|reduced genome|parasite"; "Proteobacteria|reduced genome|parasite"; "Proteobacteria| reduced genome|non-parasite", "Spirochaetes|reduced genome|parasite", "Proteobacteria|under reduction genome|parasite"; and "Actinobacteria|under reduction genome|parasite". Ortholog detection was performed according to this grouping scheme, as well as all subsequent Ka/Ks analyses. Figure 1 shows a bubbleplot of the seven genome groups based on the Ka/Ks analysis, where each data point corresponds to an ortholog pair, the X and Y axes represent Ka and Ks values, respectively, and the bubble diameter represents the Ka/Ks ratio. This analysis demonstrated that most of the genes are not under positive selection, both for reduced and under reduction genomes. However, genomes from parasite species presented significantly (t-test, $p<0.005$) higher Ka/Ks values (Supplemental Material S6) than not-parasite ones. Comparing intracellular parasite species with endosymbionts, the parasite species with small genomes have a higher Ka/Ks mean than intracellular symbionts.



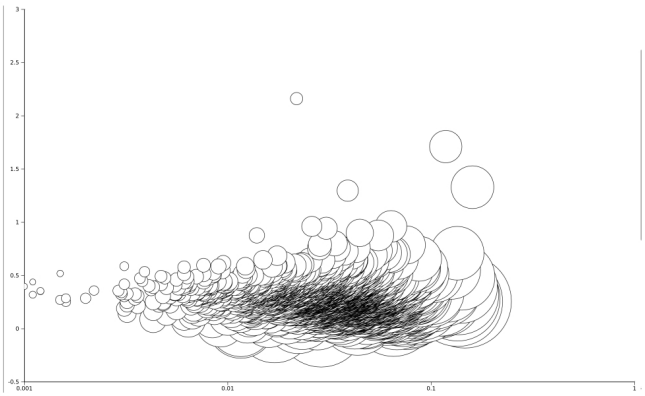
A



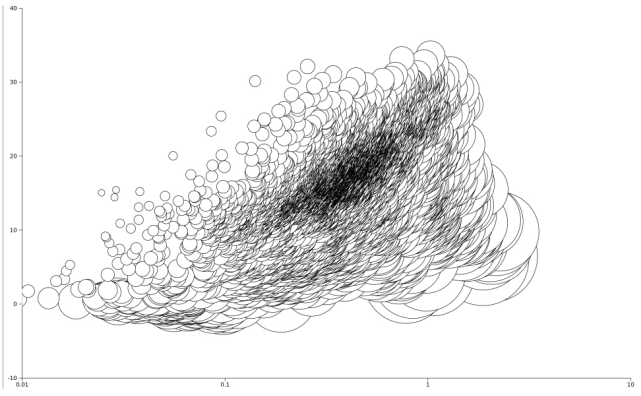
B



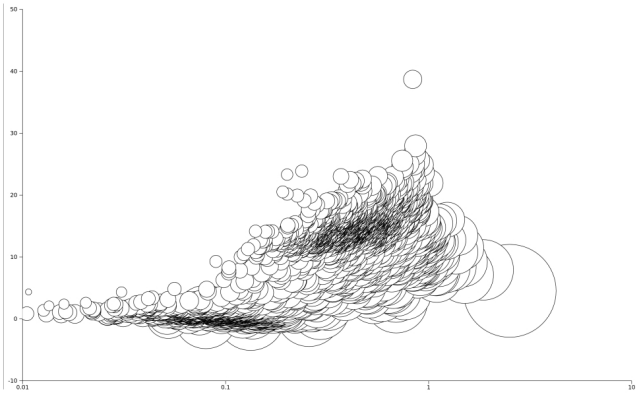
C



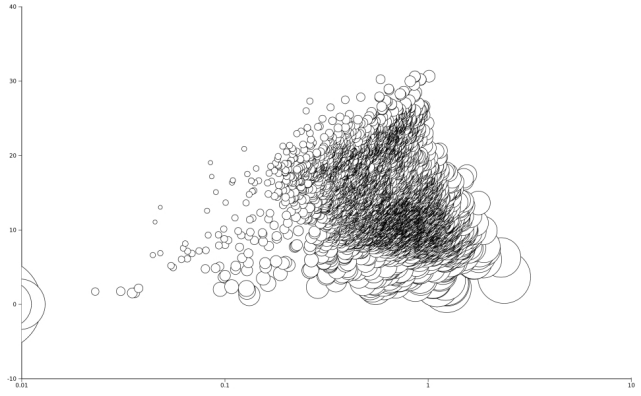
D



E



F



G

Figure 1: Bubble plot of Ka/Ks synonymous and non-synonymous distances. Ka and Ks values are represented in the X and Y axes, respectively. Bubble diameters represent Ka/Ks ratios:. The X axis was log scaled. A: Reduced parasite Chlamydiae genomes. B: Reduced Firmicutes genomes C: Parasite reduced genome Proteobacteria. D: Parasite reduced genome Spirochaetes E: Symbiont reduced genome Proteobacteria. F: Under reduction Actinobacteria genomes G: Under reduction Proteobacteria genomes

If we consider the phylogenetic classification of the analyzed genomes, the reduced genomes from parasite species of the Chlamydiae phylum own the higher Ka/Ks mean values, followed by reduced genomes from parasite species of Spirochaetes, Proteobacteria, and Firmicutes. Genomes from parasite Proteobacteria that are undergoing reduction show higher Ka/Ks values than the reduced ones from endosymbionts, and parasite Firmicutes and Spirochaetes. However, no clear correlation was observed between Ka/Ks values and parasite intracellular location.

High Ka/Ks ratios have already been described for endosymbionts in comparison those of free-living species with non-reduced genomes, suggesting that genomes from endosymbionts are undergoing a process of genome deterioration due to host-association (Moran, 1996). Population bottlenecks and mutational biases were also suggested as drive forces behind the accelerated evolution of genomes from endosymbionts. These forces would lead to a higher fixation rate of deleterious mutations, with the consequent higher Ka/Ks values (Woolfit and Bromham, 2003). On the other hand, in reduced genomes from parasite species, the higher number of genes under positive selection could be evidence of an adaptive process driven by the constant competition between the parasite and its host (Emelianov, 2007). This coevolutionary relationship would not be as strong for bacteria with symbiont lifestyles, for which selection would acts on genes that contribute to the overall fitness of the host (Brownlie *et al*, 2007).

To test the hypothesis that reduced genomes form parasite bacteria are undergoing an adaptive evolutionary process, the diversity of values generated in the Ka/Ks

study was analyzed using a multivariate hierarchical clustering, where the Ka/Ks ratio was used as a primary variable. As auxiliary variables for discriminating genes with similar evolutionary behavior we used Ks/Ka; Ka-Ks; the identity of protein sequences minus the identity of DNA sequences (an index that correlates with 1/*log* Ka); the identity of DNA sequences over the identity of protein sequences (an index linearly correlated with Ka); and a variable defined as (*PROT**-*DNA**)/(*DNA**/*PROT**), where *PROT** is the protein identity and *DNA** is the DNA sequence identity for each ortholog pair. Figure 2 shows the resulting hierarchical clustering for reduced genomes from parasite Firmicutes and Chlamydia (a color version of Figure 2 is provided as Supplemental Material S7). Reduced genomes from both Firmicutes and Chlamydia parasite species show heterogeneous evolution patterns with distinct groups for genes under positive and negative selection.

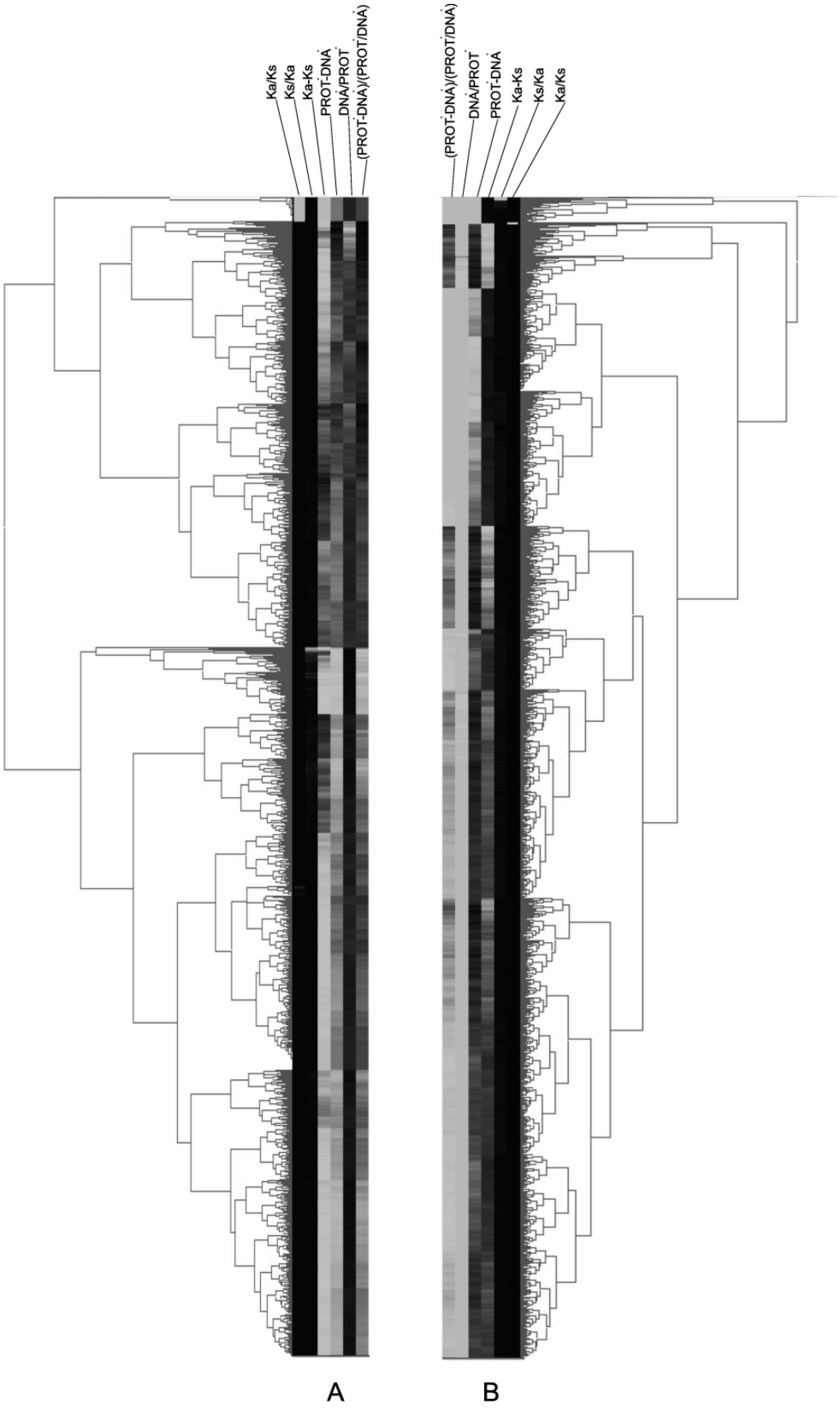

Ka/Ks
Ks/Ka
Ka-Ks
PROT-DNA
DNA/PROT
(PROT-DNA)/(PROT/DNA)
(PROT-DNA)/(PROT/DNA)
DNA/PROT
PROT-DNA
Ka-Ks
Ks/Ka
Ka/Ks
A
B

Figure 2: Hierarchical clustering of six variables associated with ortholog gene pairs. Variable names are indicated on the top of the figure. A: ortholog gene pairs values from parasite Chlamydia reduced genomes. B: ortholog gene pairs from parasite Firmicutes reduced genomes. Different shades of grey denote the normalized values (scaled from 0 to 1) of analyzed variables, with black corresponding to 0 and the lightest gray to 1. *PROT** and *DNA** refers to the sequence identity for aminoacid and DNA sequences respectively.

Interestingly, both Chlamydiae and Mycoplasma genomes presented positive and purifying evolution on several different ribosomal proteins (Supplemental material S8). For *Mycoplasma hyopneumoniae*, the presence of non-neutral mutations on ribosomal proteins could have a correlation with antibiotic resistance, as mutations on L22 ribosomal proteins could confers resistance to macrolides (Zaman, 2007), a common antibiotic used against Mycoplasmas.

Parasite Chlamydia reduced genomes present 574 ortholog pairs with Ka/Ks > 1, 167 of them annotated as hypothetical proteins. Annotated genes with a putative role in pathogenicity that were found to be under positive selection included those coding for invasin proteins (Burall *et al*, 2007), type III secretion protein SctC (Beeckman *et al*, 2008),YopC/Gen secretion protein D (Francis *et al* 2001), and numerous membrane transporters and exporting proteins. A reannotation of previously unidentified genes was also conducted, and allowed the identification of additional pathogenicity-related genes that would be under positive selection, including those coding for IncA inclusion membrane proteins (Brown *et al*, 2002), MarC multiple antibiotic transporter (Maneewannakul and Levy, 1996), CesT chaperone (Thomas *et al*, 2005), ComL DNA uptake lipoprotein (Oldfield *et al*, 2008), and several other membrane proteins, such as putative type III secretion system proteins, YjgP/YjgQ permeases (Ruiz *et al*, 2008), and FtsX permeases (Mir *et al*, 2006).

In the group of parasite Firmicutes with reduced genomes, *Mycoplasma hyopneumoniae* genomes presented the largest number of genes under positive selection. Among these genes, GreA, which codes a bacterial RNA polymerase

(RNAP) regulating protein, was found to be under positive selection in the 3 sequenced *M. hyopneumoniae* genomes (with Ka/Ks above 1 for all comparisons between strains), providing an interesting relationship with pathogenicity. GreA protein functions inducing nucleolytic activity on RNAP, thus enhancing promoter clearance and allowing the recycling of arrested transcription (Fish and Kane, 2002; Borukhov, Lee, Laptenko, 2005). In different species, GreA has been considered involved with an important component of cellular stress responses, such as exposition to antibiotics and pH culture medium changes (Stepanova *et al*, 2007; Hutchison, 1999; Singh *et al*, 2001; Len *et al,* 2004). The coupling of transcription and translation in *M. hyopneumoniae* was not directed studied yet, however it is reasonable to suppose that during the complex phase of infection there should be an important efficiency correlation between the transcription and translation machineries.

The positive selection acting on GreA correlates with the overall genomic architecture of *M. hyopneumoniae* genome. It was suggested that transcription efficiency is related to gene disposition on the leading or lagging strain, with genes positioned in the lagging strand being more prone to be involved in collision events between the replication and transcription machineries (Elías-Arnanz, 1997). Such collisions could stall the transcriptional processes or even interrupt them at all, which would force essential genes, with imperative expression requirements, to be preferentially located in the leading strand (Rocha and Danchin, 2003). Using GESPI as a quantitative index of gene strand bias, we have already shown (Oliveira de Carvalho and Ferreira, 2007) that the *M. hyopneumoniae* genome is the less structured one among Mollicutes species

regarding gene strand bias. This loss of strand preference could be caused by constant rearrangements and recombinations. Differently from the reduced genomes from symbiont bacteria, that remain stable for long periods of time (Tamas *et al*, 2002), those from parasite bacteria have larger amounts of rearrangements (Rocha, 2003). In *M. hyopneumoniae*, for instance, both small and large rearrangements have been found among the genomes of 3 different strains (Vasconcelos *et al*, 2004). Since genome rearrangements are likely to be a strong evolutionary mechanism in mycoplasmas (Rocha and Blanchard, 2002), the positive selection evidenced for the GreA gene points to a putative compensatory adaptation, for improvement of transcription efficiency in a genomic environment with a high propensity to stall during transcript elongation.

The adenylate kinase (ADK) gene from *Mycoplasma hyopneumoniae* 232 constitutes another interesting example of also positively selected gene from a reduced genome that is possibly related to pathogenicity. ADK was found to be correlated with virulence in *Pseudomonas aeruginosa*, inducing host macrophage cell death, possibly though the generation of a toxic mixture of AMP, ADP, and ATP (Markaryan *et al*, 2001). However, its expression in *M. hyopneumoniae* is decreased upon infection (Madsen *et al*, 2001) , which can be attributed to host immunity modulation by *M. hyopneumoniae*. ADK of *M. hyopneumoniae* has also been shown to be one of the most variable proteins in a study of field isolates from diverse geographic locations (Mayor *et al*, 2008), which can indicate its importance to parasite adaptation.

Synonymous codon analysis has shown that codon bias correlates with GC content both for small (R^2 = 0.90) and under reduction (R^2=0.86) genomes using

the SCUO metric (Wan *et al*, 2004). However, codon bias is more pronounced in small Mollicutes (Supplemental material S9), significantly differing from all other studied genomes (t-test, $p<0.05$). This could be related to the extreme low GC content of Mollicutes genomes. However, under reduction genomes of the Actinobacteria phylum, that are biased toward GC do not have higher SCUO values as Mollicutes. Also, small symbiont Proteobacteia genomes, with low GC content, have significantly lower codon bias values (t-test, p= 4.36e-118) than Mollicutes genomes. Higher codon bias can be the result of diverse biological factors, such as translational selection (dos Reis *et al*, 2004) mRNA structure stabilization (Seffens and Digby, 1999) and gene product protein composition (Lobry and Gautier, 1994). Codon bias in Mollicutes, though, appears to be the result of mutational pressure towards AT, as evidenced by the high G to A and C to T transition substitutions (Supplemental Material S10) and the reading of the stop codon UGA as tryptophan. Small genomes can have about 50% less tRNA genes as genomes under reduction (Supplemental material S1), and if the stochastic nature of tRNA pool is considered, we should have a correlation between tRNA gene number and codon bias (Bulmer, 1987). In agreement with that hypothesis, codon bias in small and under reduction genomes is correlated with the number of tRNAs present in the genome with a power law fit (R^2=0.42). As a higher nucleotide transition substitution rate towards AT is also found in other small genomes with higher codon bias, such as *Ehrlichia* species genomes (Supplemental Material S10), it is possible that codon bias in small reduced genomes are being driven both by mutational bias as well as optimization to the reduced tRNA gene repertoire.

Some genome features common to both parasite and symbiont bacteria are suggestive that the process of genome reduction started in the same manner for both. Genome reduction could have started by the strong competition between the respective ancestral unreduced genome and their host in a parasitic relationship, as a form to counter-act the pressures of the host immune system and overall optimization of the parasitism process through adaptative loss of function (Olson, 1999). This can be chronologically fit specially in the Mycoplasma case, as their diversification match with the emergence of the tetrapods (Vasconcelos *et al*, 2005). Another common feature of both parasites and symbionts is their lack of complete DNA repair metabolic routes (Carvalho *et al*, 2005), that has severe implications on genome compositional, as was already show for endosymbiont genomes, where loss of DNA recombinational repair enzymes coincide with mutational bias (Klasson and Andersson, 2006). Symbionts also share the signature of a great level of genome rearrangements as evidenced by its overall genome disordering. Comparing the average GESPI index for both symbionts and parasite reduced genomes, we found that symbionts genomes are in some cases even more disorganized in relation to gene collocation in the lag or lead strand than parasite ones (Table 1). This can be interpreted as an ancestral signature of genome rearrangement. Such accelerated evolution could have provided the necessary variation in the initial unreduced bacterial population to establish a successful parasitic lifestyle. As time passed, an equilibrium state between the host and the parasite could be reached and in agreement with the host biology, the initial parasitism state could be reversed to symbiosis in the case of insects, with less elaborated immune systems. In the case of, i.e. mammalian parasites like

Mycoplasmas the parasitism state could have been maintained as the result of an increase in selection pressure offered by its host evolving complex immune system. In this last case, a great level of refining in the parasitism was achieved, with the Mycoplasmas developing complex mechanisms for the evading of host immune system, like phase-variation of membrane proteins(Denison *et al*, 2005), constant genome rearrangements (as a form of diversity generation), and host immune response modulation (Simecka *et al*, 1993).

Table 1: GESPI values for selected symbionts and parasite genomes used in this work. The “Genome” column indicates if the genome is considered a reduced genome or a genome under reduction.

Species	Phylum	Genome	GESPI	Life Style
Chlamydophila felis Fe/C-56	Chlamydiae	reduced	0.01	parasite
Baumannia cicadellinicola Hc	Proteobacteria	reduced	0.61	symbiont
Mycoplasma synoviae 53	Firmicutes	reduced	1.28	parasite
Wolbachia endosymbiont of D. melanogaster	Proteobacteria	reduced	2.51	symbiont
Mycoplasma hyopneumoniae 232	Firmicutes	reduced	3.38	parasite
Chlamydia muridarum Nigg	Chlamydiae	reduced	3.57	parasite
Buchnera aphidicola APS	Proteobacteria	reduced	4.63	symbiont
Buchnera aphidicola Bp	Proteobacteria	reduced	4.64	symbiont
Mycoplasma hyopneumoniae J	Firmicutes	reduced	4.79	parasite
Rickettsia bellii RML369-C	Proteobacteria	reduced	4.98	parasite
Bordetella pertussis Tohama I	Proteobacteria	under	5.44	parasite
Buchnera aphidicola Sg	Proteobacteria	reduced	5.46	symbiont
Ehrlichia canis Jake	Proteobacteria	reduced	5.89	parasite
Mycoplasma hyopneumoniae 7448	Firmicutes	reduced	6.14	parasite
Anaplasma marginale St. Maries	Proteobacteria	reduced	6.23	parasite
Buchnera aphidicola Cc	Proteobacteria	reduced	7.1	symbiont
Bartonella quintana Toulouse	Proteobacteria	under	8.23	parasite
Anaplasma phagocytophilum HZ	Proteobacteria	reduced	10.14	parasite
Mycoplasma mobile 163K	Firmicutes	reduced	10.18	parasite
Mycoplasma pulmonis UAB CTIP	Firmicutes	reduced	13.36	parasite
Mycoplasma mycoides mycoides SC PG1	Firmicutes	reduced	14.43	parasite
Mycobacterium leprae TN	Actinobacteria	under	15	parasite
Campylobacter concisus 13826	Proteobacteria	under	16.6	parasite
Ureaplasma parvum sv 3 ATCC 700970	Firmicutes	reduced	17.49	parasite
Borrelia garinii PBi	Spirochaetes	reduced	20.19	parasite
Borrelia burgdorferi B31	Spirochaetes	reduced	20.27	parasite
Borrelia afzelii PKo	Spirochaetes	reduced	20.95	parasite
Onion yellows phytoplasma OY-M	Firmicutes	reduced	25.01	parasite
Aster yellows witches-broom phytoplasma AYWB	Firmicutes	reduced	26.02	parasite
Mycoplasma capricolum capricolum ATCC 27343	Firmicutes	reduced	32.19	parasite
Mycoplasma pneumoniae M129	Firmicutes	reduced	45.44	parasite
Mycoplasma gallisepticum R	Firmicutes	reduced	45.9	parasite
Mycoplasma genitalium G37	Firmicutes	reduced	46.05	parasite
Mycoplasma penetrans HF-2	Firmicutes	reduced	50.33	parasite
Mesoplasma florum L1	Firmicutes	reduced	69.62	heterotroph

Another evidence that genome reduction could have started with parasitism is the fact that symbionts of plants, that do not possess active immune systems also do not have small genomes , such as *Bradyrhizobium japonicum*, with a genome of 9.1 Mb (Kaneko *et al*, 2002), despite its intimate contact with their hosts. Actually, symbionts and parasites present peculiar characteristics that can be correlated with theirs lifestyles. The increased rate of non-synonymous mutations over synonymous mutations for parasites over symbionts also indicates that its "arms race" with the immune system of their host has pushed parasite genomes to a constantly accelerated evolution, in contrast to the degenerative processes that seems established in symbionts, inhabitants of a highly stable environment. For small genome parasites, especially for Mollicutes that do not have a cell wall and thus, could not survive for long periods in the environment, the bottlenecks that their population are subjected in each infection round, due to immune system attack and infection transmission, could lead to extinction if they are not capable of quickly responses to counter-act the host immune system. A great diversity has been shown for field isolates of Mycoplasmas (Mayor *et al*, 2008; Calus *et al*, 2007), showing that populations have a great diversity even for closely related strains, ranging from phase variation in surface genes (de Castro *et al*, 2006) to large rearrangements (Vasconcelos *et al*, 2005). Indeed, an analysis of rearrangement distances conducted for 28 genomes, both reduced and under reduction, has shown that parasite bacteria possesses the most shuffled genomes. Genomes were fully aligned and their similarity breakpoints converted to rearrangement distances, which were subsequently normalized against the 16S Kimura distance (Supplemental material S11). Interestingly, reduced genomes

were more rearranged than genomes under reduction. This supports the theory that genome reduction starts with mutational degeneration of redundant genes that were afterwards lost due to rearrangements, as stated by the Domino theory of genome reduction (Dagan *et al*, 2006).

We thus trace a parallel of the genome reduction process with the physical entropy of gas molecules in a closed system, where reduction in the volume of a gas leads to an overall increase in the disorganization of the system. Once started, the genome reduction process pushed the genomes to adapt, increasing the "entropy" of the genome in the form of rearrangements, translocations, and increased mutation rate. This progressed through an equilibrium point, where adaptation to the host was achieved. From that point, symbionts and parasites behave like two different systems. For parasites, the "entropy" of rearrangements and increased mutation rate was maintained, directing to the adaptation for parasitism state and sustaining equilibrium with the pressures from the host immune system. On the other hand, the symbiosis state carried in a high stable environment, acted as an "escape valve" on the closed system, allowing for an overall decrease in entropy of genome evolution on symbionts compared to parasites. A recent work has demonstrated that symbiosis could indeed start as parasitism (Pannebakker *et al*, 2007). Wolbachia species are facultative intracellular parasites that manipulate their host reproductive behavior in favor of infected females, thus promoting its own surviving inside the host population. However, the parasitoid wasp *Asobara tabida* has show an increased dependence on the manipulations carried by Wolbachia to point where it could not reproduce without the association with it. In light of this, we could consider the Wolbachia at an intermediate step in the

processes of host adaptation where the equilibrium entropy has not been reached yet. Wolbachia also presents an interesting example of how transposable elements participate in the processes of genome reduction. All endosymbiont genomes of insects that present a stable genome do not harbor transposons (Supplemental material S12), while small parasite genomes have maintained such elements for a long time, including large ones like integrative-conjugative elements (Loreto *et al*, 2007; Vasconcelos, 2005). It is thought that transposons have a great importance in genome evolution, by providing variability through rearrangements and horizontal transfer (Loreto *et al*, 2007). This indicates that transposons could act as a source of "heat" that increases de entropy during genome reduction and helps maintain its equilibrium in parasite genomes. Also, the presence of transposons in the Wolbachia genome, despite it not being a parasite under de same evolutionary process as other parasite small genomes like Mycoplasmas, can be interpreted as an ancestral signature of the process of genome reduction.

Tandem repeats can also act as another source of variability that helps increment the entropy of the genome. The distribution of tandem repeats on reduced genomes correlates with the species lifestyle (Suplemental material S13) as parasites and symbionts. Although symbionts preserve a considerable amount of tandem repeats, this can be interpreted as remnants of the early genome reduction process, since their repeat units are small and less conserved than tandem repeats of parasite genomes (Suplemental material S13). Specifically in Mycoplasmas, tandem repeats provides an efficient mechanism of host immune evasion through phase variation (de Castro, 2006), expression switching (Liu *et al*, 2000) and rearrangements hotspots. This model of evolution for reduced genomes are in

agreement with the “Domino Theory” (Dagan, *et al*, 2006), that suggests that reduction occurred both by deletions of genes, possibly leading to disruption of metabolic pathway, and posterior mutational erosion of the orphaned genes. However, this process appears to not be occurring after the equilibrium between parasites and its hosts, although for symbionts, as diverse highly depleted Buchnera strains have been found (Pérez-Brocal *et al*, 2006; Gil *et al*, 2002). In addition, the Mullers rachet hypothesis does not readily apply to parasite reduced genomes as mutational process appears to be correlated to an adaptation to the parasite lifestyle instead of genome deterioration. This is evidenced by the presence of a higher non-synonymous over synonymous rate of substitutions on genes with possible implication for pathogenicity. Taken together with genomic features such as rearrangements, repeat presence, transposons and horizontal transfer we suggest that small parasite genomes are under a different evolutionary process than symbionts species.

The genome reduction process show to be a complex subject, especially for parasite small genomes, despite the existing representation that small genomes have undergoing a simplification over their evolutionary development. Through the analysis of Ka/Ks ratios and genome architecture analysis becomes clear that a great level of sophistication has been achieved by parasite reduced genomes, with a tightly adjustment to their lifestyle. Although more investigation is needed to address the origins of genome reduction, the proposed model of genome entropy provides a framework for future investigations on the subject, aggregating the currently apparent paradoxal differences observed in both symbiont and parasite reduced genomes.

Supplementary Material

S1: Complete list of genome sequences used in this work along with genome properties

S2: Complete list of outgroup species used to build the phylogenetic tree S3

S3: Phylogenetic tree of the species used in this work, along with 29 outgroups. The tree was generated using the NJ method, Kimura 2-parameters for the distance calculus and a bootstrap value of 1000 iterations.

S4: List of genomes and their grouping as used in the Ka/Ks analysis

S5: Complete data of the Ka/Ks analysis

S6: Average Ka. Ks and Ka/Ks grouped accordingly lifestyle and species Phyla

S7: Color version of figure 2 where the values of variables are represented by the variation of color from black (value = 0) to light green (value = 1).

S8: Tabulated data derived from the clustering analysis of the Firmicutes and Chlamydiae genomes. Both for Firmicutes and Chlamydiae, the cluster with high diversity where extracted, being respectively the one with genes under positive selection and the one with genes under highly negative selection.

S9: Boxplot of the codon bias index SCUO for all the phyla analysed in this work. The Y axis represent SCUO values as determined by the INCA software on the complete CDS for the genomes of the phyla analyzed in this work.

S10: Maximum composite likelihood estimate of the pattern of nucleotide substitution, as estimated by the software MEGA for 15 genomes of reduced and under reduction genomes.

S11: Normalized and raw breakpoint and inversion distances for 14 parasite and symbionts reduced genomes

S12: Transposon count and transposons over CDS number for all genomes studied in this work.

S13: Tandem repeat count, tandem total size, tandem repeat macth, tandem consensus size, tandem copy number and tandem period size for the 81 genomes used in this work

Acknowledgments

We thank CAPES and CNPq for support on this work.

References

Beeckman DS, Tomgeens, Timmermans JP, Van Oostveldt P, Vanrompay DC. 2008. Identification and characterization of a type III secretion system in Chlamydophila psittaci. Vet Res **39**: 23-27.

Benson G. 1999.Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27**: 573-80.

Borukhov S, Lee J, Laptenko O. 2005. Bacterial transcription elongation factors: new insights into molecular mechanism of action. Mol Microbiol. **55**:1315-24.

Brown WJ, Skeiky YA, Probst P, Rockey DD. 2002. Chlamydial antigens colocalize within IncA-laden fibers extending from the inclusion membrane into the host cytosol. Infect Immun. **70**:5860-4.

Brownlie JC, Adamski M, Slatko B, McGraw EA. 2007. Diversifying selection and host adaptation in two endosymbiont genomes. BMC Evol Biol. **7**:68.

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance.. Nature, **325**:728-730.

Burall LS, Liu Z, Rank R, Bavoil PM. 2007. The chlamydial invasin-like protein gene conundrum. Microbes Infect. **9**:873-80.

Calus D, Baele M, Meyns T, de Kruif A, Butaye P, Decostere A, Haesebrouck F, Maes D. 2007. Protein variability among Mycoplasma hyopneumoniae isolates. . Vet Microbiol.**120**:284-91.

Carvalho FM, Fonseca MM, Batistuzzo De Medeiros S, Scortecci KC, Blaha CA, Agnez-Lima LF. 2005. DNA repair in reduced genome: the Mycoplasma model. Gene **360**:111-9.

Conord C, Despres L, Vallier A, Balmand S, Miquel C, Zundel S, Lemperiere G, Heddi A. 2008 Long-term Evolutionary Stability of Bacterial Endosymbiosis in Curculionoidea: Additional Evidence of Symbiont Replacement in the Dryophthoridae Family. Mol Biol Evol. **25**:859-868.

Dagan T, Blekhman R, Graur D. 2006. The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Mol Biol Evol. **23**:310-6.

de Carvalho MO, Ferreira HB. 2007. Quantitative determination of gene strand bias in prokaryotic genomes. Genomics. **90**:733-40.

de Castro LA, Rodrigues Pedroso T, Kuchiishi SS, Ramenzoni M, Kich JD, Zaha A, Henning Vainstein M, Bunselmeyer Ferreira H. 2006. Variable number of tandem aminoacid repeats in adhesion-related CDS products in Mycoplasma

hyopneumoniae strains.. Vet Microbiol. **116**:258-69.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. **30**:2478-83.

Denison AM, Clapper B, Dybvig K. 2005. Avoidance of the host immune system through phase variation in Mycoplasma pulmonis.. Infect Immun. **73**:2033-9.

dos Reis M, Savva R, Wernisch L.Nucleic 2004. Solving the riddle of codon usage preferences: a test for translational selection. Acids Res.**32**:5036-44.

Elías-Arnanz M, Salas M. 1997Bacteriophage phi29 DNA replication arrest caused by codirectional collisions with the transcription machinery. . EMBO J. **16**:5775-83.

Emelianov I. 2007. How adaptive is parasite species diversity?. Int J Parasitol. **37**:851-60.

Fish RN, Kane CM. 2002. Promoting elongation with transcript cleavage stimulatory factors. Biochim Biophys Acta. **1577**:287-307.

Francis MS, Lloyd SA, Wolf-Watz H. 2001. The type III secretion chaperone LcrH co-operates with YopD to establish a negative, regulatory loop for control of Yop synthesis in Yersinia pseudotuberculosis. Mol Microbiol. **42**:1075-93.

Gil R, Sabater-Munoz B, Latorre A, Silva FJ, Moya A. 2002. Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci U S A. **99**:4454-8.

Goto M, Washio T, Tomita M. 2000. Causal analysis of CpG suppression in the Mycoplasma genome. Microb Comp Genomics.**5**:51-8.

Hutchison, C. A., S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, and J. C. Venter. 1999. Global transposon mutagenesis and a minimal Mycoplasma genome. Science **286**:2165-2169.

Ihaka, R.; Gentleman, R. 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5**:299-314.

Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S, Watanabe A, Idesawa K, Iriguchi M, Kawashima K, Kohara M, Matsumoto M, Shimpo S, Tsuruoka H, Wada T, Yamada M, Tabata S. 2002. Complete genomic sequence of nitrogen-fixing symbiotic bacterium Bradyrhizobium japonicum USDA110. DNA Res.**9**:189-97.

Khachane AN, Timmis KN, Martins dos Santos VA. 2007. Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes. . Mol Biol Evol. **2**:449-56.

Klasson L, Andersson SG. 2006. Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. Mol Biol Evol. **5**:1031-9.

Len, A. C., D. W. Harty, and N. A. Jacques. 2004. Stress-responsive proteins are upregulated in Streptococcus mutans during acid tolerance. Microbiology **150**:1339-1351.

Li H., Fu Y., 2007. NeutralityTest: novel software for performing tests of neutrality. Bioinformatics, submitted.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **9**:2178-89.

Liu L, Dybvig K, Panangala VS, van Santen VL, French CT. 2000. GAA trinucleotide repeat region regulates M9/pMGA gene expression in Mycoplasma gallisepticum. Infect Immun. **2**:871-6.

Lobry JR, Gautier C. 1994. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. **22**:3174-80.

Loreto EL; Ortiz MF; Porto JR. 2007. Insertion sequences as variability

generators in the *Mycoplasma hyopneumoniae* and *M. synoviae genomes.* Genet. Mol. Biol. **30**

Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S 2004. Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res. **32**:3781-91.

Madsen ML, Puttamreddy S, Thacker EL, Carruthers MD, Minion FC. 2008. Transcriptome changes in Mycoplasma hyopneumoniae during infection. Infect Immun. **76**:658-63.

Maneewannakul K, Levy SB. 1996. Identification for mar mutants among quinolone-resistant clinical isolates of Escherichia coli. . Antimicrob Agents Chemother. **40**:1695-8.

Markaryan A, Zaborina O, Punj V, Chakrabarty AM. J 2001. Adenylate kinase as a virulence factor of Pseudomonas aeruginosa. Bacteriol. **183**:3345-52.

Mayor D, Jores J, Korczak BM, Kuhnert P. 2008. Multilocus sequence typing (MLST) of Mycoplasma hyopneumoniae: a diverse pathogen with limited clonality. Vet Microbiol. **127**:63-72.

Mir MA, Rajeswari HS, Veeraraghavan U, Ajitkumar P. 2006. Molecular characterisation of ABC transporter type FtsE and FtsX proteins of Mycobacterium tuberculosis.Arch Microbiol. **185**:147-58

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. **93**:2873-8.

Moret, B.M.E., Wyman, S., Bader, D.A., Warnow, T., and Yan, M., 2001. A new implementation and detailed study of breakpoint analysis, Proc. 6th Pacific Symp. on Biocomputing (PSB 2001), 583-594.

Oldfield NJ, Donovan EA, Worrall KE, Wooldridge KG, Langford PR, Rycroft AN, Ala'aldeen DA. 2008.. Identification and characterization of novel antigenic vaccine candidates of Actinobacillus pleuropneumoniae. Vaccine **26**:1942-54.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. **64**:18-23.

Pannebakker BA, Loppin B, Elemans CP, Humblot L, Vavre F. 2007. Parasitic inhibition of cell death facilitates symbiosis. Proc Natl Acad Sci U S A. **104**:213-5.

Pinto PM, Chemale G, de Castro LA, Costa AP, Kich JD, Vainstein MH, Zaha A, Ferreira HB. 2007. Proteomic survey of the pathogenic Mycoplasma

hyopneumoniae strain 7448 and identification of novel post-translationally modified and antigenic proteins.Vet Microbiol. **121**:83-93.

Price AL, Jones NC, Pevzner PA. . 2005. De novo identification of repeat families in large genomes. Bioinformatics **1**:351-8.

Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A. 2006. A small microbial genome: the end of a long symbiotic relationship? Science **5797**:312-3.

Rocha EP, Blanchard A. 2002. Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. Nucleic Acids Res. **30**:2031-42.

Rocha EP, Danchin A. 2003. Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet. **34**:377-8.

Rocha EP. 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res. **13**:1123-32.

Ruiz N, Gronenberg LS, Kahne D, Silhavy TJ. 2008. Identification of two inner-membrane proteins required for the transport of lipopolysaccharide to the outer membrane of Escherichia coli. Proc Natl Acad Sci U S A. **105**:5537-42.

Seffens W, Digby D: 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences.. Nucleic Acids Res, **27**:1578-1584.

Seo J., Gordish-Dressman H., Hoffman E., 2006. An Interactive Power Analysis Tool for Microarray Hypothesis Testing and Generation, Bioinformatics **22**:808-814.

Silva FJ, Latorre A, Moya 2003. Why are the genomes of endosymbiotic bacteria so stable? Trends Genet. **4**:176-80.

Simecka JW, Ross SE, Cassell GH, Davis JK. 1993. Interactions of mycoplasmas with B cells: antibody production and nonspecific effects. Clin Infect Dis. l **1**:176-82.

Singh, V. K., R. K. Jayaswal, and B. J. Wilkinson. 2001. Cell wall-active antibiotic induced proteins of Staphylococcus aureus identified using a proteomic approach. FEMS Microbiol. Lett. **199**:79-84.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehv?slaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. **10**:1611-8.

Stepanova E, Lee J, Ozerova M, Semenova E, Datsenko K, Wanner BL, Severinov K, Borukhov S. 2007. Analysis of promoter targets for Escherichia coli transcription elongation factor GreA in vivo and in vitro. J Bacteriol. **189**:8772-85.

Tamas I, Klasson L, Canbuck B, Noslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science. **5577**:2376-9.

Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0 2007. Mol Biol Evol. 2007 **8**:1596-9.

Thomas NA, Deng W, Puente JL, Frey EA, Yip CK, Strynadka NC, Finlay BB. 2005. CesT is a multi-effector chaperone and recruitment factor required for the efficient type III secretion of both LEE- and non-LEE-encoded effectors of enteropathogenic Escherichia coli. Mol Microbiol. **6**:1762-79.

Vasconcelos AT, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, along with 81 authors. 2005. Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. J Bacteriol. **16**:5568-77.

Wan XF, Xu D, Kleinhofs A, Zhou 2004.. Quantitative relationship between

synonymous codon usage bias and GC composition across unicellular genomes. J.BMC Evol Biol. **4**:19.

Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. Mol Biol Evol. **9**:1545-55.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. **8**:1586-91.

Zaman S, Fitzpatrick M, Lindahl L, Zengel J. 2007. Novel mutations in ribosomal proteins L4 and L22 that confer erythromycin resistance in Escherichia coli. Mol Microbiol. **66**:1039-50.

3. Discussão geral

Apesar de implicitamente representarem um sistema simples, os genomas reduzidos de bactérias, tanto de espécies simbiontes quanto parasitas, representam um grande desafio em termos de compreensão dos seus processos evolutivos. Notadamente, bactérias simbiontes e parasitas possuem atualmente diferentes mecanismos evolutivos, apesar de no passado, poderem ter compartilhado dos mesmos fatores que levaram à redução genômica.

Isto é sugerido pelas observações feitas no segundo artigo que faz parte desta dissertação. Entre elas encontra-se a perda de genes relacionados ao reparo de DNA, baixo conteúdo GC e a falta de colocação preferencial de genes nas fitas senso e anti-senso, o que pode refletir uma assinatura ancestral de rearranjos nos genomas simbiontes. Os genomas reduzidos de espécies de bactérias simbiontes e parasitas também compartilham maiores taxas mutacionais que espécies relacionadas de genoma não reduzido. No entanto, as bactérias de genoma reduzido parasitas, apresentam taxas mutacionais ainda maiores que as simbiontes, como descrito também no segundo trabalho desta dissertação.

De especial importância no combate aos micoplasmas é o fato de muitas proteínas localizadas ou externas a membrana celular destes apresentar elevada taxa evolutiva. Esta observação tem implicações diretas para o desenvolvimento de vacinas para humanos e animais, onde a intensa geração de variabilidade, desenvolvida pelos micoplasmas como compensação à limitada capacidade de adaptação de seu pequeno genoma, pode tornar ineficazes vacinas recombinantes que não levarem em consideração pontos altamente variáveis do genoma. Da mesma forma como bacterinas provenientes de uma única linhagem ou de linhagens originadas de pontos geográficos próximos.

A grande heterogeneidade evolutiva observada no genoma de micoplasmas pode

explicar os resultados encontrados na análise filogenômica realizada por Yotoko & Bonatto (2007). Tal trabalho encontrou inconsistências na geração de uma árvore filogenética de Mollicutes utilizando dados concatenados de genes ortólogos. Como os genes de micoplasmas apresentam taxas evolutivas divergentes, que estão intimamente acopladas a suas características funcionais, o uso conjunto desse grupo de dados provavelmente produziu um nível de ruído tal que dificultou a resolução dos métodos de classificação filogenética.

Dessa forma, para que se tenha uma maior probabilidade de resolução na análise filogenômica de Mollicutes, sugere-se que seja utilizado um conjunto de genes ortólogos com taxas evolutivas semelhantes. Este grupo de genes pode ser determinado através da técnica descrita no artigo que corresponde ao capítulo 2.2 deste trabalho.

Esta técnica, empregada pioneiramente neste trabalho, consiste da análise multivariada de um conjunto de variáveis relacionadas ao perfil evolutivo de cada gene ortólogo em um grupo específico de genomas relacionados evolutivamente relacionados. Uma analogia pode ser traçada com a análise de expressão de grandes conjuntos de genes através de microarranjos, onde variáveis como nível de expressão, tecido e condição experimental são analisadas em conjunto através de agrupamento hierárquico para determinar padrões de expressão gênica (Choi & Do, 2008).

No caso da técnica usada no trabalho do capítulo 2.2, as variáveis utilizadas relacionam-se ao perfil evolutivo de cada gene ortólogo em um grupo filogenético definido. Assim, foi possível identificar conjuntos de genes que apresentam as mesmas características evolutivas, através do uso de variáveis como número de mutações sinônimas, número de mutações não-sinônimas, identidade média da seqüência de proteínas e DNA, além de suas relações como a proporção de mutações não-sinônimas sobre mutações sinônimas.

Esta última é um importante índice que permite identificar genes sob pressão seletiva diferencial positiva ou negativa (Liberles & Wayne, 2002) e foi usada como variável primária no processo de determinação dos grupos identificados por agrupamento hierárquico. Aplicando esta técnica em 118.170 pares de genes ortólogos divididos em 7 grupos funcionais, foi possível traçar o perfil evolutivo de genes em genomas reduzidos de parasitas e simbiontes, evidenciando diferenças ainda não relatadas em relação à dinâmica evolutiva das espécies em estudo.

Devido às características populacionais já descritas de genomas de endossimbiontes (Moran, 1996), estes apresentam uma maior taxa de fixação de mutações deletérias em função dos gargalos populacionais aos quais suas espécies estão sujeitas do que espécies extracelulares (não-parasitas). Este fenômeno conduz, portanto, a um aumento generalizado das taxas evolutivas dos genomas de endossimbiontes em relação aos de outros organismos relacionados filogeneticamente, mas que apresentam nicho ecológico diferenciado.

No entanto, os dados publicados no segundo artigo descrito nesta dissertação evidenciam que genomas reduzidos de bactérias parasitas apresentam taxas evolutivas médias ainda maiores do que as de genomas reduzidos de endossimbiontes, apesar de

estarem submetidos a gargalos populacionais menos intensos que estes. A explicação para tal paradoxo pode estar na diferença de estilo de vida (parasitário/patogênico ou de endossimbiose) das diferentes espécies comgenomas reduzidos.

Enquanto as espécies endossimbiontes gozam de uma relativa estabilidade em seu habitat e, portanto, não necessitam de mecanismos atuantes para incremento de adaptabilidade (*fitness*), as espécies parasitas estão em constante conflito com seus hospedeiros. Esta condição das espécies parasitas as levam a uma situação de "adaptação ou extinção", mantida pela constante pressão seletiva derivada dos mecanismos de proteção dos respectivos hospedeiros.

No caso de bactérias parasitas de mamíferos como *M. hyopneumoniae*, a pressão de seleção exercida pelo sistema imune é extramemente alta (Wodarz, 2003) e uma grande plasticidade é exigida do parasita para que o processo de infecção seja completado com sucesso.

Em Mollicutes esta plasticidade se reflete na hipervariabilidade observada em genes de proteínas de membrana, que estão diretamente relacionadas aos processos de reconhecimento por parte do sistema imune do hospedeiro. Tal variabilidade se manifesta tanto em modificações relacionadas a variação de fase, mediada por repetições em tandem (de Castro *et al*, 2006), quanto a variações derivadas de processos mutacionais.

Portanto, para micoplasmas, pode-se supor que os processos mutacionais derivados dos gargalos populacionais, assim como a perda de genes relacionados ao reparo de DNA (ocorrida durante o processo de redução genômica), podem ter um caráter adaptativo, facilitando o escape do sistema imune e provendo a espécie parasita de genoma reduzido com uma rápida capacidade de adaptação a um ambiente hospedeiro hostil.

A geração de variabilidade parece ser particularmente importante para genomas de

micoplasmas do grupo hominis e, dentro deste grupo, para *Mycoplasma hyopneumoniae*. Os genomas de linhagens desta espécie são os que apresentam as maiores médias de mutações sinônimas sobre não-sinônimas, além do maior número de genes que apresentam valores de Ka/Ks acima de 1, evidenciando seleção positiva. Adicionalmente, são os genomas com maior número de rearranjos e com presença de recombinação (Mayor *et al*, 2008), fatores importantes para rápida geração de diversidade genética. Também são os genomas com o maior nível de desorganização em relação a preferência de colocação dos genes nas fitas senso e anti-senso (de Carvalho e Ferreira, 2007), evidenciando uma longa história evolutiva de rearranjos.

Interessantemente, o genoma da espécie de micoplasma não-parasita *Mesoplasma florum* é o que apresenta o maior nível de organização, inclusive que de outras espécies fora do filo Firmicutes (de Carvalho e Ferreira, 2007). Estas observações estão em concordância com a hipótese de que genomas de espécies de micoplasmas parasitas necessitariam de uma maior freqüência evolutiva de rearranjos e recombinação como estratégia geradora de variabilidade, a qual teria como conseqüência uma perda da organização cromossômica dos genes, inclusive em relação à distribuição deles nas fitas senso e anti-senso.

Entretanto, na análise em grande escala de genes ortólogos, descrita no capítulo 2.2, verificou-se que o gene GreA está sob seleção positiva. O produto do gene GreA tem por principal função estabelecer a reciclagem de complexos de transcrição interrompidos, estimulando a atividade endonucleotídica da RNA-polimerase (Stepanova, 2007). O fato de apresentar um genoma teoricamente propício à formação de complexos de transcrição interrompidos, somado à evolução positiva da proteína GreA, leva à hipótese de que este gene estaria em processo evolutivo compensatório, a fim de contrabalançar a possível alta

taxa de colisões entre as maquinarias de replicação e transcrição. Ao mesmo tempo, isso permitiria que o genoma de *M. hyopneumoniae* mantivesse um estado dinâmico devido a rearranjos e recombinação.

Essa propriedade dos genomas de *M. hyopneumoniae* acrescenta uma vantagem evolutiva importante para a manutenção da sua condição de parasita permitindo que mecanismos geradores de variabilidade, como freqüentes rearranjos, não interfiram na eficiência dos processos celulares de replicação e transcrição.

Levando em consideração as observações relacionadas a taxas mutacionais e estrutura genômica, foi proposto um modelo evolutivo para genomas reduzidos de bactérias parasitas, que contrapõe-se ao modelo de degeneração genômica governado pelo princípio de "Mullers rachet" (Moran, 1996). Este modelo baseia-se em uma comparação entre os genomas de parasitas e o processo físico de entropia de um sistema fechado contendo um gás onde é aplicada determinada pressão.

Procura-se dessa forma explicar a relação entre a redução do genoma e o grande número de rearranjos observados em genomas reduzidos de espécies parasitas em contraponto com a estabilidade mostrada por genomas de espécies simbiontes. Assim, o processo de redução genomica é acompanhado de uma constante pressão para geração de rearranjos como forma de contrabalançar o tamanho reduzido do genoma frente a necessidade de rápida adaptação requerida pela modo de vida parasitário. Contribuem como "geradores de entropia" nesse sistema taxas mutacionais mais altas que as evidenciadas em simbiontes e a presença de recombinação.

Genomas reduzidos de espécies parasitas estariam, portanto, de acordo com esse modelo. Eles estariam submetidos a uma pressão de redução que produz propriedades adaptativas e incrementam a adaptabilidade das espécies parasitas. Estas propriedades adaptativas se refletiriam na alta taxa de rearranjos e na maior taxa mutacional.

É importante notar que a comparação entre o modo de vida de bactérias simbiontes e parasitas de genoma reduzido refletem mecanismos biológicos que se encaixam em um arcabouço Darwiniano clássico, sendo o modelo proposto uma metáfora a fim de salientar o papel dos mecanismos geradores de variabilidade para os genomas reduzidos de espécies parasitas.

Em outras palavras, na ausência de uma intensa ação selecionadora para adaptação a um ambiente hostil (como no caso do parasitismo), as bactérias simbiontes não evoluiram fortes mecanismos de geração de variabilidade, apresentando uma baixa taxa de rearranjos assim como um menor índice de substituições não-sinônimas sobre substituições sinônimas. Por outro lado a natureza da relação parasita-hospedeiro exige que a população parasita apresente diversidade suficiente para que a espécie possa adaptar-se à seleção aplicada pelo hospedeiro através do seus mecanismos de defesa. Em genomas reduzidos de parasitas, tal geração de variabilidade ocorre através de rearranjos, recombinação e alta taxa mutacional, se comparados com outros genomas reduzidos simbiontes assim como bactérias de genomas não-reduzido.

4. Conclusões e Perspectivas

A partir dos resultados gerados nesta dissertação, novas perspectivas se desdobram para a análise funcional e evolutiva de genomas reduzidos de espécies bacterianas parasitas.O grande número de informações geradas em relação aos perfis evolutivos dos genes de *M. hyopneumoanie* e outras espécies de Mollicutes poderá esclarecer diversos aspectos da biologia destas espécies, como no caso do gene GreA, que indica uma evolução adaptativa acoplada à arquitetura genômica de *M. hyopneumoniae*.

Do ponto de vista mais aplicado, a escolha de alvos para o desenvolvimento de vacinas poderá ser aprimorada se forem levados em consideração os aspectos evolutivos de cada gene alvo. Dessa forma, genes com regiões altamente variáveis poderão ser identificados e excluídos do processo de desenvolvimento de vacinas recombinantes, uma vez que a probabilidade de escape mutacional, ocasionado pela grande diversidade gerada pelos genomas de Mollicutes pode rapidamente tornar inefetiva uma estratégia de vacinação.

Além dos dados exibidos mais detalhadamente para o filo Firmicutes, durante este trabalho foram analizados outros quatro filos de eubacteria (Actinobacteria, Proteobacteria, Chlamydiae e Spirochaetes) cuja discussão será implementada em trabalhos adicionais.

Dado o exposto é possível concluir que índices numéricos derivados da transformação de dados simbólicos representativos da sequência de nucleotídeos de genomas completos constituem-se como importantes ferramentas para a análise comparativa de genomas.

O índice GESPI, desenvolvido como parte desta dissertação para análise da estrutura genômica de procariotos, indicou importantes diferenças estruturais na organização da disposição de genes nas fitas senso e anti-senso. Estas diferenças foram

correlacionadas claramente com a classificação filogenética das espécies em estudo assim como com características evolutivas, no caso da espécie *M. hyopneumoniae* e a presença de evolução positiva no gene GreA.

Da mesma forma, a disponibilidade crescente de genomas completos de procariotos abre a perspectiva de realização de análises evolutivas em larga escala, tal como a desenvolvida na seção 2.2 deste trabalho. Tal análise permitiu a identificação de padrões evolutivos diferenciados entre espécies de genoma reduzido parasitas e endossinbiontes, ampliando a compreensão sobre a evolução destas espécies.

Especificamente, foi possível identificar que dentre as espécies de genoma reduzido do filo Firmicutes, a espécie *M. hyopneumoniae* apresenta as maiores taxas de evolução positiva, assim como o maior número de genes sob essa condição. Esta conclusão tem importante repercusão no estudo da biologia desta bactéria, uma vez que, tanto o desenvolvimento de vacinas quanto o de fármacos para o combate de infecções crônicas causadas por *M. hyopneumoniae,* deve levar em consideração sua grande capacidade de adaptação.

De maneira ampla, é importante salientar que apesar do diminuto tamanho dos seus genomas, as espécies procarióticas parasitas com genoma reduzido apresentam um grande desafio em termos de saúde pública humana e animal. Desafio este, imposto principalmente pelas suas particularidades evolutivas e de organização genômica.

5. Referências bibliográficas

Bak AL, Black FT, Christiansen C, Freundt EA. Genome size of mycoplasmal DNA.*Nature*. 224:1209-10, 1969.

Calus D, Baele M, Meyns T, de Kruif A, Butaye P, Decostere A, Haesebrouck F, Maes D. Protein variability among Mycoplasma hyopneumoniae isolates.*Vet Microbiol*. 120:284-91, 2007.

Choi DK, Do JH. Clustering Approaches To Identifying Gene Expression Patterns from DNA Microarray Data. *Mol Cells*. 25, 2008.

de Castro LA, Pedroso T. R., Kuchiishi SS, Ramenzoni M, Kich JD, Zaha A, Vainstein M.H. Ferreira H.B. Variable number of tandem aminoacid repeats in adhesion-related CDS products in Mycoplasma hyopneumoniae strains. *Vet Microbiol.* 116:258-69, 2006.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 5223:496-512, 1995.

Grigoriev A., Freeman J. M., Plasterer T. N., Smith T. F., and Mohr S. C. Genome Arithmetic. *Science* 281:1923, 1998.

Herbeck JT, Funk DJ, Degnan PH, Wernegreen JJ. A conservative test of genetic drift in the endosymbiotic bacterium Buchnera: slightly deleterious mutations in the chaperonin groEL. *Genetics*. 165:1651-60, 2003.

Hsiao W, Wan I, Jones SJ, Brinkman FS. IslandPath: aiding detection of genomic islands in prokaryotes. .*Bioinformatics* ,19:418-20, 2003.

Iverson-Cabral SL, Astete SG, Cohen CR, Rocha EP, Totten PA. Intrastrain heterogeneity of the mgpB gene in Mycoplasma genitalium is extensive in vitro and in vivo and suggests that variation is generated via recombination with repetitive chromosomal sequences.Infect Immun;74:3715-26, 2006.

Kingsbury DT. Estimate of the genome size of various microorganisms *J Bacteriol*. 98:1400-1, 1969.

Liberles DA, Wayne ML. Tracking adaptive evolutionary events in genomic sequences. *Genome Biol*;3, 2002.

Lin Z, Madan D, Rye HS. GroEL stimulates protein folding through forced unfolding. *Nat Struct Mol Biol*;15:303-11, 2008.

Loreto EL; Ortiz MF; Porto JR. Insertion sequences as variability generators in the Mycoplasma hyopneumoniae and M. synoviae genomes *Genet. Mol. Biol*. 30. 2007.

Mayor D, Jores J, Korczak BM, Kuhnert P. Multilocus sequence typing (MLST) of Mycoplasma hyopneumoniae: a diverse pathogen with limited clonality *Vet Microbiol*. 127:63-72,2008.

Moran NA.. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*. 93:2873-8,1996.

Nagylaki T. The evolution of one- and two-locus systems. *Genetics*. 83:583-600, 1976.

Necsulea A, Lobry JR A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol Biol Evol* 10:2169-79. 2007.

Pinto, Paulo Marcos ; Carvalho, Marcos Oliveira de ; Alvez-Junior, L. ; Brocchi, M. ; Schrank, I. S. . Molecular analysis of an Integrative Conjugative Element, ICEH, present in the chromosome of different strains of Mycoplasma hyopneumoniae. *Genet. Mol. Biol*. 30: 256-263, 2007.

Rocha EP, Danchin A: Ongoing evolution of strand composition in bacterial genomes.*Mol Biol Evol*, ,18:1789-99, 2001.

Rocha EP, Danchin A. Essentiality, not expressiveness, drives gene-strand bias in bacteria. . *Nat Genet,* 34:377-8, 2003.

Rocha EP: The replication-related organization of bacterial genomes. *Microbiology*. 150:1609-27, 2004.

Rogers MJ, Simmons J, Walker RT, Weisburg WG, Woese CR, Tanner RS, Robinson IM, Stahl DA, Olsen G, Leach RH, *et al*. Construction of the mycoplasma evolutionary tree from 5S rRNA sequence data. *Proc Natl Acad Sci U S A* ; 82:1160-4,1985.

Song J, Ware A, Liu SL.Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. *BMC Genomics*. 4:17, 2003.

Stepanova E, Lee J, Ozerova M, Semenova E, Datsenko K, Wanner BL, Severinov K, Borukhov S. Analysis of promoter targets for Escherichia coli transcription elongation factor GreA in vivo and in vitro. *J Bacteriol*. 189:8772-85. 2007.

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ, Sandstrom JP, Moran NA, Andersson SG. 50 million years of genomic stasis in endosymbiotic bacteria. *Science*. 296:2376-9. 2002.

Tillier ER, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes,50:249-57. . *J Mol Evol*. 2000.

Thurman RE, Noble WS, Stamatoyannopoulos JA.Multi-scale correlations in continuous genomic data. *Pac Symp Biocomput*.;201-15, 2008.

van Ham RC, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A.Reductive genome evolution in Buchnera aphidicola.*Proc Natl Acad Sci U S A*.; 100:581-6, 2003.

Van Valen L. "A new evolutionary law". *Evolutionary Theory* 1:1-30. 1973.

Vasconcelos AT, Ferreira HB, Bizarro CV, Bonatto SL, Carvalho MO, Pinto PM, along with 81 authors.Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. . *J Bacteriol*. 2005 187:5568-77.

Yotoko KSC.; Bonatto SL, A phylogenomic appraisal of the evolutionary relationship of mycoplasmas. *Genet. Mol. Biol*. 30, 2007.

Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*. ;7:142. 2006.

Wodarz D. Ecological and evolutionary principles in immunology. *Ecol Lett* ;9:694-705. 2006.

Woese C. R.,Maniloff J.,Zablen L.B.; Phylogenetic analysis of the mycoplasmas. *Proc. Natl. Acad. Sci. USA*; 77:494-498. 1980.

6. Apêndices

Como apêndice encontra-se em anexo um CD com os dados resultantes das análises realizadas neste trabalho. Abaixo consta a lista de arquivos a descrição do seu conteúdo.

Pasta codon-bias – Contém os dados de tendência de uso de códon para os genomas descritos nesta dissertação. Os arquivos estão em formato CSV que pode ser lido no software Microsoft Excel ou em editor de textos comum.

Pasta dnds-data – Dados de análise de substituição sinônima e não-sinônima para os diferentes grupos genômicos apresentados nesta dissertação. Os dados estão agrupados em pastas de acordo com o grupo genômico em estudo sendo os grupos de ortólogos separados em sub-pastas assim como suas respectivas seqüências nucleotídicas e de aminoácidos. Os resultados da análise de substituição sinônima e não-sinônima de acordo com o software codeml encontra-se na sub-pasta "groups" nomeada de acordo com o grupo ortólogo a que se refere.

Pasta duplication – Dados de duplicação genômica inferidos através de alinhamento utilizando-se o software nucmer do pacote Mummer. Os arquivos estão nomeados de acordo com o nome do genoma em análise.

Pasta ortholog-groups – Grupos de ortólogos definidos de acordo com o software ORTHOMCL. Cada sub-pasta contém os grupos genômicos de acordo com a definição exposta nesta dissertação, sendo o arquivo com a extensão ".out" o que contém a relação dos genes ortólogos para o grupo em cada sub-pasta.

Pasta rearrangements – Contém em suas subpastas o resultado dos alinhamentos genomicos utilizados para determinar as coordenadas dos pontos de quebra de rearranjos no formato tabular definido pelo software nucmer e os resultados do software de determinação de distância de rearranjos GRAPPA.

Pasta repeats – A sub-pasta tandem contém os resultados da buscam por repetições em tandem de acordo com o software TRF. Os resultados estão compactados em formato ".tar.gz" por razões de espaço. Cada um dos dois arquivos contém respectivamente os resultados para genomas reduzidos e genomas em processo de redução. A subpasta "direct-inverse-repeats" contém os resultados da busca por repetições diretas e inversas para genomas reduzidos e genomas em processo de redução de acordo com a nomeação de cada arquivo e sub-pasta.

s1.xls – Material Suplementar S1 do artigo apresentado na seção 2.2

s2.xls – Material Suplementar S2 do artigo apresentado na seção 2.2

s3.jpg – Material Suplementar S3 do artigo apresentado na seção 2.2

s4.xls – Material Suplementar S4 do artigo apresentado na seção 2.2

s5.zip – Material Suplementar S5 do artigo apresentado na seção 2.2

s6.xls – Material Suplementar S6 do artigo apresentado na seção 2.2

s7.png – Material Suplementar S7 do artigo apresentado na seção 2.2

s8.xls – Material Suplementar S8 do artigo apresentado na seção 2.2

s9.png – Material Suplementar S9 do artigo apresentado na seção 2.2

s10.xls – Material Suplementar S10 do artigo apresentado na seção 2.2

s11.xls – Material Suplementar S11 do artigo apresentado na seção 2.2

s12.xls – Material Suplementar S12 do artigo apresentado na seção 2.2

s13.xls – Material Suplementar S13 do artigo apresentado na seção 2.2

filelist – Lista completa de arquivos presentes na mídia e sua localização nos subdiretórios

7. Curriculum Vitæ resumido

MARCOS OLIVEIRA DE CARVALHO

Publicações:

Livro, capítulo de livro ou artigo científico publicado em revista indexada, como autor principal

CARVALHO, Marcos Oliveira de ; SILVA, Joana ; LORETO, Elgion Lucio . Analyses of P-like transposable element sequences from the genome of Anopheles gambiae. Insect Molecular Biology, Inglaterra, v. 13, n. 1, p. 55-63, 2004

CARVALHO, Marcos Oliveira de ; FERREIRA, H. B. . Quantitative determination of gene strand bias in prokaryotic genomes. Genomics (San Diego), v. 90, p. 733-740, 2007.

Livro, capítulo de livro ou artigo científico publicado em revista indexada, como co-autor

VASCONCELOS, A. T. R. ; FERREIRA, H. B. ; BIZARRO, C. V. ; BONATTO, S. L. ; CARVALHO, Marcos Oliveira de ; PINTO, Paulo Marcos *et al* . Swine and poultry pathogens: the complete genome sequences of two strains of Mycoplasma hyopneumoniae and a strain of Mycoplasma synoviae. Journal of Bacteriology, v. 187, n. 16, p. 5568-5577, 2005.

PINTO, Paulo Marcos ; CARVALHO, Marcos Oliveira de ; ALVES-JUNIOR, L. ; BROCCHI, M. ; SCHRANK, I. S. . Molecular analysis of an Integrative Conjugative Element, ICEH, present in the chromosome of different strains of Mycoplasma hyopneumoniae. Genetics and Molecular Biology, v. 30, p. 256-263, 2007.

Resumos em anais de congresso como autor principal

CARVALHO, Marcos Oliveira de ; FERREIRA, H. B. ; SILVA, S. C. . Genome rearrangement analysis of three Mycoplasma hyopneumoniae strains.. In: 1st International Conference of the AB3C, 2005, Caxambu. Annals of the 1st International Conference of the AB3C, 2005.

CARVALHO, Marcos Oliveira de ; LORETO, Elgion Lucio ; SEPEL, Lenira . Caracterização das mutações somáticas observadas em um clone de violeta africana (Saintpaulia ionantha). In: XV Jornada Acadêmica Integrada, 2000, Santa Maria. Anais da XV Jornada Acadêmica Integrada, 2000.

CARVALHO, Marcos Oliveira de ; SEPEL, Lenira ; LORETO, Elgion Lucio . Detecção do luteovírus potato leafroll virus (PLRV) através da técnica de RT-PCR. In: XV Jornada Acadêmica Integrada, 2000, Santa Maria. Anais da XV Jornada Acadêmica Integrada, 2000.

CARVALHO, Marcos Oliveira de . Software para cálculos referentes a reações de PCR. In: 46 Congresso Nacional de Genética, 2000, Águas de Lindóia. Genetics and Molecular Biology, 2000. v. 23. p. 179-179.

CARVALHO, Marcos Oliveira de ; SILVEIRA, Lia Rejane Machado ; POLLETO, Nara ; SEPEL, Lenira ; STROBEL, Taísa . Efeito da acidez do meio nutritivo sobre o comportamento in vitro de batata. In: XIII Jornada Acadêmica Integrada, 1998, Santa Maria. Anais da XIII Jornada Acadêmica Integrada, 1998.

Resumos em anais de congresso como co-autor

GOLOMBIESKI, Ronaldo ; PIVETTA, Lucinéia ; GRAICHEN, Daniel ; CARVALHO, Marcos Oliveira de ; ROCHA, João Batista ; LORETO, Elgion Lucio ; NOGUEIRA, C W . Effects of diphenyl diselenide (PhSe)2 on aminolevulinate dehydratase (ALAD) from mouse (Mus musculus) and fruit fly (Drosophila melanogaster): A comparative study. In:

XXXI Reunião Anual da SBBq, 2002, Caxambu. XXXI Reunião Anual da SBBq - Programa e Resumos, 2002.

SEPEL, Lenira ; LORETO, Elgion Lucio ; CARVALHO, Marcos Oliveira de . Modificações no padrão de organização floral em quatro clones de violeta africana (Saintpaulia ionantha). In: 46 Congresso Nacional de Genética, 2000, Águas de Lindóia. Genetics and Molecular Biology, 2000. v. 23. p. 412-412.

SEPEL, Lenira ; CARVALHO, Marcos Oliveira de ; PINTO, Paulo Marcos ; GOLOMBIESKI, Ronaldo ; ROBE, Lizandra ; LORETO, Elgion Lucio . Tópicos contextualizados de Genética e Biologia Molecular. In: XIV Jornada Acadêmica Integrada, 1999, Santa Maria. Anais da XIV Jornada Acadêmica Integrada, 1999.

STROBEL, Taísa ; SILVEIRA, Lia Rejane Machado ; CARVALHO, Marcos Oliveira de ; SEPEL, Lenira ; SONZA JR, Arcelindo . Situação atual da produção de batata-semente livre de viroses para a região central do RS. In: XIII Jornada Acadêmica Integrada, 1998, Santa Maria. Anais da XIII Jornada Acadêmica Integrada, 1998.

Formação Acadêmica

Graduação em Farmácia pela Universidade Federal de Santa Maria, 2004

Bolsa de Iniciação científica ou aperfeiçoamento

Bolsista FIEX/UFSM com o projeto "Programa regional integrado de produção de batata-semente". Orientador: Profa. Lia Rejane Machado Silveira - Período: Junho a Dezembro de 1997

Bolsista PIBIC/CNPq com o projeto "Detecção do luteovírus "Potato Leafroll Virus" (PLRV) através da técnica de RT-PCR. Orientador: Prof. Elgion Lúcio da Silva Loreto - Período: Setembro de 1999 a Julho de 2000

Bolsista PIBIC/CNPq com o projeto "Avaliação de mutações florais em clones de violeta africana (Saintpaulia ionantha) e caracterização das formas mutantes em um clone hipermutável para morfologia floral". Orientador: Profa. Lenira Sepel - Período: Agosto de 2000 a Julho de 2001

Estágios:

Laboratório de Biotecnologia Aplicada à Produção Vegetal. Período: Março de 1997 a Setembro de 1998. Total de 160 horas. Responsáveis: Profa. Lenira Sepel e Profa. Lia Rejane Machado da Silveira

Empresa Procampo - Planejamento e Levantamentos Rurais. Período: Janeiro de 1998 a Setembro de 1998. Total de 550 horas. Responsável: Engº Agrº Oscar Luiz Moreira de Carvalho

Laboratório de Introdução e Expressão de Genes da Embrapa Recursos Genéticos e Biotecnologia. Período: 5 a 23 de Fevereiro de 2001. Total de 100 hroas. Responsável: Francisco J. L. Aragão

Curso de extensão, durante a graduação

Curso de Extensão "Produção de Plantas transgênicas" - 19 a 22 de Janeiro de 1999 - Total de 16 horas

Curso de Extensão "Clonagem Gênica e sua aplicação na Biotecnologia" - 10 a 11 de Abril de 2003 - Total de 8 horas

Curso "Aplicação dos marcadores moleculares no melhoramento genético de plantas" - 45º Congresso Brasileiro de Genética - 3 a 6 de outubro de 1999 - Gramado - RS

Curso "Uso da tecnologia do DNA recombinante para produção de plantas geneticamente

modificadas " - 45º Congresso Brasileiro de Genética - 3 a 6 de outubro de 1999 - Gramado - RS

Curso "Produção e análise de produtos transgênicos " - 46º Congresso Brasileiro de Genética - 3 a 6 de outubro - Aguas de Lindóia - SP

Curso "Bases moleculares do desenvolvimento vegetal " - 46º Congresso Brasileiro de Genética - 3 a 6 de outubro - Aguas de Lindóia - SP

Participação em congresso sem apresentação de trabalho

X Reunião Estadual de Biotecnologia Vegetal - REDBIO - 10 a 11 de dezembro de 1998 - Passo Fundo - RS

Seminário Estadual Biotecnologia e Produtos Transgênicos - 29 a 30 de Abril de 1999 - Santa Maria - RS

Jornada Farmacêutica Prof. Zósymo Lopes dos Santos - 25 a 28 de setembro de 2001 - Santa Maria - RS

Reunião de Anotação do genoma de Mycoplasma hyopneumoniae - 25 e 26 de novembro de 2006.UFRGS - Porto Alegre - RS

I Semana de Estudos Farmacêuticos da UFSM - 7 a 11 de abril de 2003 - Santa Maria - RS

3. EXPERIÊNCIA PROFISSIONAL

Seminários e palestras

Palestra "Bioinformática" durante aula da disciplina BIO99099 do Curso de Ciências

Biológicas da UFRGS - 4 de maio de 2005

Palestra "Bioinformática" durante aula da disciplina BIO99099 do Curso de Ciências Biológicas da UFRGS - 21 de setembro de 2005

Palestra "Genômica e Bioinformática" - Centro de Ciências da Saúde - UFSM - 4 de maio de 2004

Atividade docente em cursos de graduação, extensão, aperfeiçoamento, especialização ou pós-graduação sem vínculo empregatício

Curso "Introdução a Biologia Molecular Computacional" promovido pelo Departamento de Biologia Molecular e Biotecnologia e Laboratório de Genômica Funcional da Universidade Federal do Rio Grande do Sul– 06 de Julho de 2007 a 10 de Agosto de 2007 – Porto Alegre – RS – 75 horas

Mini-curso "Bioinformática e tecnologias emergentes em biologia molecular" - XX semana Acadêmica do Curso de Ciências Biológicas da UFSM - 28 de Julho de 2007 - 8 horas

Curso "Genomica Funcional e Estrutural" - Promovido pela Universidade Federal de Santa Maria - Coordenação do Curso de Ciências Biológicas - Centro de Ciências Naturais e Exatas - Laboratório de Biologia Molecular - LabDros. - 20 horas

Curso "Seqüenciamento e Análise de Genomas II" - Promovido pelo Grupo de Genômica Estrutural e Funcional - Centro de Biotecnologia - UFRGS - Total de 32 horas

Atividades em projetos de pesquisa, relacionados ao campo de atividade profissional

Bolsista DTI/CNPq no projeto Genoma Funcional de Mycoplasma synoviae e Mycoplasma

hyopneumoniae no período de agosto de 2004 a março de 2006.

Participação em cargos de direção de associações ou sociedades científicas

Integrante da Board of Directors da The Bioinformatics Organization de Fevereiro de 2003 a Março de 2006